Syntax: Linguistische Grundlagen

Katja Markert

Institut für Computerlinguistik
Uni Heidelberg
markert@cl.uni-heidelberg.de
mit einigen Folien von Yannick Versley

December 18, 2019

Übersicht

Übersicht

Was ist Syntax?

Syntax beschäftigt sich mit der **Grammatikalität** (Wohlgeformtheit) von Sätzen

- Colourless green ideas sleep furiously
- * Colourless ideas green furiously sleep

Sprecher einer Sprache verfügen über grammatisches Wissen

- Lexikon: Wörter und ihre Wortartenzugehörigkeit
- Regeln für die Bildung grammatisch wohlgeformter Sätze
- Struktur von Sätzen

Syntaktische Begriffe I: Konstituenz

Syntaktische Konstituente oder Phrase

Wortsequenz, die sich als eine syntaktische Einheit verhält

- [NP She] saw [NP him] [PP on Friday].
- [NP Women] saw [NP men with hats] [PP on Friday the seventeenth]
- Women flying on broomsticks saw old men with hats on Friday the seventeenth, January 2004

Syntaktische Begriffe I: Konstituenz

Permutationstest/Movability test: Konstituenten können als Einheiten unterschiedliche Positionen in einem Satz einnehmen

- She saw him [PP on Friday the seventeenth]
- [PP on Friday the seventeenth], she saw him
- * On Friday, she saw him the seventeenth.

Substitutionstest: Konstituenten können als Einheiten durch alternative Phrasen ersetzt werden.

- [Nach dem Spiel] waren viele Fans begeistert in die Innenstadt gezogen
- [Danach] waren viele Fans begeistert in die Innenstadt gezogen
- [Nach dem Spiel, auf das sie seit Monaten hingefiebert hatten], waren viele Fans

Syntaktische Eigenschaften I : Konstituenz

- Nomen oder Pronomina bilden Nominalphrase (NP)
  das Buch; das dicke Buch im Regal; er; ein Buch, das fasziniert
- Adjektive bilden Adjektivphrasen (AP)
  schön, schöner als das Wetter; sehr stolz auf seine Arbeit
- Präpositionen bilden Präpositionalphrasen (PP)
  hinter dem Haus; gleich nach der Vorlesung
- Verben bilden Verbalphrasen (VP)
  schläft: kauft dem Mädchen ein Buch; verspricht Max zu kommen

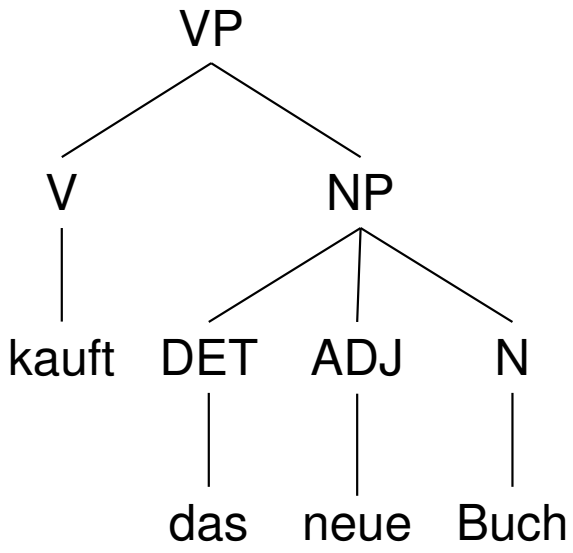
Syntaktische Begriffe I: Konstituenz

Hierarchische Anordnung von Konstituenten. Grammatik einer Sprache legt fest, welche Wörter/Wortarten/Phrasen nach welchen Prinzipien umfassendere Phrasen bilden können.

Syntaktische Begriffe I: Konstituenz

Hierarchische Anordnung von Konstituenten. Grammatik einer Sprache legt fest, welche Wörter/Wortarten/Phrasen nach welchen Prinzipien umfassendere Phrasen bilden können.

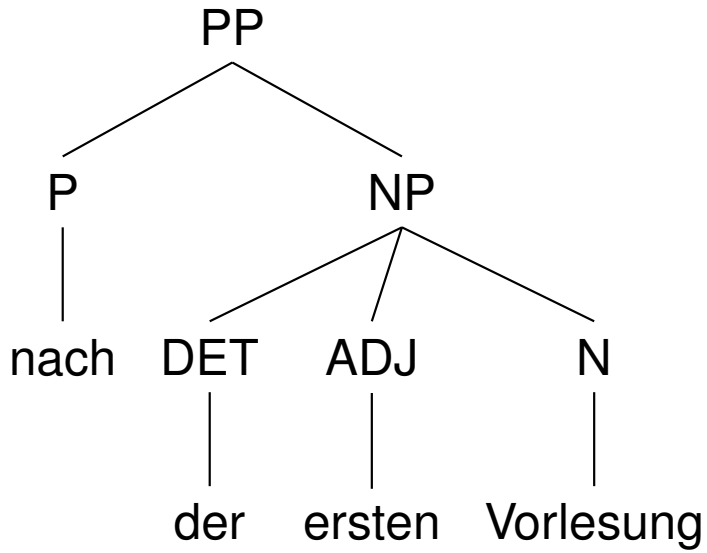
Phrasenstrukturbaum:



Klammerschreibweise: [$_{VP}$kauft [$_{NP}$das neue Buch]]

Syntaktische Begriffe I: Konstituenz

Hierarchische Anordnung von Konstituenten. Grammatik einer Sprache legt fest, welche Wörter/Wortarten/Phrasen nach welchen Prinzipien umfassendere Phrasen bilden können.



Klammerschreibweise: [$_{PP}$nach [$_{NP}$der ersten Vorlesung]]

Syntaktische Begriffe I: Konstituenz

- Vergleich mit n-grams: N-grams können innerhalb einer Konstituente aufhören oder beginnen und Konstituentengrenzen überschreiten: *The secretary types*
- Vergleich mit POS-Sequenzen: Kann man mit Sequenzen vonm POS-TAGs (REs über POS-tags) alle NPs in einem Text finden?
- Wir versuchen: DT* JJ* NN+. Passt zu
  - the lecturer
  - the weird German lecturer
  - university lecturers
- Probleme?

Syntaktische Begriffe II: Wortordnung

Die Grammatik einer Sprache legt (meist) auch fest, in welcher Reihenfolge die Konstituenten einer Phrase anzuordnen sind. $x < y$: x immediately precedes y

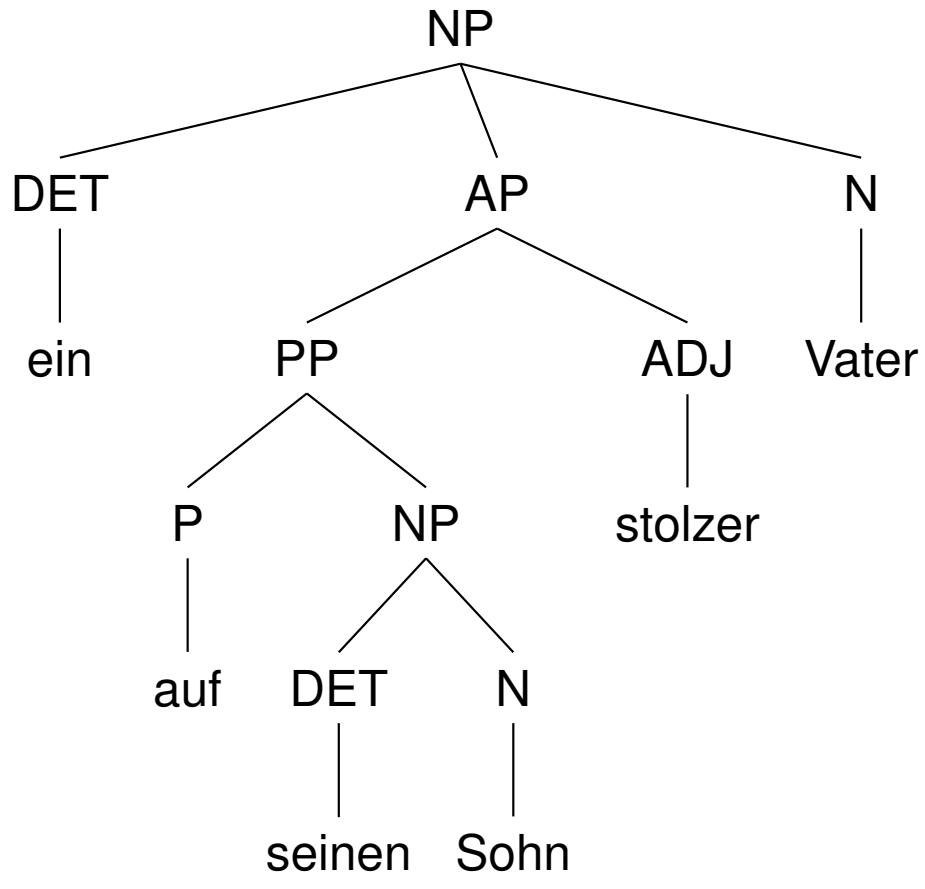
- NP
  - Deutsch: *Art* $<$ *Adj* $<$ *Nomen* $<$ *Relativsatz*
  - Franz.: *Art* $<$ *Adj* $<$ *Nomen* $<$ *Adj* $<$ *Relativsatz*
- PP: *Praeposition* $<$ *NP*
- AP:
  - Deutsch: *PP* $<$ *Adj*. *ein auf seinen Sohn stolzer Vater*
  - Franz.: *Adj* $<$ *PP*. *un homme fier de son fils*

Syntaktische Begriffe: Wortordnung

Phrasenstrukturbäume kodieren sowohl Konstituenz (Dominanz) als auch Präzedenz.

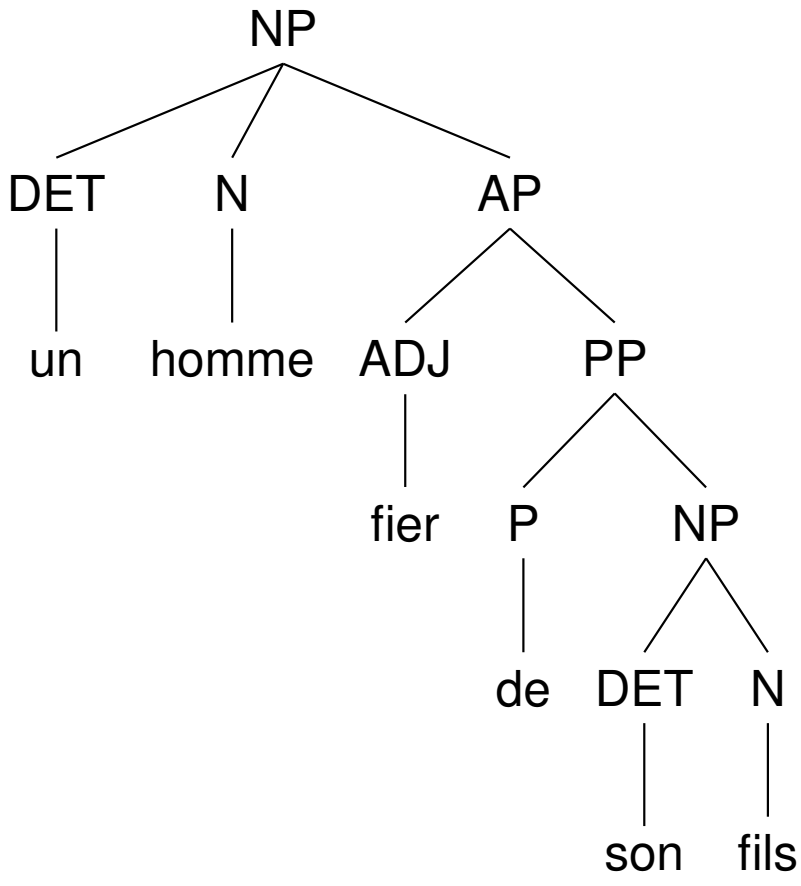
Syntaktische Begriffe: Wortordnung

Phrasenstrukturbäume kodieren sowohl Konstituenz (Dominanz) als auch Präzedenz.



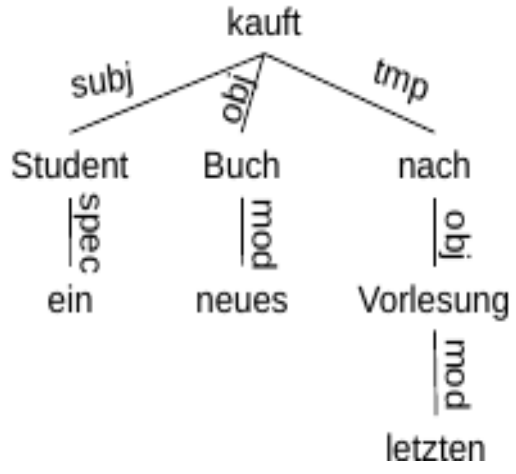
Syntaktische Begriffe: Wortordnung

Phrasenstrukturbäume kodieren sowohl Konstituenz (Dominanz) als auch Präzedenz.



Syntaktische Begriffe III: Dependenz

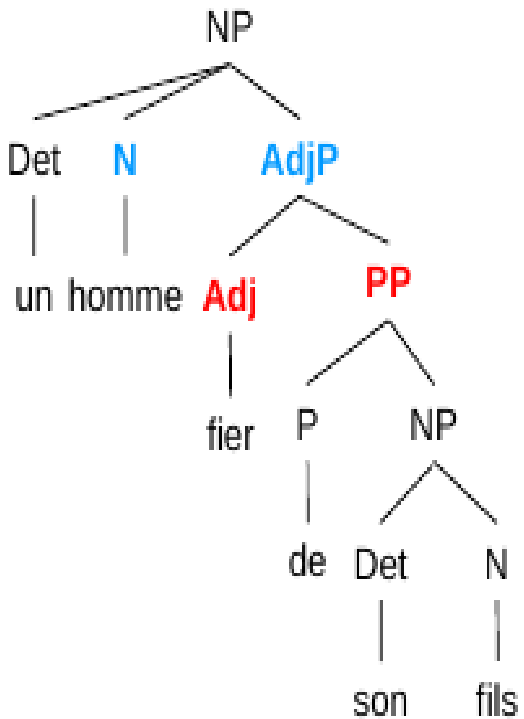
Dependenzbeziehungen bzw grammatische Funktionen über Sprachen hinweg relativ konstant. Prädikate verlangen bestimmte Argumente. Werden durch Adjunkte ergänzt.

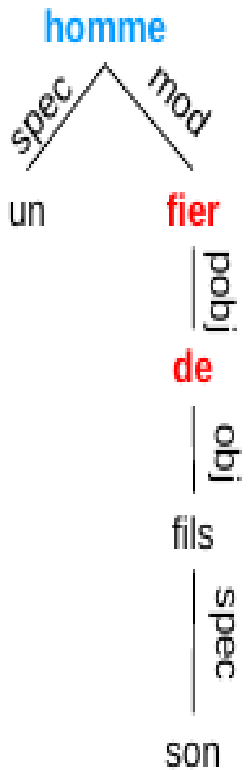


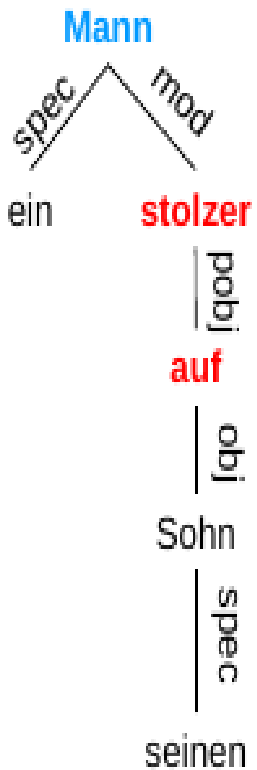
Dependency	Description
subj	syntactic subject
obj	direct object (incl. sentential complements)
dat	indirect object
pcomp	complement of a preposition
comp	predicate nominals (complements of copulas)
tmp	temporal adverbials
loc	location adverbials
attr	premodifying (attributive) nominals (genitives, etc.)
mod	nominal postmodifiers (prepositional phrases, etc.)

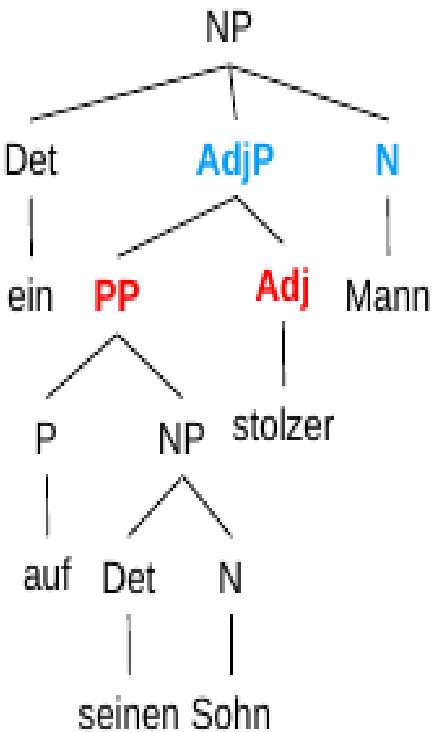
Syntaktische Begriffe III: Dependenz vs. Konstituenz

Dependenzstrukturen kodieren syntaktische Abhängigkeiten, aber keine Konstituenz und keine lineare Abfolge.









Grammatik und Parsing

- Die syntaktischen Strukturen einer Sprache werden durch Grammatik (Phrasenstruktur oder Dependenz) spezifiziert
- Auf Basis der Grammatik kann syntaktische Wohlgeformtheit eines Satzes überprüft werden.
- Diesen Prozess nennt man **Erkennung**.
- Automatische Strukturzuweisung = Parsing

Übersicht

Phrasenstrukturgrammatiken (kontextfreie Grammatiken)

- CFGs (context-free grammars) erlauben eine formale, deklarative Definition syntaktischer Strukturen einer Sprache nach den Strukturprinzipien Konstituenz und Präzedenz.
- Gibt Regeln zur Klammerung einer Sprache vor: NP $\rightarrow$ DT NNS
- Enthält Lexikon mit POS: NNS $\rightarrow \{$women, men, tables$\}$
- **Generative Grammatik:** sollte idealerweise alle grammatischen Sätze einer Sprache generieren und keine der ungrammatischen.

Definition einer kontextfreien Grammatik

$G =< NT, \Sigma, R, S >$, wobei

- NT: endliche Menge von **Nichtterminalsymbolen**.
  Phrasenkategorien NP; VP ... und Wortkategorien; NN, JJ ...
- Σ: endliche Menge von **Terminalsymbole**, d.h. Wörter
- R: endliche Regelmenge der Form $A \rightarrow \alpha$, wobei A ein Nichtterminal und α eine Sequenz von Symbolen $\alpha \in (NT \cup \Sigma)^*$
- S: Startsymbol (aus NT)

$L(G)$: formale Sprache, die durch G erzeugt wird

Beispiel: Kontextfreie Grammatik G_1

$G_1 =< NT_1, \Sigma_1, R_1, S_1 >$

- $NT_1 = \{S, NP, VP, Det, N, V\}$
- $\Sigma_1 = \{der, bellt, Hund, Katze, die, sieht\}$
- R_1 :

$$\begin{aligned} S &\rightarrow NP\ VP \\ VP &\rightarrow V \\ VP &\rightarrow V\ NP \\ NP &\rightarrow Det\ N \end{aligned}$$

$$\begin{aligned} Det &\rightarrow der \\ Det &\rightarrow die \\ N &\rightarrow Katze \\ N &\rightarrow Hund \\ V &\rightarrow sieht \\ V &\rightarrow bellt \end{aligned}$$

Ableitung

Ableitung

Die Menge aller wohlgeformten Sätze einer Sprache ist ableitbar auf Basis der Grammatik G für diese Sprache, ausgehend vom Startsymbol S, durch Anwendung der Ersetzungsregeln der Regelmenge R

Ableitung von *der Hund bellt* aus G_1

Ableitung

Ableitung

Die Menge aller wohlgeformten Sätze einer Sprache ist ableitbar auf Basis der Grammatik G für diese Sprache, ausgehend vom Startsymbol S, durch Anwendung der Ersetzungsregeln der Regelmenge R

Ableitung von *der Hund bellt* aus G_1

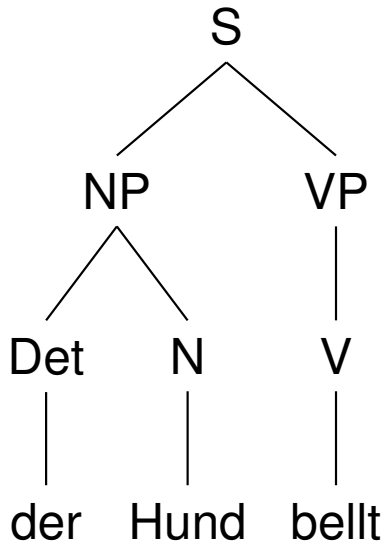
S $\Rightarrow$ NP VP $\Rightarrow$ Det N VP $\Rightarrow$ Det N V $\Rightarrow$ der N V $\Rightarrow$ der Hund V $\Rightarrow$ der Hund bellt

Ableitung

Ableitung

Die Menge aller wohlgeformten Sätze einer Sprache ist ableitbar auf Basis der Grammatik G für diese Sprache, ausgehend vom Startsymbol *S*, durch Anwendung der Ersetzungsregeln der Regelmenge *R*

Ableitung von *der Hund bellt* aus G_1



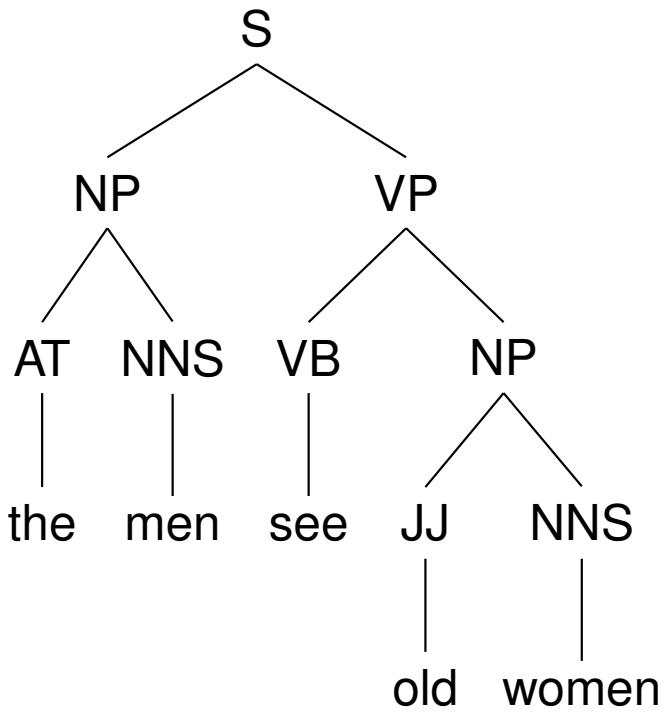
Noch ein Beispiel: G_2

$S \rightarrow NP\ VP$
$NP \rightarrow DT\ NNS$
$NP \rightarrow DT\ NN$
$NP \rightarrow NP\ PP$
$NP \rightarrow NNS$
$NP \rightarrow JJ\ NNS$
$VP \rightarrow VB\ NP$
$VP \rightarrow VBZ\ NP$
$VP \rightarrow VB\ PP$
$PP \rightarrow IN\ NP$

$DT \rightarrow \{$the, a, an, that $\}$
$NNS \rightarrow \{$men, women, tables, hats, feathers, birds $\}$
$NN \rightarrow \{$man, woman, table, flight $\}$
$JJ \rightarrow \{$old, young, pretty $\}$
$VB \rightarrow \{$see, meet, identify, cancel$\}$
$VBZ \rightarrow \{$sees, meets, identifies $\}$
$IN \rightarrow \{$with, for, against, of $\}$

Noch ein Beispiel: G_2

Derive sentences:



(labelled) Bracketing:
[S [NP [AT the] [NNS men]] [VP [VB see] [NP [JJ old] [NNS women]]]]

Üben mit G_2

1. Zeichnen Sie den Baum für *The men with the hats see old women*
2. Erkennungsproblem: Welche der folgenden Sätze wird von Grammatik G_2 erkannt?
    1. *The men with the hats with the feathers see old women*
    2. *The children see old women*
    3. *Old women sees old men*
    4. *See old women*

Üben mit G_2

- Regeln wie NP $\rightarrow$ NP PP sind rekursiv! Wenn man sie mehrfach anwendet, kann man unendlich viele Sätze generieren
- Versuchen Sie, neue Regeln zu schreiben für
  1. Imperative: *See old women*
  2. Konjunktionen von Nominalphrasen: *men and women*
  3. Relativsätze: *a man who sees old women*

Erkennung und Ableitung

- Eine durch eine Grammatik generierte Sprache
  $L(G) = \{\alpha | \alpha \in \Sigma^* \textit{ und } S_G{}^* \Rightarrow \alpha\}$
- Erkennungsproblem: Ist Satz *s* ein Element der durch *G* definierten Sprache?
- G_2 generiert Sätze, die im Englischen ungrammatisch sind (Übergenerierung)
- G_2 generiert viele grammatischen Sätze des Englischen nicht (Untergenerierung)

Problem: Agreement

- **Agreement:** Zwei Konstituenten müssen grammatikalisch in Person, Anzahl, Geschlecht, Kasus übereinstimmen
- Agreementproblem: G_2 *Old men sees old women*
- Wie kann man das in Grammatik integrieren?

$$\text{3SgNP} \rightarrow \text{DT SgNN}$$
$$\text{Non3SgNP} \rightarrow \text{DT PlNN}$$
$$\text{S} \rightarrow \text{3sgNP 3SgVP}$$
$$\text{S} \rightarrow \text{Non3SgNP non3SgVP}$$

Problem: Freie Wortstellung

I sleep on Friday generiert durch S $\rightarrow$ NP VP sowie VP $\rightarrow$ V PP

Was machen wir mit *On Friday, I sleep*

Penn Treebank (1m W Korpus mit Syntaxbäumen in Klammerformat): daraus herauszulesende Grammatik enthält alleine 4 500 Regeln für VP Expandierung.

Kontextfreie Sprachen vs. Reguläre Sprachen

Man kann zeigen, dass alle regulären Sprachen kontextfrei sind.

Aber nicht alle kontextfreien Sprachen sind regulär.
Beispiel, Grammatik G mit

- $NT = \{S, A\}$
- $\Sigma = \{a, b\}$
- R :

$$\begin{aligned} \mathrm{S} &\rightarrow \mathrm{A} \\ \mathrm{A} &\rightarrow \mathrm{a\ A\ b} \\ \mathrm{A} &\rightarrow \mathrm{a\ b} \end{aligned}$$

erzeugt die Sprache $a^n b^n$ mit $n > 1$

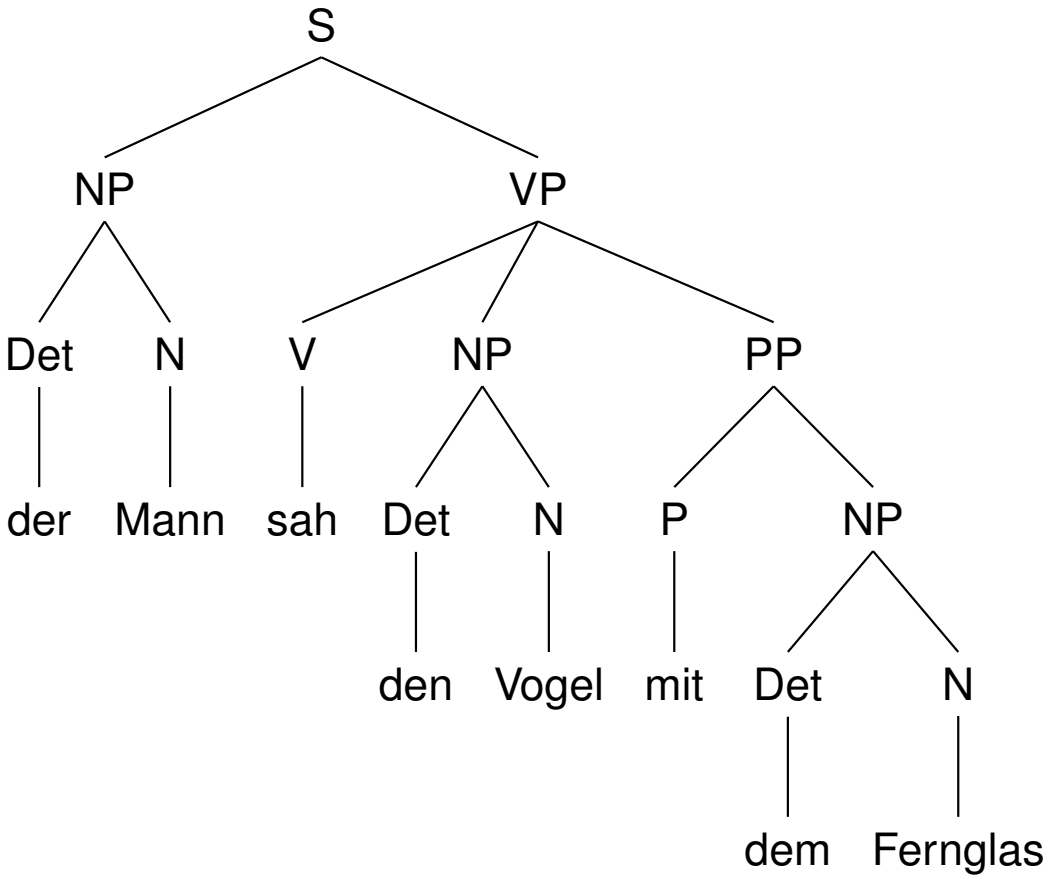
Natürliche Sprachen benötigen mindestens kontextfreie Grammatiken zur Beschreibung (Beispiel: eingebettete Relativsätze)

Übersicht

Ambiguität

Innerhalb einer Grammatik kann es für einen Satz unterschiedliche Ableitungen geben.

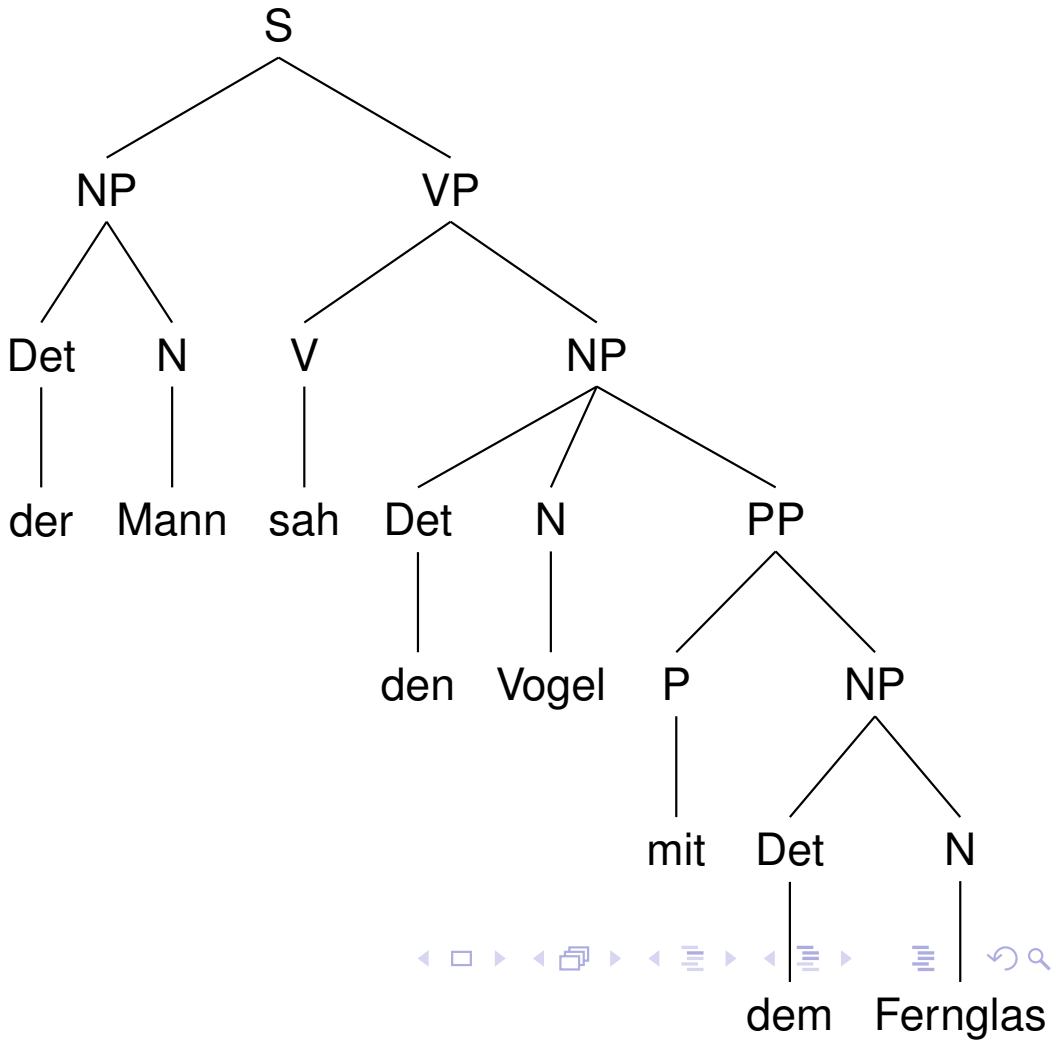
$S \rightarrow NP\ VP$
$NP \rightarrow Det\ N$
$NP \rightarrow Det\ N\ PP$
$VP \rightarrow V\ NP$
$VP \rightarrow V\ NP\ PP$
$PP \rightarrow P\ NP$
$Det \rightarrow \{der, den, dem\}$
$N \rightarrow \{Mann, Vogel, Fernglas\}$
$V \rightarrow sah$
$P \rightarrow mit$



Ambiguität

Innerhalb einer Grammatik kann es für einen Satz unterschiedliche Ableitungen geben.

$S \rightarrow NP\ VP$
$NP \rightarrow Det\ N$
$NP \rightarrow Det\ N\ PP$
$VP \rightarrow V\ NP$
$VP \rightarrow V\ NP\ PP$
$PP \rightarrow P\ NP$
$Det \rightarrow \{der, den, dem\}$
$N \rightarrow \{Mann, Vogel, Fernglas\}$
$V \rightarrow sah$
$P \rightarrow mit$



Ambiguität

- POS Ambiguität: Ein Wort mit verschiedenen Wortarten z.B. *house*
- Lexikalische Ambiguität: Ein Wort (mit einer POS) aber verschiedenen Bedeutungen z.B. *port*
- Syntaktische Ambiguität: Satz mit mehr als einem Parse aufgrund verschiedener Regelanwendungen in der Grammatik (siehe vorherige Folie)

Wie viele Parses für:

1. *old men and women*
2. *I saw men with a telescope*
3. *I saw old men and women with a telescope*

Übersicht

Chomsky Normalform

Chomsky Normalform (CNF)

Eine Grammatik in Chomsky Normalform besteht ausschliesslich aus Regeln der Form $A \rightarrow B\ C$ oder $A \rightarrow \alpha$, wobei A, B, C Nichtterminale (mit B, C nicht Startsymbol) und α ein einzelnes Terminal.

Jede kontextfreie Grammatik kann in eine schwach äquivalente Normalform transformiert werden (d.h. in eine Grammatik mit Normalform, die die gleichen Sätze erkennt/erzeugt).

Umwandlung in Chomsky Normalform

Prozedur zur Umwandlung:

1. Kopiere alle Regeln, die schon in CNF sind
2. Problem 1: Terminale in Regeln wie NP $\rightarrow$ NP und NP. Füge Dummy-Terminale hinzu

$$\text{NP} \rightarrow \text{NP CONJ NP}$$
$$\text{CONJ} \rightarrow \text{und}$$

3. Problem 2: Unit Productions S $\rightarrow$ VP. Folge und expandiere alle unit productions.

$$\text{S} \rightarrow \text{VP}$$
$$\text{VP} \rightarrow \text{V N}$$
$$\text{VP} \rightarrow \text{V}$$
$$\text{V} \rightarrow \text{stehle}$$
$$\text{N} \rightarrow \text{Elefanten}$$

$$\text{S} \rightarrow \text{V N}$$
$$\text{S} \rightarrow \text{stehle}$$
$$\text{V} \rightarrow \text{stehle}$$
$$\text{N} \rightarrow \text{Elefanten}$$

4. Problem 3: Binarisiere, indem ich neue Nichtterminale einfüge.

$$\text{VP} \rightarrow \text{V NP PP}$$

$$\text{VP} \rightarrow \text{V XP1}$$
$$\text{XP1} \rightarrow \text{NP PP}$$

Zusammenfassung

- Man kann Wörter hierarchisch zu Phrasen gruppieren (Konstituenzparsing)
- Man kann Wörter in syntaktische Beziehungen zueinander setzen (Dependenzparsing)
- Kontextfreie Grammatiken definieren formale Sprachen, die mächtiger sind als reguläre Sprachen
- Man sollte können: Konstituenten erkennen, Regeln aufstellen, Bäume zeichnen (aus vorgegebener Grammatik), CNF Konversion, Ambiguitäten darstellen

Lektüre

- Jurafsky und Martin, 3rd online edition Kapitel 12 (nicht 12.6)
- Sag et al (2003). Syntactic Theory: A formal Introduction.
- Van Valin et al (1997): Syntax: Structure, Meaning and Function.