

# Einführung in die Computerlinguistik – Formale Grammatiken

Dozentin: Wiebke Petersen

24.6.2010

# Formale Grammatik

## Definition

Eine *formale Grammatik* ist ein 4-Tupel  $G = (N, T, S, P)$  aus

- einem Alphabet von Terminalsymbolen  $T$  (häufig auch  $\Sigma$ )
- einem Alphabet von Nichtterminalsymbolen  $N$  mit  $N \cap T = \emptyset$
- einem Startsymbol  $S \in N$
- einer Menge von Regeln/Produktionen  
 $P \subseteq \{ \langle \alpha, \beta \rangle \mid \alpha, \beta \in (N \cup T)^* \text{ und } \alpha \notin T^* \}$ .

Für eine Regel  $\langle \alpha, \beta \rangle$  schreiben wir auch  $\alpha \rightarrow \beta$ .

Formale Grammatiken werden auch *Typ0-* oder *allgemeine Regelgrammatiken* genannt.

S	→	NP VP	VP	→	V	NP	→	D N
D	→	the	N	→	cat	V	→	sleeps

Generiert: the cat sleeps

# Terminologie

$$G = \langle \{S, NP, VP, N, V, D, EN\}, \{the, cat, peter, chases\}, S, P \rangle$$
$$P = \left\{ \begin{array}{lll} S & \rightarrow & NP VP \\ NP & \rightarrow & EN \\ EN & \rightarrow & peter \end{array} \quad \begin{array}{lll} VP & \rightarrow & V NP \\ D & \rightarrow & the \\ V & \rightarrow & chases \end{array} \quad \begin{array}{lll} NP & \rightarrow & D N \\ N & \rightarrow & cat \end{array} \right\}$$

# Terminologie

$G = \langle \{S, NP, VP, N, V, D, EN\}, \{the, cat, peter, chases\}, S, P \rangle$

$P = \left\{ \begin{array}{lll} S & \rightarrow & NP VP \\ NP & \rightarrow & EN \\ EN & \rightarrow & peter \end{array} \quad \begin{array}{lll} VP & \rightarrow & V NP \\ D & \rightarrow & the \\ V & \rightarrow & chases \end{array} \quad \begin{array}{lll} NP & \rightarrow & D N \\ N & \rightarrow & cat \end{array} \right\}$

“NP VP” ist **in einem Schritt ableitbar** aus S

# Terminologie

$$G = \langle \{S, NP, VP, N, V, D, EN\}, \{the, cat, peter, chases\}, S, P \rangle$$

$$P = \left\{ \begin{array}{lll} S & \rightarrow & NP VP & VP & \rightarrow & V NP & NP & \rightarrow & D N \\ NP & \rightarrow & EN & D & \rightarrow & the & N & \rightarrow & cat \\ EN & \rightarrow & peter & V & \rightarrow & chases & & & \end{array} \right\}$$

“NP VP” ist **in einem Schritt ableitbar** aus  $S$

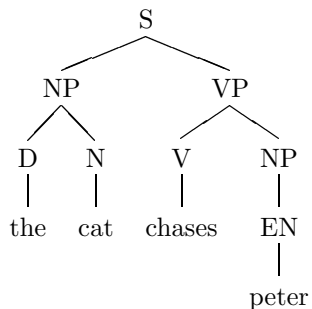
“the cat chases peter” ist **ableitbar** aus  $S$ :

$$\begin{array}{lll} S & \rightarrow & NP VP & \rightarrow & NP V NP & \rightarrow & NP V EN \\ & \rightarrow & NP V peter & \rightarrow & NP chases peter & \rightarrow & D N chases peter \\ & \rightarrow & D cat chases peter & \rightarrow & the cat chases peter & & \end{array}$$

Die Menge aller aus dem Startsymbol  $S$  ableitbarer Wörter ist die von der Grammatik  $G$  **erzeugte Sprache**  $L(G)$ .

$$L(G) = \left\{ \begin{array}{ll} the\ cat\ chases\ peter & peter\ chases\ the\ cat \\ peter\ chases\ peter & the\ cat\ chases\ the\ cat \end{array} \right\}$$

# Ableitungsbaum



# reguläre Sprachen (Typ 3-Sprachen) und rechtslineare Grammatiken

## Definition

Eine Grammatik  $(N, T, S, P)$  heißt **rechtslinear**, wenn alle Regeln/Produktionen die folgende Form haben:

$A \rightarrow a$  oder  $A \rightarrow aB$  wobei  $a \in T \cup \{\epsilon\}$  und  $A, B \in N$ .

Eine durch eine rechtslineare Grammatik erzeugte Sprache heißt **rechtslinear**.

(Vorsicht, häufig werden schärfere [aber äquivalente] Bedingungen gefordert!)

# reguläre Sprachen (Typ 3-Sprachen) und rechtslineare Grammatiken

## Definition

Eine Grammatik  $(N, T, S, P)$  heißt **rechtslinear**, wenn alle Regeln/Produktionen die folgende Form haben:

$A \rightarrow a$  oder  $A \rightarrow aB$  wobei  $a \in T \cup \{\epsilon\}$  und  $A, B \in N$ .

Eine durch eine rechtslineare Grammatik erzeugte Sprache heißt **rechtslinear**.

(Vorsicht, häufig werden schärfere [aber äquivalente] Bedingungen gefordert!)

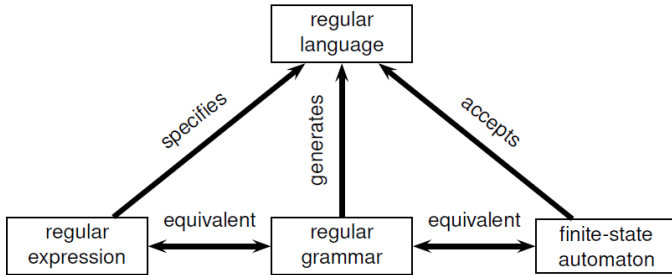
## Theorem

Sei  $L$  eine formale Sprache, dann sind die folgenden Aussagen äquivalent:

- 1  $L$  ist regulär.
- 2 Es gibt eine rechtslineare Grammatik  $G$ , die  $L$  erzeugt.
- 3 Es gibt einen endlichen Automaten  $A$ , der  $L$  akzeptiert.
- 4 Es gibt einen regulären Ausdruck  $R$ , der  $L$  beschreibt.



# Zusammenfassung: reguläre Sprachen



## Definition

Eine Grammatik  $(N, T, S, P)$  heißt **kontextfrei**, wenn alle Regeln/Produktionen die folgende Form haben:

$$A \rightarrow \alpha, \text{ wobei } A \in N \text{ und } \alpha \in (T \cup N)^*.$$

Eine durch eine kontextfreie Grammatik erzeugte Sprache heißt **kontextfrei**.

# kontextfreie Grammatik

## Definition

Eine Grammatik  $(N, T, S, P)$  heißt **kontextfrei**, wenn alle Regeln/Produktionen die folgende Form haben:

$$A \rightarrow \alpha, \text{ wobei } A \in N \text{ und } \alpha \in (T \cup N)^*.$$

Eine durch eine kontextfreie Grammatik erzeugte Sprache heißt **kontextfrei**.

Die Menge der kontextfreien Sprachen ist eine echte Obermenge der Menge der regulären Sprachen

# kontextfreie Grammatik

## Definition

Eine Grammatik  $(N, T, S, P)$  heißt **kontextfrei**, wenn alle Regeln/Produktionen die folgende Form haben:

$$A \rightarrow \alpha, \text{ wobei } A \in N \text{ und } \alpha \in (T \cup N)^*.$$

Eine durch eine kontextfreie Grammatik erzeugte Sprache heißt **kontextfrei**.

Die Menge der kontextfreien Sprachen ist eine echte Obermenge der Menge der regulären Sprachen

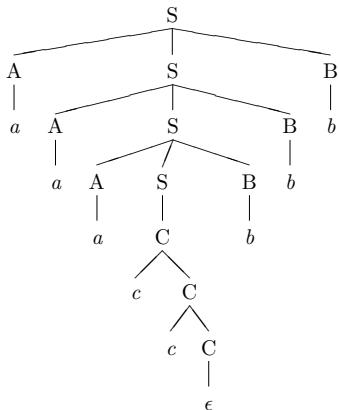
**Beweis:** Jede reguläre Sprache ist per Definition auch kontextfrei und es gibt mindestens eine kontextfreie Sprache, nämlich  $a^n b^n$ , die nicht regulär ist. ( $S \rightarrow aSb, S \rightarrow \epsilon$ )

# Beispiel einer kontextfreien Sprache

$$G = \langle \{S, A, B, C\}, \{a, b, c\}, S, P \rangle$$
$$P = \left\{ \begin{array}{l} S \rightarrow ASB \quad S \rightarrow C \quad S \rightarrow S \\ A \rightarrow a \quad B \rightarrow b \\ C \rightarrow cC \quad C \rightarrow \epsilon \end{array} \right\}$$

# Beispiel einer kontextfreien Sprache

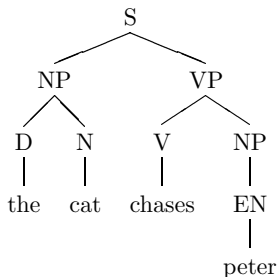
$$G = \langle \{S, A, B, C\}, \{a, b, c\}, S, P \rangle$$
$$P = \left\{ \begin{array}{l} S \rightarrow ASB \quad S \rightarrow C \quad S \rightarrow S \\ A \rightarrow a \quad B \rightarrow b \\ C \rightarrow cC \quad C \rightarrow \epsilon \end{array} \right\}$$



# Linksableitung

Gegeben eine kontextfreie Grammatik G. Eine Ableitung bei der stets das am weitesten links stehende nichtterminale Symbol ersetzt wird, heißt **Linksableitung**

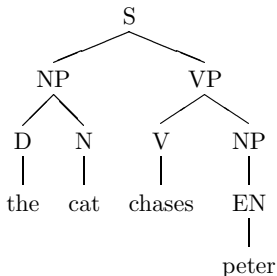
S	→ NP VP	→ D N VP	→ the N VP
	→ the cat VP	→ the cat V NP	→ the cat chases NP
	→ the cat chases EN	→ the cat chases peter	



# Linksableitung

Gegeben eine kontextfreie Grammatik  $G$ . Eine Ableitung bei der stets das am weitesten links stehende nichtterminale Symbol ersetzt wird, heißt **Linksableitung**

$S$	$\rightarrow$ NP VP	$\rightarrow$ D N VP	$\rightarrow$ the N VP
	$\rightarrow$ the cat VP	$\rightarrow$ the cat V NP	$\rightarrow$ the cat chases NP
	$\rightarrow$ the cat chases EN	$\rightarrow$ the cat chases peter	



Zu jeder Linksableitung gibt es genau einen Ableitungsbaum und zu jedem Ableitungsbaum gibt es genau eine Linksableitung.



# ambige Grammatik

Eine Grammatik  $G$  heißt **ambig**, wenn es für ein Wort  $w \in L(G)$  mehr als eine Linksableitung gibt.

$G = (N, T, NP, P)$  mit  $N = \{S, EN, NP, VP, PP, D, N, P\}$ ,

$T = \{\text{Eva, sieht, den, Mann, mit, dem, Fernglas}\}$ ,

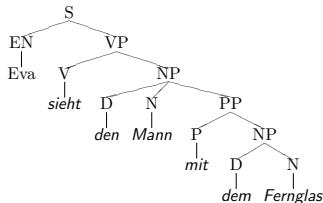
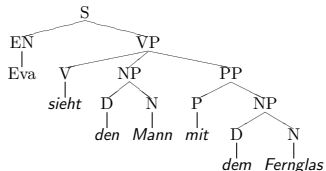
$$P = \left\{ \begin{array}{lll} S \rightarrow EN VP & VP \rightarrow V NP & VP \rightarrow V NP PP \\ NP \rightarrow D N & NP \rightarrow D N PP & PP \rightarrow P NP \\ EN \rightarrow \text{Eva} & P \rightarrow \text{mit} & V \rightarrow \text{sieht} \\ D \rightarrow \text{den} & D \rightarrow \text{dem} & N \rightarrow \text{Mann} \\ N \rightarrow \text{Fernglas} & & \end{array} \right\}$$

# ambige Grammatik

Eine Grammatik  $G$  heißt **ambig**, wenn es für ein Wort  $w \in L(G)$  mehr als eine Linksableitung gibt.

$G = (N, T, NP, P)$  mit  $N = \{S, EN, NP, VP, PP, D, N, P\}$ ,

$T = \{\text{Eva, sieht, den, Mann, mit, dem, Fernglas}\}$ ,

$$P = \left\{ \begin{array}{lll} S \rightarrow EN VP & VP \rightarrow V NP & VP \rightarrow V NP PP \\ NP \rightarrow D N & NP \rightarrow D N PP & PP \rightarrow P NP \\ EN \rightarrow \text{Eva} & P \rightarrow \text{mit} & V \rightarrow \text{sieht} \\ D \rightarrow \text{den} & D \rightarrow \text{dem} & N \rightarrow \text{Mann} \\ N \rightarrow \text{Fernglas} & & \end{array} \right\}$$


# Hausaufgaben (Abgabe 1.7.2010)

- ① Sei  $L$  die Sprache, die aus allen nichtleeren Wörtern über dem Alphabet  $\{a, b\}$  besteht, in denen auf jedes  $a$  unmittelbar ein  $b$  folgt. Beispiele für Wörter dieser Sprache:  $bbbab$ ,  $abababab$ ,  $bb$ ,  $babbbbab$ .
  - geben Sie eine rechtslineare Grammatik  $G$  an, die  $L$  erzeugt und zeichnen Sie den Ableitungsbaum für das Wort  $bbababb$
  - geben Sie einen endlichen Automaten  $A$  an, der  $L$  akzeptiert.
  - geben Sie einen regulären Ausdruck  $R$  an, der  $L$  beschreibt.
- ② Geben sie jeweils eine kontextfreie Grammatik zu den folgenden Sprachen an:
  - ①  $L_1 = \{a^i b^j \mid i < j\}$
  - ②  $L_2 = \{w \in \{a, b\}^* \mid w \text{ ist ein Palindrom}\}$

Wählen Sie pro Sprache ein Wort, das mindestens die Länge 5 hat, und zeichnen Sie den Ableitungsbaum in Bezug auf Ihre Grammatik.