Einführung in die Computerlinguistik – Formale Grammatiken

Dozentin: Wiebke Petersen

22.12.2009

Formale Grammatik

Definition

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

- *einem Alphabet von Terminalsymbolen* T *(häufig auch* Σ*)*
- *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* $\alpha \to \beta$.
Formale Grammatiken werden auch Typ0- oder allgemeine Regelgrammatiken genannt.

$$
\begin{array}{rcl rcl rcl}
\text{S} & \to & \text{NP VP} & \text{VP} & \to & \text{V} & \text{NP} & \to & \text{D N} \\
\text{D} & \to & \text{the} & \text{N} & \to & \text{cat} & \text{V} & \to & \text{sleeps}
\end{array}
$$

Generiert: the cat sleeps

Terminologie

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

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

"NP VP" ist in einem Schritt ableitbar aus S

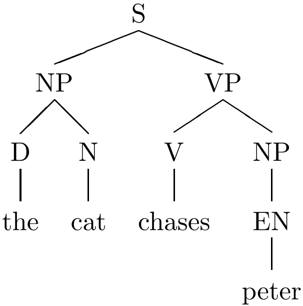
"the cat chases peter" ist ableitbar aus S:

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

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

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

Ableitungsbaum



reguläre Sprachen (Typ 3-Sprachen) und rechtslineare Grammatiken

Definition

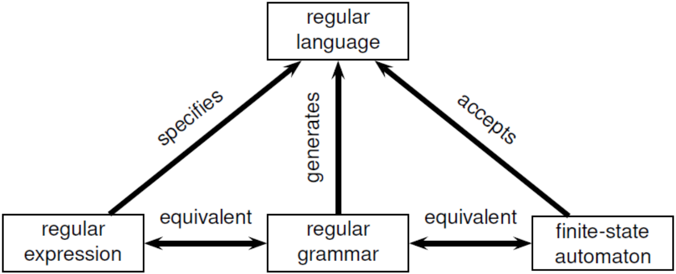
Eine Grammatik (N, T, S, P) heißt rechtslinear*, wenn alle Regeln/Produktionen die folgende Form haben:*
$A \to a$ oder $A \to aB$ wobei $a \in T \cup \{\epsilon\}$ und $A, B \in N$.
Eine durch eine rechtslineare Grammatik erzeugte Sprache heißt rechts- bzw. linkslinear.

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 akzteptiert.*
4. *Es gibt einen regulären Ausdruck R, der L beschreibt.*

Zusammenfassung: reguläre Sprachen



kontextfreie Grammatik

Definition

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

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

Eine durch eine kontextfreie Grammatik erzeugte Sprache heißt kontextfrei.

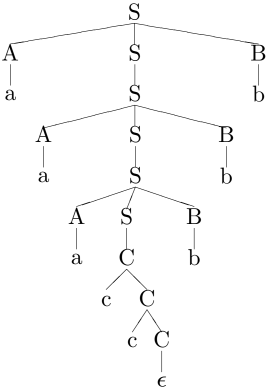
Die Menge der kontextfreien Sprachen ist eine echte Obermenge der Menge der regulären Sprachen

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. $(S \to aSb, S \to \epsilon)$

Beispiel einer kontextfreien Sprache

$$G = \langle \{\mathsf{S}, \mathsf{A}, \mathsf{B}, \mathsf{C}\}, \{a, b, c\}, \mathsf{S}, P \rangle$$

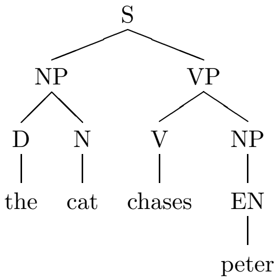
$$P = \left\{ \begin{array}{lllllllll} \mathsf{S} & \rightarrow & \mathsf{ASB} & \mathsf{S} & \rightarrow & \mathsf{C} & \mathsf{S} & \rightarrow & \mathsf{S} \\ \mathsf{A} & \rightarrow & a & \mathsf{B} & \rightarrow & b & & & \\ \mathsf{C} & \rightarrow & c\mathsf{C} & \mathsf{C} & \rightarrow & \epsilon & & & \end{array} \right\}$$



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



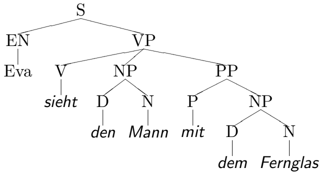
Zu jeder Linksableitung gibt es genau einen Ableitungsbaum und zu jedem Ableitungsbaum gibt es genau eine Linksableitung.

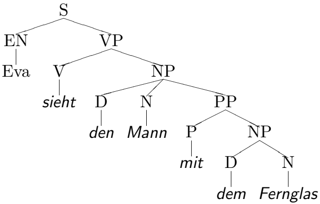
ambige Grammatik

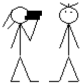
Eine Grammatik G heißt ambig, wenn es für ein Wort $w \in L(G)$ mehr als eine Linksableitung gibt.

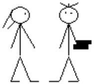
$G = (N, T, \mathrm{NP}, P)$ mit $N = \{\mathrm{S}, \mathrm{EN}, \mathrm{NP}, \mathrm{VP}, \mathrm{PP}, \mathrm{D}, \mathrm{N}, \mathrm{P}\}$,
$T = \{\mathrm{Eva}, \mathrm{sieht}, \mathrm{den}, \mathrm{Mann}, \mathrm{mit}, \mathrm{dem}, \mathrm{Fernglas}\}$,

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









Chomsky-Hierarchie

Eine formale Grammatik (N, T, S, P) ist eine

Typ3 / rechtslineare Grammatik (REG): Regeln die Form
$A \to bB$ oder $A \to b$ mit $A, B \in N$ und $b \in T \cup \{\epsilon\}$

Typ2 / kontextfreie Grammatik (CFG): Regeln der Form
$A \to \beta$ mit $A \in N$ und $\beta \in (N \cup T)^*$.

Typ1 / kontextsensitive Grammatik (CS): Regeln der Form
$\gamma A \delta \to \gamma \beta \delta$ mit $\gamma, \delta, \beta \in (N \cup T)^*, A \in N$ und $\beta \neq \epsilon$;

Typ0 / rekursiv aufzählbere Grammatik (RE): Regeln der Form $\alpha \to \beta$ mit $\alpha, \beta \in (N \cup T)^*$ und $\alpha \notin T^*$

(Vorsicht: aus Platzgründen wurden die Regelbedingungen zum Teil vereinfacht.)

Hausaufgaben (Abgabe 7.1.2010)

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 akzteptiert.
   - geben Sie einen regulären Ausdruck R an, der L beschreibt.
2. Geben sie jeweils eine kontextfreie Grammatik zu den folgenden Sprachen an:
   1. $L_1 = \{a^i b^j | i > j\}$
   2. $L_2 = \{w \in \{a, b\}^* | w \text{ ist ein Palindrom}\}$

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