Kontextfreie Grammatiken

Struktur von Sprache

Wie werden korrekte Sätze einer Sprache gebildet?

Gemäß den Regeln einer Grammatik.

Beispiel für eine kontextfreie Grammatik G

Startsymbol

Satz → S P O

S → A N

P → sieht | frisst

A → der | die | das | den

N → Haus | Hund | Katze

O → A N

Variablen (Nichtterminalsymbole)

Terminalsymbole

Regeln (Produktionen)

Satz ⇒ S P O ⇒ A N P O

⇒ der N P O

⇒ der N sieht O

⇒ der N sieht A N

⇒ der Hund sieht A N

⇒ der Hund sieht A Haus

⇒ der Hund sieht das Haus

Ableitungsschritt

Satzform

Satz

Ableitung

Satz $\overset{*}{\Rightarrow}$ die Katze frisst das Haus

Ableitung in 0 oder mehr Schritten

Satz $\overset{*}{\Rightarrow}$ den Katze sieht das Katze

Satz $\overset{*}{\Rightarrow}$ den Hund frisst die Katze

sollte eigentlich O sein

sollte eigentlich S sein

Variante von G:

Satz → SPO | OPS

Satz ⇒ OPS $\overset{*}{\Rightarrow}$ …

<u>Definition</u> Eine kontextfreie Grammatik $G = (\Sigma, V, P, S)$ besteht aus

- einer Menge Σ von <u>Terminalsymbolen</u>
- einer Menge V von <u>Variablen</u> (Nichtterminalsymbolen, syntaktischen Kategorien) $(V \cap \Sigma = \emptyset)$
- einer Menge P von <u>Regeln</u> (Produktionen)
  der Form $\alpha \to \beta$
  wobei die <u>linke Seite</u> $\alpha \in V$
  und die rechte Seite $\beta \in (V \cup \Sigma)^*$ ist.
- ein <u>Startsymbol</u> $S \in V$

<u>Konventionen</u>: Mehrere Regeln $\alpha \to \beta_1, \alpha \to \beta_2, \ldots$ mit derselben linken Seite werden zusammengefasst: $\alpha \to \beta_1 | \beta_2 | \ldots$
- Variablen: Großbuchstaben, Terminalsymbole: der Rest
- Startsymbol: linke Seite der ersten Regel.

Es reicht aus, die Regeln der Grammatik anzugeben.

Eine Zeichenkette $\alpha \in (\Sigma \cup V)^*$ heißt <u>Satzform</u>
Eine Zeichenkette $\alpha \in \Sigma^*$ heißt <u>Satz</u>

Bei Grammatiken spricht man eher von „Sätzen" $\alpha \in \Sigma^*$ als von „Wörtern" $\alpha \in \Sigma^*$.

Die Elemente von $\Sigma = \{$der, die, Hund, Hans, if, else, $\ldots\}$ sind die „Wörter", aus denen die Sprache aufgebaut ist.

(Bei theoretischen „Spielzeugsprachen" $L \in \{0,1\}^*$ spricht man auch im Zusammenhang von Grammatiken von Wörtern $x \in \{0,1\}^*$)

Definition Ableitung:

Wir schreiben $\alpha \beta \gamma \Rightarrow \alpha \delta \gamma$, (Ableitungsschritt)

wenn es eine Regel $\beta \to \delta \in P$ gibt. $\alpha, \beta, \gamma, \delta \in (\Sigma \cup V)^*$

Wir schreiben $\alpha \overset{*}{\Rightarrow} \omega$, (Ableitung)

wenn es eine Folge $\alpha = \alpha_0 \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \dots \Rightarrow \alpha_k = \omega$ gibt

$k \geq 0$, $\alpha_0, \alpha_1, \dots, \alpha_k \in (\Sigma \cup V)^*$

Ableitung ist <u>nicht</u> deterministisch: Man darf eine beliebige Variable auswählen und eine passende Regel anwenden.

Bei einer <u>Linksableitung</u> wird immer die erste (linkeste) in α_i vorkommende Variable ersetzt.

Die von $G = (\Sigma, V, P, S)$ <u>erzeugte</u> Sprache $L(G)$ ist

$$L(G) := \{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \}$$

Ableitungsbaum (Syntaxbaum, parse tree)

Satz $\Rightarrow$ S P O $\Rightarrow$ A N P O

$\Rightarrow$ der N P O

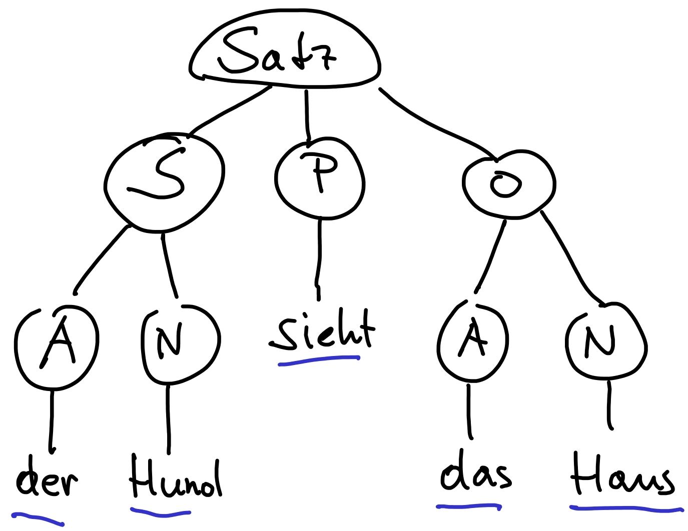
$\Rightarrow$ der N sieht O

$\Rightarrow$ der N sieht A N

$\Rightarrow$ der Hund sieht A N

$\Rightarrow$ der Hund sieht A Haus

$\Rightarrow$ der Hund sieht das Haus



Kontextfreie Sprachen

Eine Sprache L heißt kontextfrei, wenn sie von einer kontextfreien Grammatik erzeugt wird:

$$\exists G: \quad L = L(G)$$

Bsp 2. $S \to 0S1 \mid \varepsilon$

$$S \Rightarrow 0S1 \Rightarrow 00S11 \Rightarrow 000S111 \Rightarrow 000111 \in L(G_2)$$

$$L(G_2) = \{0^n 1^n \mid n \geq 0\}$$ nicht regulär

Bsp 3. $S \to 0S0 \mid 1S1 \mid \varepsilon$

$$S \Rightarrow 0S0 \Rightarrow 00S00 \Rightarrow 001S100 \Rightarrow 001100 \in L(G_3)$$

$$L(G_3) = \{w w^R \mid w \in \{0,1\}^*\}$$ = Palindrome gerader Länge

$00100 \notin L(G_3)$

Bsp 4. $S \to 0T \mid 1T \mid \varepsilon$

$T \to 0S \mid 1S$

$$S \Rightarrow 1T \Rightarrow 10S \Rightarrow 100T \Rightarrow 1000S \Rightarrow 1000 \in L(G_4)$$

$$L(G_4) = \{00, 01, 10, 11\}^*$$ (alle Wörter gerader Länge)

Bsp 5: $S \to 0S1S \mid 1S0S \mid \varepsilon$

$$S \Rightarrow 0S1S \Rightarrow 00S1S1S \Rightarrow 001S1S \Rightarrow 0011S0S1S$$

$$\overset{*}{\Rightarrow} 001101 \in L(G_5)$$

Behauptung: Jede Satzform in der Ableitung hat gleich viele Nullen und Einsen. ✓

$L(G_5) = L_5 := \{ w \in \{0,1\}^* \mid w$ enthält gleich viele Nullen und Einsen. $\}$

$$\boxed{S \to 0S1S \mid 1S0S \mid \varepsilon}$$

Beweis: $L(G_5) \subseteq L_5$ ✓

$L_5 \subseteq L(G_5)$: $\quad w \in L_5$

vollständige Induktion nach $|w|$

Induktionsbasis: $\quad w = \varepsilon \in L(G_5)$

Induktionsschritt: $\quad w \neq \varepsilon$

die erste Position, wo #0 und #1 wieder ausgeglichen sind.

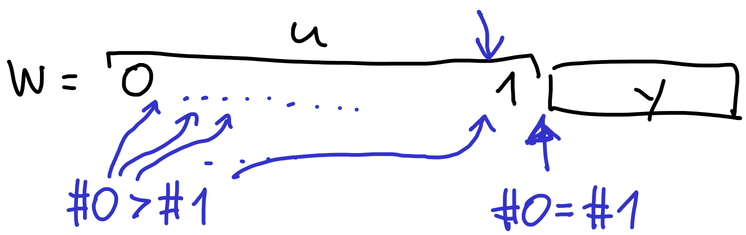
$$w = 0\overbrace{0011001101}^{x}1\underbrace{11110000}_{y}$$

$$S \Rightarrow \underline{0}S1S$$

Zu zeigen: $\quad w = 0x1y \qquad x, y \in L_5$

Fall 1: w fängt mit 0 an.

Suche den kürzesten nichtleeren Präfix u von w, der ausgeglichen ist. (#0 = #1) $\quad$ ($u = w$ ist möglich)



u muss mit einer 1 aufhören.

$w = u.y = \overline{0\,x\,1}\,y$, $\quad x \in L_5$, $y \in L_5$

Nach Induktionsvoraussetzung: $\quad S \overset{*}{\Rightarrow} x$, $S \overset{*}{\Rightarrow} y$

$$S \Rightarrow 0S1S \overset{*}{\Rightarrow} 0x1S \overset{*}{\Rightarrow} 0x1y = w \quad \in L(G_5)$$

Fall 2: analog

Variante: $\quad S \to 0S1 \mid 1S0 \mid SS \mid \varepsilon$