Kapitel 6
Kontextfreie Sprachen

Abschnitt 6.1
Kontextfreie Grammatiken

Die Syntax komplexer formaler Sprachen wie Programmiersprachen oder Spezifikationssprachen ist oft induktiv definiert.

Solche induktiven Definitionen lassen sich kompakt und präzise durch Grammatiken beschreiben. In der Regel handelt es sich dabei um kontextfreie Grammatiken.

Backus-Naur Form

In der Praxis werden kontextfreie Grammatiken oft in Backus-Naur Form (BNF) angegeben.

Beispiel 6.1

Folgende Grammatik in BNF beschreibt unsere induktiv definierten Mengen `BIN` der Binärzahlen (vgl. Beispiel 3.6) und und `ART` der arithmetischen Terme (vgl. Beispiel 3.20):

```
<digit>   ::= "0" | "1"
<bstring> ::= "" | <digit> <bstring>
<bin>     ::= "0" | "1" <bstring>
<op2>     ::= "+" | "*" | "-"
<art>     ::= <bin>
            | "-" <art>
            | "(" <art> <op2> <art> ")"
```

Definition 6.2

Eine kontextfreie Grammatik ist ein Quadrupel

$$(N, \Sigma, P, S),$$

bestehend aus

- einer endlichen Menge N, deren Elemente wir als Nichtterminalsymbole bezeichnen,
- einer zu N disjunkten endlichen Menge Σ, dem Terminalalphabet, dessen Elemente wir als Terminalsymbole bezeichnen,
- einer endlichen Menge P von Regeln (oder Produktionen) der Form $A \to \alpha$ mit $A \in N$ und $\alpha \in (N \cup \Sigma)^*$,
- einem Startsymbol $S \in N$.

Beispiel 6.3

Wir betrachten noch einmal die Grammatik für die Menge $\texttt{ART} \subseteq \Sigma^*$, die in BNF wie folgt aussah (vgl. Beispiel 6.1):

```
<digit>   ::= "0" | "1"
<bstring> ::= "" | <digit> <bstring>
<bin>     ::= "0" | "1" <bstring>
<op2>     ::= "+" | "*" | "-"
<art>     ::= <bin> | "-" <art>
            | "(" <art> <op2> <art> ")"
```

In unserem Formalismus entspricht sie der Grammatik

$$\mathcal{G}_{\texttt{ART}} = (N, \Sigma_{\text{ASCII}}, P, A)$$

mit

- $N = \{A, B, C, D, O\}$,
- $P = \{ D \to 0, \quad D \to 1, \quad C \to \varepsilon, \quad C \to DC,$
  $B \to 0, \quad B \to 1C, \quad O \to +, \quad O \to *,$
  $O \to -, \quad A \to B, \quad A \to -A, \quad A \to (AOA) \quad \}$

- Grammatiken bezeichnen wir mit $\mathcal{G}$ und Varianten wie $\mathcal{G}'$.
- Großbuchstaben $A, B, \ldots$, die wir zur besseren Lesbarkeit farbig darstellen, bezeichnen Nichterminalsymbole.
- Kleinbuchstaben $a, b, \ldots$ bezeichnen Terminalsymbole.
- Kleinbuchstaben $u, v, \ldots$ bezeichnen Wörter in Σ^*.
  Wir bezeichnen solche Wörter als Terminalwörter.
- Griechische Buchstaben $\alpha, \beta, \ldots$ bezeichnen Wörter in $(N \cup \Sigma)^*$.
  Wir bezeichnen solche Wörter als Satzformen.

Wenn Terminal- und Nichtterminalsymbole klar sind, genügt die Auflistung der Regeln

- beginnend mit einer Regel, die links das Startsymbol hat,
- zusammengefasst nach gleichen linken Seiten, dann rechte Seiten durch „|“ getrennt

Beispiel 6.3 (Forts.)

Die Grammatik $\mathcal{G}_{\text{ART}}$ lässt sich damit wie folgt darstellen:

$$
\begin{aligned}
A &\to B \mid -A \mid (AOA) \\
O &\to + \mid * \mid - \\
B &\to 0 \mid 1C \\
C &\to \varepsilon \mid DC \\
D &\to 0 \mid 1
\end{aligned}
$$

Definition 6.4

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik, und seien $\alpha, \beta \in (N \cup \Sigma)^*$ Satzformen.

1. β ist direkt herleitbar aus α (kurz: $\alpha \to_{\mathcal{G}} \beta$), wenn es eine Regel $A \to \delta$ in P und Satzformen $\gamma_1, \gamma_2 \in (N \cup \Sigma)^*$ gibt, so dass

$$\alpha = \gamma_1 A \gamma_2 \quad \text{und} \quad \beta = \gamma_1 \delta \gamma_2.$$

   Intuitiv entsteht also β aus α, indem man ein A durch δ ersetzt.
2. Eine Ableitung von β aus α ist eine Folge $(\alpha_0, \ldots, \alpha_n)$ von Satzformen, so dass $\alpha_0 = \alpha$, $\alpha_n = \beta$ und

$$\alpha_{i-1} \to_{\mathcal{G}} \alpha_i \qquad \text{für } 1 \leq i \leq n.$$

3. Eine Ableitung von β ist eine Ableitung von β aus S.
4. β ist ableitbar aus α in $\mathcal{G}$ (wir schreiben $\alpha \overset{*}{\to}_{\mathcal{G}} \beta$), wenn es eine Ableitung von β aus α gibt.
5. β ist ableitbar in $\mathcal{G}$, wenn es eine Ableitung von β gibt.

Definition 6.5

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik.

Die von $\mathcal{G}$ erzeugte Sprache ist

$$L(\mathcal{G}) = \{w \in \Sigma^* \mid S \overset{*}{\to}_{\mathcal{G}} w\},$$

also die Menge aller in $\mathcal{G}$ ableitbaren Terminalwörter.

Terminologie und Notation

Sei $\mathcal{G}$ eine kontextfreie Grammatik.

- Die Länge einer Ableitung $(\alpha_0, \ldots, \alpha_n)$ in $\mathcal{G}$ ist n.
  Wir schreiben $\alpha \overset{n}{\to}_{\mathcal{G}} \beta$, wenn es eine Ableitung der Länge n von β aus α gibt und $\alpha \overset{\leq n}{\longrightarrow}_{\mathcal{G}} \beta$, wenn es eine Ableitung der Länge höchstens n von β aus α gibt.
- In den Notationen $\to_{\mathcal{G}}$, $\overset{n}{\to}_{\mathcal{G}}$, $\overset{\leq n}{\longrightarrow}_{\mathcal{G}}$ und $\overset{*}{\to}_{\mathcal{G}}$ lassen wir den Index $\mathcal{G}$ weg, wenn die Grammatik $\mathcal{G}$ aus dem Kontext hervorgeht.

Definition 6.6

Eine Sprache ist kontextfrei, wenn sie von einer kontextfreien Grammatik erzeugt wird.

Beispiel 6.7

Sei $L_1 = \{a^n b^n \mid n \geq 0\}$.
L_1 wird von folgender Grammatik $\mathcal{G}_1$ erzeugt und ist damit kontextfrei:

$$S \to aSb, \qquad S \to \varepsilon.$$

Die Ableitungen einiger Wörter in $\mathcal{G}_1$ sind:

$$S \to \varepsilon, \qquad S \to aSb \to a\varepsilon b = ab,$$
$$S \to aSb \to aaSbb \to aaaSbbb \to aaabbb.$$

Mit Hilfe des Pumpinglemmas lässt sich leicht zeigen, dass L_1 nicht regulär ist.

Beispiel 6.3 (Forts.)

Die Sprache ART der arithmetischen Terme ist kontextfrei. Sie wird erzeugt von der Grammatik

$$A \to B \mid -A \mid (AOA) \qquad O \to + \mid * \mid -$$
$$B \to 0 \mid 1C \qquad C \to \varepsilon \mid DC$$
$$D \to 0 \mid 1$$

Eine Ableitung des Terms -(10*(0+1)) ist

$$\underline{A} \to -\underline{A} \to -(\underline{A}OA) \to -(\underline{B}OA) \to -(1\underline{C}OA) \to -(1\underline{D}COA)$$
$$\to -(10\underline{C}OA) \to -(10\underline{O}A) \to -(10*\underline{A}) \to -(10*(\underline{A}OA))$$
$$\to -(10*(\underline{B}OA)) \to -(10*(0\underline{O}A)) \to -(10*(0+\underline{A}))$$
$$\to -(10*(0+\underline{B})) \to -(10*(0+1\underline{C})) \to -(10*(0+1))$$

Hier ist jeweils das Nichtterminalsymbol, das als nächstes ersetzt wird, unterstrichen.

Beispiel 6.8

Ein Palindrom ist ein Wort, das von vorne und hinten gelesen gleich ist.

Wir betrachten die Sprache aller Palindrome über dem Alphabet $\{a, b\}$:

$$L_P := \{a_1 \dots a_n \in \{a, b\}^* \mid i \in \mathbb{N}, a_i = a_{n+1-i} \text{ für } 1 \leq i \leq n\}.$$

L_P wird von folgender kontextfreier Grammatik $\mathcal{G}_P$ erzeugt und ist damit kontextfrei:

$$S \to aSa \mid bSb \mid a \mid b \mid \varepsilon.$$

Zwei Lemmata

Lemma 6.9 (Kombinationslemma)

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik. Seien $\alpha, \beta, \gamma, \delta \in (N \cup \Sigma)^*$ und $A \in N$, so dass

$$\alpha \xrightarrow{*} \beta A \delta \quad \text{und} \quad A \xrightarrow{*} \gamma.$$

Dann gilt $\alpha \xrightarrow{*} \beta\gamma\delta$.

Lemma 6.10 (Zerlegungslemma)

Sei $n \in \mathbb{N}$. Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik. Seien $\alpha_1, \alpha_2, \beta, \in (N \cup \Sigma)^*$ Satzformen, so dass

$$\alpha_1\alpha_2 \xrightarrow{n} \beta$$

Dann gibt es eine Zerlegung $\beta = \beta_1\beta_2$, so dass

$$\alpha_1 \xrightarrow{\leq n} \beta_1 \quad \text{und} \quad \alpha_2 \xrightarrow{\leq n} \beta_2.$$

Beweise I

Beweis des Kombinationslemmas. Sei $(\gamma_0 = A, \gamma_1, \ldots, \gamma_m = \gamma)$ eine Ableitung von γ aus A. Dann ist

$$(\beta\gamma_0\delta, \beta\gamma_1\delta, \ldots, \beta\gamma_m\delta)$$

eine Ableitung von $\beta\gamma\delta$ aus $\beta A\delta$.
Sei $(\alpha_0 = \alpha, \alpha_1, \ldots, \alpha_\ell = \beta A\delta)$ eine Ableitung von $\beta A\delta$ aus α. Dann ist

$$(\alpha_0 = \alpha, \alpha_1 \ldots, \alpha_\ell = \beta\gamma_0\delta, \beta\gamma_1\delta, \ldots, \beta\gamma_m\delta)$$

eine Ableitung von $\beta\gamma\delta$ aus α. □

Beweis des Zerlegungslemmas. Induktion über die Länge n der Ableitung.

Induktionsanfang: $n = 0$
Gelte $\alpha_1\alpha_2 \xrightarrow{0} \beta$. Dann gilt $\beta = \alpha_1\alpha_2$, und wir setzen $\beta_1 := \alpha_1$ und $\beta_2 := \alpha_2$.

Beweise II

Induktionsschritt: $n = n + 1$
Gelte $\alpha_1\alpha_2 \xrightarrow{n+1} \beta$. Dann gibt es ein $\beta' \in (N \cup \Sigma)^*$, so dass

$$\alpha_1\alpha_2 \xrightarrow{n} \beta' \to \beta$$

Nach Induktionsannahme gibt es eine Zerlegung $\beta' = \beta'_1\beta'_2$, so dass

$$\alpha_1 \xrightarrow{\leq n} \beta'_1 \quad \text{und} \quad \alpha_2 \xrightarrow{\leq n} \beta'_2.$$

Betrachten wir den letzten Schritt der Ableitung, $\beta' \to \beta$. Dann gibt es eine Zerlegung $\beta' = \gamma_1 A\gamma_2$ und eine Regel $A \to \delta \in P$, so dass $\beta = \gamma_1\delta\gamma_2$.

Wir unterscheiden zwei Fälle: Das A in $\gamma_1 A\gamma_2$ gehört zu β'_1, und das A in $\gamma_1 A\gamma_2$ gehört zu β'_2.

Fall 1: Das A in $\gamma_1 A\gamma_2$ gehört zu β'_1.
Dann gibt es eine Zerlegung $\gamma_2 = \gamma_{21}\gamma_{22}$, so dass

$$\beta'_1 = \gamma_1 A\gamma_{21} \quad \text{und} \quad \beta'_2 = \gamma_{22}.$$

Es gilt $\beta = \gamma_1\delta\gamma_{21}\gamma_{22}$ und $\beta_1' \to \gamma_1\delta\gamma_{21}$. Wir setzen

$$\beta_1 := \gamma_1\delta\gamma_{21} \quad \text{und} \quad \beta_2 := \gamma_{22}$$

Dann gilt $\beta = \beta_1\beta_2$ und $\alpha_1 \xrightarrow{\leq n} \beta_1' \to \beta_1$, also $\alpha_1 \xrightarrow{\leq n+1} \beta_1$ und $\alpha_2 \xrightarrow{\leq n} \beta_2' = \beta_2$.

Fall 2: Das A in $\gamma_1 A\gamma_2$ gehört zu β_2'.
Dann gibt es eine Zerlegung $\gamma_1 = \gamma_{11}\gamma_{12}$, so dass

$$\beta_1' = \gamma_{11} \quad \text{und} \quad \beta_2' = \gamma_{12}A\gamma_2$$

Es gilt $\beta = \gamma_{11}\gamma_{12}\delta\gamma_2$ und $\beta_2' \to \gamma_{12}\delta\gamma_2$. Wir setzen

$$\beta_1 := \gamma_{11} \quad \text{und} \quad \beta_2 := \gamma_{12}\delta\gamma_2$$

Dann gilt $\beta = \beta_1\beta_2$ und $\alpha_1 \xrightarrow{\leq n} \beta_1' = \beta_1$ und $\alpha_2 \xrightarrow{\leq n} \beta_2' \to \beta_2$, also $\alpha_2 \xrightarrow{\leq n+1} \beta_2$. □

Beispiel 6.11 I

Die Sprache $L_K \subseteq \{(,)\}$ aller korrekten Klammerausdrücke ist kontextfrei. Sie wird von folgender Grammatik $\mathcal{G}_k$ erzeugt:

$$S \to (A) \mid \varepsilon$$
$$A \to S \mid AA$$

Behauptung

$L_K = L(\mathcal{G}_K)$

Beweis.

Beispiel 6.11 II

$L_K \subseteq L(\mathcal{G}_k)$: Wir zeigen induktiv über den Aufbau von $w \in L_K$, dass $w \in L(\mathcal{G}_k)$.

Induktionsanfang: $S \to \varepsilon$ ist eine Ableitung von ε in $\mathcal{G}_k$, also gilt $\varepsilon \in L(\mathcal{G}_k)$.

Induktionsschritt: Seien $w_1, \ldots, w_k \in L_K$ und $w = (w_1 \ldots w_k)$.
Nach Induktionsannahme gilt für $1 \leq i \leq k$

$$S \stackrel{*}{\to} w_i.$$

Außerdem ist $(\overbrace{S \ldots S}^{k \text{ mal}})$ ableitbar:

$$\underline{S} \to (\underline{A}) \to (A\underline{A}) \to (AA\underline{A}) \to \ldots \to (\underline{A}A \ldots A)$$
$$\to (S\underline{A} \ldots A) \to \ldots \to (S \ldots S)$$

Jetzt ergibt k-maliges Anwenden des Kombinationslemmas eine Ableitung von w.

Beispiel 6.11 III

$L(\mathcal{G}_k) \subseteq L_K$: Wir zeigen per Induktion über die Länge n einer Ableitung, für alle Terminalwörter $w \in \Sigma^*$:

(i) Wenn $S \stackrel{n}{\to} w$, dann $w \in L_K$.

(ii) Wenn $A \stackrel{n}{\to} w$, dann $w \in L_K^*$.

Induktionsanfang: $n = 1$

(i) Wenn $S \to w$ für ein $w \in \Sigma^*$, so $w = \varepsilon \in L_K$.

(ii) $A \to w \in \Sigma^*$ ist nicht möglich.

Induktionsschritt: $1, \ldots, n \to n + 1$.

Beispiel 6.11 IV

(i) Gelte $S \xrightarrow{n+1} w$. Dann gilt $S \to (A) \xrightarrow{n} w$. Nach dem Zerlegungslemma (zweimal angewandt) gibt es eine Zerlegung $w = xyz$, so dass

$$(\xrightarrow{\leq n} x \quad \text{und} \quad A \xrightarrow{\leq n} y \quad \text{und} \quad) \xrightarrow{\leq n} z$$

Weil (und) keine Nichtterminale enthalten, gilt $x = (\,$ und $z =)$.
Nach Induktionsanahme gilt $y \in L_K^*$, also $y = y_1 \dots y_k$ für ein $k \geq 1$ und $y_1, \dots, y_k \in L_K$. Hier können wir annehmen, dass $k \geq 1$, weil $\varepsilon \in L_K$.
Insgesamt gilt also

$$w = (y_1 \dots y_k)$$

für ein $k \geq 1$ und $y_1, \dots, y_k \in L_K$.
Weil L_K abgeschlossen ist unter der rekursiven Regel

Beispiel 6.11 V

„Wenn $w_1, \dots, w_k \in L_K$, dann $(w_1 \dots w_k)$."

gilt damit $w \in L_K$.

(ii) Gelte $A \xrightarrow{n+1} w$. Dann gilt entweder $A \to S \xrightarrow{n} w$ oder $A \to AA \xrightarrow{n} w$.

Falls $S \xrightarrow{n} w$, so $w \in L_K$ nach Induktionsannahme.

Nehmen wir also an, dass $AA \xrightarrow{n} w$. Nach dem Zerlegungslemma gibt es eine Zerlegung $w = xy$, so dass

$$A \xrightarrow{\leq n} x \quad \text{und} \quad A \xrightarrow{\leq n} y.$$

Nach Induktionsannahme gilt dann $x \in L_K^*$ und $y \in L_K^*$ und damit $w = xy \in L_K^*$.

□

Satz 6.12

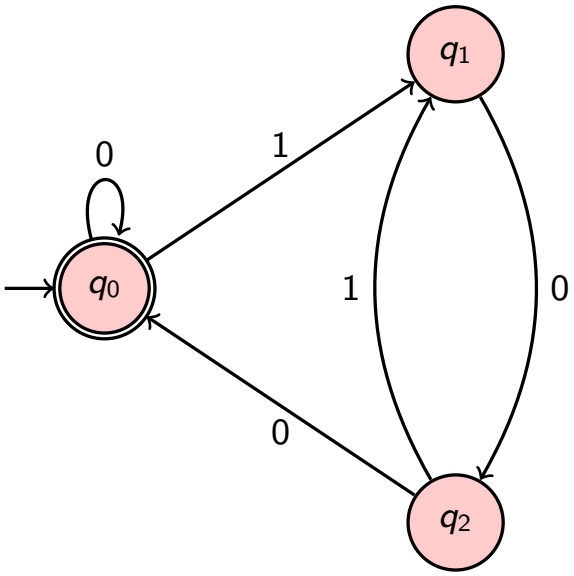
Alle regulären Sprachen sind kontextfrei.

Bemerkung 6.13

Die Umkehrung dieses Satzes gilt nicht. Wir haben bereits einige Beispiele von nichtregulären Sprachen gesehen, die kontextfrei sind.

Beispiel 6.14 I

Sei $L = L(\mathcal{A})$ für folgenden NFA $\mathcal{A}$.



Beispiel 6.14 II

Wir wollen eine kontextfreie Grammatik angeben, die L erzeugt.

Wir notieren dazu Läufe des NFAs durch Präfixe des gelesenen Wortes und die jeweils erreichten Zustände.

Beispiel: Für den Lauf

$$(q_0, 1, q_1, 0, q_2, 1, q_1, 0, q_2, 0, q_0)$$

schreiben wir

$$q_0 \to 1q_1 \to 10q_2 \to 101q_1 \to 1010q_2 \to 10100q_0 \to 10100 \qquad (\star)$$

Jetzt konstruieren wir eine Grammatik, deren Nichterminalsymbole die Zustände des NFAs sind, so dass $(\star)$ eine Ableitung dieser Grammatik ist:

$$\begin{aligned} q_0 &\to 0q_0 \mid 1q_1 \mid \varepsilon \\ q_1 &\to 0q_2 \\ q_2 &\to 0q_0 \mid 1q_1. \end{aligned}$$

Allgemeine Konstruktion

Definition 6.15

Für jeden NFA $\mathcal{A} = (Q, \Sigma, \Delta, q_0, F)$ sei

$$\mathcal{G}_\mathcal{A} = (Q, \Sigma, P, q_0)$$

die kontextfreie Grammatik mit

$$P := \big\{q \to ar \;\big|\; (q, a, r) \in \Delta\big\} \cup \big\{q \to \varepsilon \;\big|\; q \in F\big\}.$$

Lemma 6.16

Für jeden NFA $\mathcal{A}$ gilt $L(\mathcal{A}) = L(\mathcal{G}_\mathcal{A})$.

Satz 6.12 folgt sofort aus dem Lemma.

Beweis von Lemma 6.16 I

$\mathcal{A} = (Q, \Sigma, \Delta, q_0, F)$

$\mathcal{G}_\mathcal{A}$:

$$q \to ar \qquad \text{für alle } (q, a, r) \in \Delta,$$
$$q \to \varepsilon \qquad \text{für alle } q \in F.$$

Behauptung

$$L(\mathcal{A}) = L(\mathcal{G}_\mathcal{A})$$

Beweis.

Beweis von Lemma 6.16 II

„$\subseteq$“ Für jeden akzeptierenden Lauf

$$(q_0, a_1, q_1, \ldots, a_n, q_n)$$

auf einem Wort $w = a_1 \ldots a_n \in L(\mathcal{A})$ ist

$$q_0 \to a_1 q_1 \to \ldots \to a_1 \ldots a_i q_i \to \ldots \to a_1 \ldots a_n q_n \to a_1 \ldots a_n$$

eine Ableitung von w in $\mathcal{G}_\mathcal{A}$. Also $w \in L(\mathcal{G}_\mathcal{A})$.

„$\supseteq$“ Wir beweisen per Induktion über $n \geq 0$, dass für jede Satzform $\alpha \in (Q \cup \Sigma)^*$ mit

$$q_0 \xrightarrow{n} \alpha$$

gilt:

(i) entweder $\alpha = wq$ für ein $w \in \Sigma^*$ und ein $q \in Q$, so dass $\mathcal{A} : q_0 \xrightarrow{w} q$,

(ii) oder $\alpha = w$ für ein $w \in L(\mathcal{A})$.

Beweis von Lemma 6.16 III

Für alle ableitbaren Terminalwörter $w \in \Sigma^*$ sind wir also im Fall (ii), und es gilt $w \in L(\mathcal{A})$. Also $L(\mathcal{G}_\mathcal{A}) \subseteq L(\mathcal{A})$.

Induktionsanfang: $n = 0$.
$q_0 \xrightarrow{0} \alpha$ impliziert $\alpha = q_0$, und weil $\mathcal{A} : q_0 \xrightarrow{\varepsilon} q_0$ gilt (i).

Induktionsschritt: $n \to n + 1$.
Gelte $q_0 \xrightarrow{n+1} \alpha$. Dann gibt es ein $\alpha' \in (Q \cup \Sigma)^*$ mit

$$q_0 \xrightarrow{n} \alpha' \to \alpha$$

Nach Induktionsannahme erfüllt α' (i) oder (ii), und weil $\alpha' \to \alpha$ muss α' mindestens ein Nichtterminalsymbol enthalten, erfüllt also (i). Das heißt,

$$\alpha' = wq$$

für ein $w \in \Sigma^*$ und ein $q \in Q$, so dass $\mathcal{A} : q_0 \xrightarrow{w} q$. Die Regel im letzten Schritt $\alpha' \to \alpha$ der Ableitung ist

Beweis von Lemma 6.16 IV

- entweder $q \to ar$ für ein Tripel $(q, a, r) \in \Delta$
- oder $q \to \varepsilon$, falls $q \in F$.

Im ersten Fall ist $\alpha = war$, und es gilt

$$\mathcal{A} : q_0 \xrightarrow{w} q \xrightarrow{a} r.$$

Also ist (i) erfüllt.
Im zweiten Fall ist $\alpha = w$ und $w \in L(\mathcal{A})$, weil $\mathcal{A} : q_0 \xrightarrow{w} q \in F$. Also ist (ii) erfüllt.

□

Abschnitt 6.2

Normalformen

Ziel

In vielen Anwendungen ist es hilfreich, nur Grammatiken in einer speziellen Form (einer Normalform) zu betrachten.

Chomsky-Normalform: Alle Regeln haben die Form

$$A \to BC \quad \text{oder} \quad A \to a.$$

Greibach-Normalform: Alle Regeln haben die Form

$$A \to aB_1 \dots B_k \quad \text{für ein } k \geq 0.$$

Rechtslinear: Alle Regeln haben die Form

$$A \to aB \quad \text{oder} \quad A \to a.$$

In diesem Kapitel wollen wir untersuchen, inwieweit sich beliebige kontextfreie Grammatiken in äquivalente Grammatiken dieser Formen umwandeln lassen.

Um das leere Wort zu erzeugen, muss eine Grammatik mindestens eine Regel $A \to \varepsilon$ (eine ε-Regel) enthalten.

Das bedeutet, dass keine Grammatik in einer der auf der letzten Seite aufgeführten Formen das leere Wort erzeugen kann.

Der Einfachheit halber berücksichtigen wir in diesem Kapitel das leere Wort in den Sprachen nicht.

Aber:
Jede Grammatik $G = (N, \Sigma, R, S)$ lässt sich durch

- Hinzunahme eines neuen Startsymbols S',
- Hinzunahme der Regeln $S' \to \varepsilon \mid S$

in eine Grammatik $\mathcal{G}' = (N, \Sigma, P', S')$ mit $L(\mathcal{G}') = L(\mathcal{G}) \cup \{\varepsilon\}$ transformieren.

Elimination von ε-Regeln

Lemma 6.17

Zu jeder Grammatik $\mathcal{G}$ gibt es eine Grammatik $\mathcal{G}'$ ohne ε-Regeln, so dass $L(\mathcal{G}') = L(\mathcal{G}) \setminus \{\varepsilon\}$.

Beweisskizze.

Sei $\mathcal{G} = (N, \Sigma, P, S)$.

Sei P' die Menge von Regeln, die aus P durch wiederholtes Anwenden des folgenden Erweiterungsschrittes entsteht (bis keine Erweiterung mehr möglich ist).

Erweiterungsschritt: Sind $A \to \varepsilon$ und $B \to \alpha A \beta$ Regeln, so füge die Regel $B \to \alpha\beta$ hinzu.

Anschließend entfernen wir alle ε-Regeln. Sei P' die resultierende Regelmenge und $\mathcal{G}' = (N, \Sigma, P', S)$.

Es ist nicht schwer (wenn auch ein wenig mühsam), zu zeigen, dass $L(\mathcal{G}') = L(\mathcal{G}) \setminus \{\varepsilon\}$. □

Beispiel 6.18

Wir betrachten folgende Grammatik $\mathcal{G}_K$ (vgl. Beispiel 6.11), die die Sprache L_K aller korrekten Klammerausdrücke erzeugt.

$$S \to (A) \mid \varepsilon$$
$$A \to S \mid AA$$

Anwenden der Konstruktion zur Elimination der ε-Regeln ergibt:

Erster Eliminationsschritt (mit $S \to \varepsilon$)
Füge die Regel $A \to \varepsilon$ hinzu.

Zweiter Eliminationsschritt (mit $A \to \varepsilon$)
Füge die Regeln $S \to ()$ und $A \to A$ hinzu.

Entferne die Regeln $S \to \varepsilon$ und $A \to \varepsilon$.

Es ergibt sich folgende Grammatik $\mathcal{G}'_K$:

$$S \to (A) \mid ()$$
$$A \to S \mid AA \mid A$$

Chomsky-Normalform

Definition 6.19

Eine kontextfreie Grammatik ist in Chomsky-Normalform (CNF), wenn sie nur Regeln der Form

$$A \to BC$$
$$A \to a,$$

für Nichtterminalsymbole A, B, C und Terminalsymbole a, hat.

Satz 6.20

Jede kontextfreie Grammatik ist äquivalent zu einer kontextfreien Grammatik in CNF.

Genauer: Zu jeder kontextfreien Grammatik $\mathcal{G}$ gibt es eine kontextfreie Grammatik $\mathcal{G}'$ in CNF, so dass $L(\mathcal{G}') = L(\mathcal{G}) \setminus \{\varepsilon\}$.

Vier Schritte

1. Elimination der ε-Regeln
   (bereits durchgeführt, Lemma 6.17)
2. Terminalsymbole nur in Regeln der Form $A \to a$
   (dann nur noch weitere Regeln $A \to A_1 \dots A_m$)
3. Elimination der Regeln $A \to B$
4. Elimination der Regeln $A \to A_1 \dots A_m$ mit $m > 2$

Als Ergebnis von Schritt 1 haben wir eine kontextfreie Grammatik $\mathcal{G}_1$ ohne ε-Regeln.

Ziel

Konstruktion einer zu $\mathcal{G}_1$ äquivalenten Grammatik $\mathcal{G}_2$, die nur Regeln der Formen

$$A \to a \quad \text{und} \quad A \to A_1 \dots A_m$$

mit $A, A_1, \dots, A_m \in N$, $m \geq 1$ und $a \in \Sigma$ enthält

Beispiel 6.21

Schritt 2 II

Aus der Grammatik $\mathcal{G}_K$ für die Sprache L_K der korrekten Klammerausdrücke erhalten wir in Schritt 1 durch Elimination der ε-Transitionen folgende Grammatik $\mathcal{G}_1$:

$$S \to (A) \mid ()$$
$$A \to S \mid AA \mid A$$

(siehe Beispiel 6.18).

Für die beiden Terminalsymbole (,) führen wir zwei neue Nichtterminalsymbole $T_($ und $T_)$ ein und verwenden diese an Stelle von (bzw.).

Außerdem fügen wir Regeln $T_(\to ($ und $T_) \to)$ hinzu.

Es ergibt sich folgende Grammatik $\mathcal{G}_2$:

$$\begin{aligned} S &\to T_(A T_) \mid T_(T_) \\ A &\to S \mid AA \mid A \\ T_(&\to (\\ T_) &\to) \end{aligned}$$

Schritt 3 I

Als Ergebnis von Schritt 2 haben wir eine kontextfreie Grammatik

$$\mathcal{G}_2 = (N_2, \Sigma, P_2, S_2)$$

die nur Regeln der Formen

$$A \to a \quad \text{und} \quad A \to A_1 \dots A_m$$

mit $A, A_1, \dots, A_m \in N$, $m \geq 1$ und $a \in \Sigma$ enthält.

Ziel

Konstruktion einer zu $\mathcal{G}_2$ äquivalenten Grammatik $\mathcal{G}_3$, die nur Regeln der Formen

$$A \to a \quad \text{und} \quad A \to A_1 \dots A_m$$

mit $A, A_1, \dots, A_m \in N$, $m \geq 2$ und $a \in \Sigma$ enthält

Konstruktion

Wir bilden die Grammatik $\mathcal{G}_3$ aus $\mathcal{G}_2$ wie folgt:

Schritt 3 II

- Für jede Ableitung

$$A_1 \to_{\mathcal{G}_2} A_2 \to_{\mathcal{G}_2} \ \dots \ \to_{\mathcal{G}_2} A_n \to_{\mathcal{G}_2} \alpha$$

  mit $n \geq 2$ und $A_1, \dots A_n \in N_2$ und $\alpha \notin N_2$ fügen wir die Regel $A_1 \to \alpha$ hinzu.
- Wir entfernen alle Regeln der Form $A \to B$, für $A, B \in N_2$.

Dann lässt sich leicht beweisen, dass $L(\mathcal{G}_2) = L(\mathcal{G}_3)$.

Praktische Umsetzung

Wie finden wir alle Ableitungen $A_1 \to_{\mathcal{G}_2} A_2 \to_{\mathcal{G}_2} \ \dots \ \to_{\mathcal{G}_2} A_n \to_{\mathcal{G}_2} \alpha$ mit $n \geq 2$ und $A_1, \dots A_n \in N_2$ und $\alpha \notin N_2$?
Und sind es nicht viel zu viele (es kann ja sogar unendlich viele solcher Ableitungen geben)?

Wir konstruieren einen gerichteten Graphen $D = (V, E)$ wie folgt:

Schritt 3 III

- Die Knotenmenge V besteht aus allen Nichtterminalen von $\mathcal{G}_2$ und allen Satzformen $\alpha \notin N_2$, die auf der rechten Seite einer Regel von $\mathcal{G}_2$ auftauchen.
- Es gibt Kanten (A, B) für alle $A, B \in N_2$, so dass $A \to B$ eine Regel von $\mathcal{G}_2$ ist, und Kanten (A, α) für alle $A \in N_2$ und Satzformen $\alpha \notin N_2$, so dass $A \to \alpha$ eine Regel von $\mathcal{G}_2$ ist.

Dann gibt es für $A_1 \in N_2$ und $\alpha \in (N \cup \Sigma^*) \setminus N_2$ genau dann eine Ableitung $A_1 \to_{\mathcal{G}_2} A_2 \to_{\mathcal{G}_2} \ \dots \ \to_{\mathcal{G}_2} A_n \to_{\mathcal{G}_2} \alpha$ mit $n \geq 2$ und $A_2, \dots A_n \in N_2$, wenn es einen Weg in D von A_1 nach α gibt.

Um $\mathcal{G}_3$ zu konstruieren müssen wir also eine Regel $A \to \alpha$ für alle $A \in N_2$ und alle $\alpha \in (N \cup \Sigma^*) \setminus N_2$ hinzufügen, für die ein Weg von A nach α in D existiert.

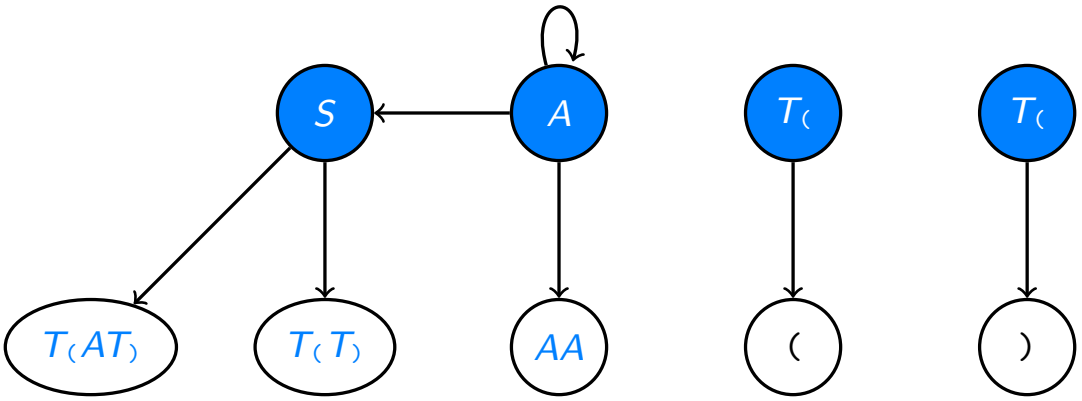
Beispiel 6.22

Schritt 3 IV

Aus der Grammatik $\mathcal{G}_K$ für die Sprache L_K der korrekten Klammerausdrücke erhalten wir in den Schritten 1 und 2 folgende Grammatik $\mathcal{G}_2$:

$$\begin{aligned} S &\to T_(A T_) \mid T_(T_) \\ A &\to S \mid AA \mid A \\ T_(&\to (\\ T_) &\to) \end{aligned}$$

Für diese Grammatik sieht der Graph D wie folgt aus:



Schritt 4 I

Als Ergebnis von Schritt 3 haben wir eine kontextfreie Grammatik

$$\mathcal{G}_3 = (N_3, \Sigma, P_3, S_3)$$

die nur Regeln der Formen

$$A \to a \quad \text{und} \quad A \to A_1 \dots A_m$$

mit $A, A_1, \dots, A_m \in N$, $m \geq 2$ und $a \in \Sigma$ enthält.

Ziel

Konstruktion einer zu $\mathcal{G}_3$ äquivalenten Grammatik $\mathcal{G}_4$, die nur Regeln der Formen

$$A \to a \quad \text{und} \quad A \to A_1 A_2$$

mit $A, A_1, A_2 \in N$ und $a \in \Sigma$ enthält

Schritt 4 II

Idee

Ersetze Regel $A \to A_1 \dots A_m$ mit $m \geq 2$ durch

$$A \to A_1 B_1, \quad B_1 \to A_2 B_2, \quad B_2 \to A_3 B_3,$$
$$\dots \; B_{m-3} \to A_{m-2} B_{m-2}, \quad B_{m-2} \to A_{m-1} A_m$$

mit neuen Nichtterminalsymbolen B_i.

Beispiel 6.23

Schritt 4 III

Aus der Grammatik $\mathcal{G}_K$ für die Sprache L_K der korrekten Klammerausdrücke erhalten wir in den Schritten 1–3 folgende Grammatik $\mathcal{G}_3$:

$$\begin{aligned} S &\to T_(A T_) \mid T_(T_) \\ A &\to T_(A T_) \mid T_(T_) \mid AA \\ T_(&\to (\\ T_) &\to) \end{aligned}$$

Wir müssen die beiden Regeln $S \to T_(A T_)$ und $A \to T_(A T_)$ ersetzen durch

$$S \to T_(B_1, \quad B_1 \to A T_)$$

bzw.

$$A \to T_(C_1, \quad C_1 \to A T_).$$

Insgesamt erhalten wir folgende Grammatik $\mathcal{G}_4$ in CNF:

$$\begin{aligned} S &\to T_(B_1 \mid T_(T_) & B_1 &\to A T_) \\ A &\to T_(C_1 \mid T_(T_) \mid AA & C_1 &\to A T_) \\ T_(&\to (\end{aligned}$$

Definition 6.24

Eine kontextfreie Grammatik ist in Greibach-Normalform, wenn sie nur Regeln der Form

$$A \to aB_1 \dots B_k,$$

für $k \geq 0$, Nichtterminalsymbole $A, B_1, \dots, B_k$ und Terminalsymbole a hat.

Satz 6.25

Jede kontextfreie Grammatik ist äquivalent zu einer kontextfreien Grammatik in Greibach-Normalform.

Genauer: Zu jeder kontextfreien Grammatik G gibt es eine kontextfreie Grammatik G' in Greibach-Normalform, so dass $L(G') = L(G) \setminus \{\varepsilon\}$.

(Ohne Beweis.)

Beispiel 6.26

Sei $\mathcal{G}'_K$ die kontextfreie Grammatik

$$\begin{aligned} S &\to (A) \mid () \\ A &\to S \mid AA \end{aligned}$$

die die Sprache $L_K \setminus \{\varepsilon\}$ erzeugt.

Eine äquivalente Grammatik in Greibach-Normalform ist

$$\begin{aligned} S &\to (AT_) \mid (T_) \\ A &\to (AT_) \mid (T_) \mid (AT_)A \mid (T_)A \\ T_) &\to) \end{aligned}$$

Beweisidee für Satz 6.25

Wir verfolgen „linkestmögliche“ Ableitungen, bis ganz links ein Terminalsymbol steht, und bilden daraus neue Regeln.
Für weiter rechts auftretende Terminalsymbole verwenden wir den gleichen Trick wie bei der Chomsky-Normalform.

Definition 6.27

Eine kontextfreie Grammatik ist rechtslinear, wenn sie nur Regeln der Form

$$A \to aB$$
$$A \to a,$$

für Nichtterminalsymbole A, B, und Terminalsymbole a, hat.

Beobachtung 6.28

Jede rechtslineare Grammatik ist in Greibach-Normalform, aber nicht umgekehrt.

Achtung

Rechtslineare Grammatiken bilden keine Normalform!

Es gibt kontextfreie Sprachen (die ε nicht enthalten), für die es keine rechtslineare Grammatik gibt.

(Das folgt aus dem nächsten Satz.)

Satz 6.29

Rechtslineare Grammatiken erzeugen gerade die regulären Sprachen.

Genauer:

1. *Zu jeder regulären Sprache L gibt es eine rechtslineare kontextfreie Grammatik $\mathcal{G}$ so dass $L(\mathcal{G}) = L \setminus \{\varepsilon\}$.*
2. *Für jede rechtslineare kontextfreie Grammatik $\mathcal{G}$ ist die Sprache $L(\mathcal{G})$ regulär.*

Beweis von Satz 6.29 I

Behauptung 1

Zu jeder regulären Sprache L gibt es eine rechtslineare kontextfreie Grammatik $\mathcal{G}$ so dass $L(\mathcal{G}) = L \setminus \{\varepsilon\}$.

Beweis. Unser Beweis, dass jede reguläre Sprache kontextfrei ist, liefert zu jedem NFA $\mathcal{A} = (Q, \Sigma, \Delta, q_0, F)$ folgende äquivalente kontextfreie Grammatik $\mathcal{G}_\mathcal{A} = (Q, \Sigma, P, q_0)$ mit Regeln

$$\begin{aligned} q &\to ar && \text{für alle } (q, a, r) \in \Delta, \\ q &\to \varepsilon && \text{für alle } q \in F. \end{aligned}$$

Wir eliminieren die ε-Regeln, indem wir für alle $(q, a, r) \in \Delta$ mit $r \in F$ die Regel $q \to a$ hinzufügen (und danach die Regeln $q \to \varepsilon$ entfernen).
Die resultierende Grammatik ist rechtslinear. □

Behauptung 2

Beweis von Satz 6.29 II

Für jede rechtslineare kontextfreie Grammatik $\mathcal{G}$ ist die Sprache $L(\mathcal{G})$ regulär.

Beweis. Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine rechtslineare kontextfreie Grammatik.
Wir beobachten zunächst, dass jede Ableitung eines Terminalwortes in $\mathcal{G}$ die Form

$$S \to a_1 A_1 \to a_1 a_2 A_2 \to \quad \dots \quad \to a_1 \dots a_{k-1} A_{k-1} \to a_1 \dots a_k \qquad (\star)$$

für ein $k \geq 0$ und Regeln $S \to a_1 A_1$, $A_1 \to a_2 A_2$, $A_2 \to a_3 A_3$, …, $A_{k-2} \to a_{k-1} A_{k-1}$, $A_{k-1} \to a_k$ hat.

Wir definieren einen NFA $\mathcal{A} = (Q, \Sigma, \Delta, q_0, F)$ durch

- $Q := N \cup \{q_f\}$, wobei $q_f \notin N$,
- $\Delta = \big\{(A, a, B) \;\big|\; A \to aB \in P\big\} \cup \big\{(A, a, q_f) \;\big|\; A \to a \in P\big\}$,
- $q_0 := S$,
- $F := \{q_f\}$.

Beweis von Satz 6.29 III

Dann entspricht die Ableitung $(\star)$ des Wortes $a_1 \dots a_k \in L(\mathcal{G})$ dem akzeptierenden Lauf

$$S \xrightarrow{a_1} A_1 \xrightarrow{a_2} A_2 \xrightarrow{a_3} \dots \xrightarrow{a_{k-1}} A_{k-1} \xrightarrow{a_k} q_f. \qquad (\star\star)$$

Also gilt $L(\mathcal{G}) \subseteq L(\mathcal{A})$.

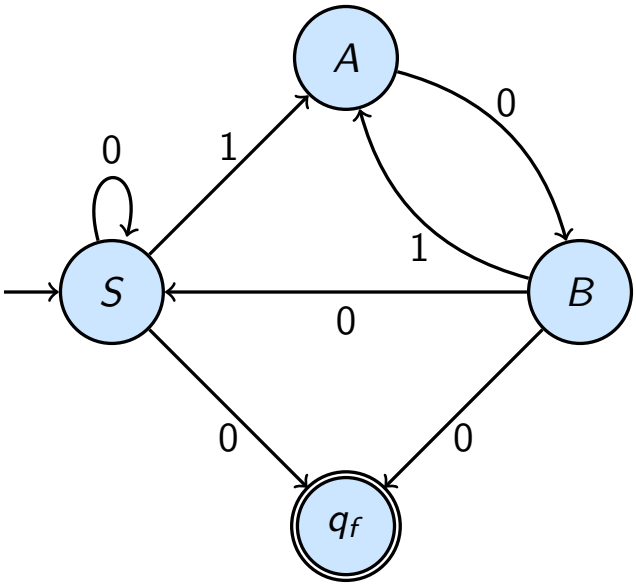
Umgekehrt hat jeder akzeptierende Lauf von $\mathcal{A}$ auf einem Wort $w = a_1 \dots a_k$ die Gestalt $(\star\star)$ und entspricht damit einer Ableitung von w in $\mathcal{G}$ der Gestalt $(\star)$. Also gilt auch $L(\mathcal{A}) \subseteq L(\mathcal{G})$. $\square$

Beispiel 6.30

Wir betrachten folgende rechtslineare Grammatik:

$$S \to 0S \mid 1A \mid 0$$
$$A \to 0B$$
$$B \to 0S \mid 1A \mid 0$$

Unsere Konstruktion liefert folgenen äquivalenten NFA:



Abschnitt 6.3

Ableitungsbäume

Grundüberlegung

Ableitungen sind eigentlich nicht linear, sondern haben eine Baumstruktur, denn die Reihenfolge, in der die Regeln angewandt werden, ist unwichtig.

Beispiel 6.31

Die Ableitungen

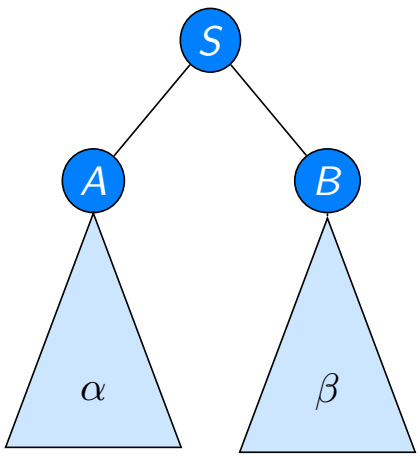
$$S \to AB \to \alpha B \to \alpha\beta$$

und

$$S \to AB \to A\beta \to \alpha\beta$$

sind „im Wesentlichen gleich“.

Für viele Zwecke geeigneter ist die Darstellung der beiden Ableitungen als Baum.



Beispiel 6.32

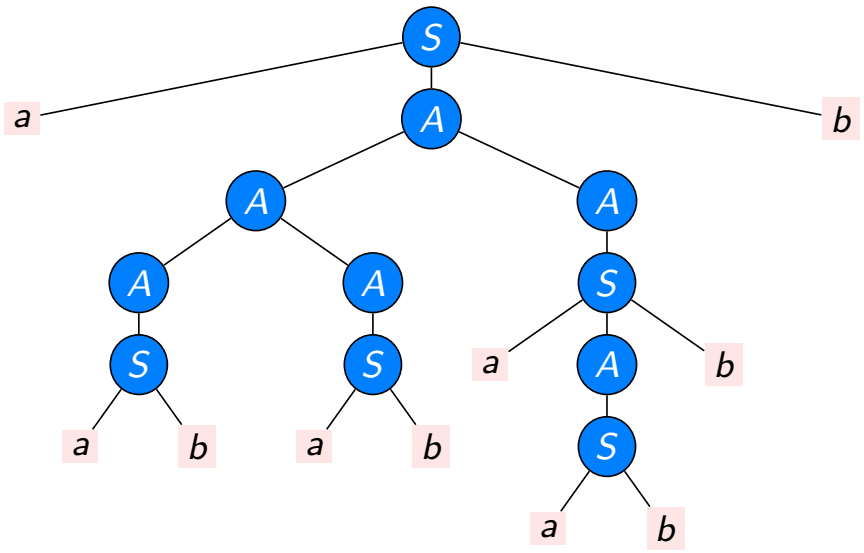
Grammatik:

$$S \to aAb \mid ab$$
$$A \to S \mid AA$$

Eine Ableitung:

$$\underline{S} \to a\underline{A}b \to a\underline{AA}b \to a\underline{AAA}b \to a\underline{S}AAb \to aab\underline{AA}b \to aab\underline{S}Ab$$
$$\to aabab\underline{A}b \to aabab\underline{S}b \to aababa\underline{A}bb \to aababa\underline{S}bb \to aababaabbb$$

Ableitungsbaum:



Format der Ableitungsbäume

Wir unterscheiden in einem Ableitungsbaum T zu einer Grammatik $\mathcal{G} = (N, \Sigma, P, S)$

- die Knotenmenge $\mathrm{dom}(T)$
- die Beschriftungsfunktion

$$T : \mathrm{dom}(T) \to (N \cup \Sigma),$$

die jedem Knoten ein Symbol aus N oder Σ zuordnet

Definition 6.33

1. Ein Baumbereich ist eine Menge $D \subseteq (\mathbb{N}_+)^*$, die
   - unter Präfixbildung abgeschlossen ist,
   - jeweils für uj auch alle ui mit $1 \leq i \leq j$ enthält.

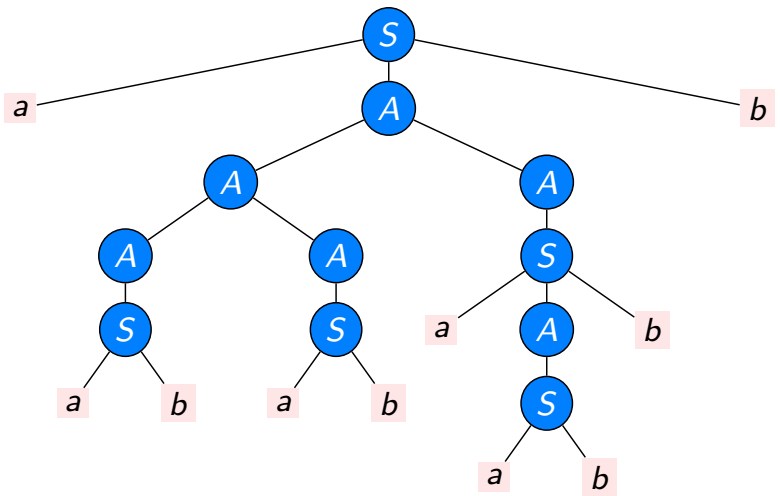
   In dieser Vorlesung betrachten wir nur endliche Baumbereiche.

2. Sei Σ ein Alphabet. Ein Σ-Baum ist eine Funktion $T : \mathrm{dom}(T) \to \Sigma$, wobei $\mathrm{dom}(T)$ ein Baumbereich ist.

3. Ein Baum T ist höchstens k-verzweigt, wenn $\mathrm{dom}(T) \subseteq \{1, \ldots, k\}^*$.

Beispiel

Ableitungsbaum T



Knotenmenge

$$\mathrm{dom}(T) = \{\varepsilon, 1, 2, 21, 211, 2111, 21111, 21112, 212, 2121, 21211, 21212, 22, 221, 2211, 2212, 22121, 221211, 221212, 2213, 3\}$$

Beschriftung

v	ε	1	2	21	211	2111	21111	21112	212	2121	$\cdots$
$T(v)$	S	a	A	A	A	S	a	b	A	S	$\cdots$

Baumterminologie

Sei T ein Σ-Baum.

- ε ist die Wurzel von T.
- Ein Knoten $ui \in \mathrm{dom}(T)$ ist $i$ter Nachfolger (oder $i$tes Kind) von u, umgekehrt ist u der Vorgänger von ui.
- Die Blätter von T sind die Knoten ohne Nachfolger, alle andere Knoten sind innere Knoten.
- Seien $b_1, \ldots, b_n \in \mathrm{dom}(T)$ die Blätter von T in lexikographischer Reihenfolge, d.h., von links nach rechts. Dann ist das Wort

$$\beta(T) := T(b_1) \ldots T(b_n) \in \Sigma^*$$

die Blattbeschriftung von T.

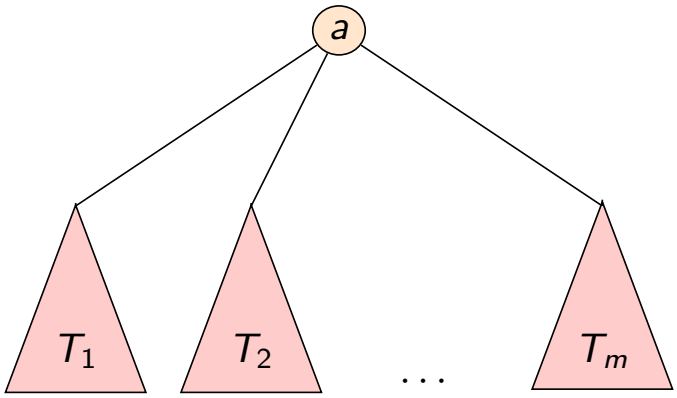
Baumterminologie (Forts.)

Sei T ein Σ-Baum.

- Ein Ast (der Länge n) in T ist eine Knotenfolge $v_0, \ldots, v_n$ mit $v_0 = \varepsilon$, v_{i+1} jeweils Nachfolger von v_i, und Blatt v_n,
- die Höhe $h(T)$ des Baumes T ist die Länge eines längsten Astes, also

$$h(T) := \max\{|v| \mid v \in \mathrm{dom}(T)\}.$$

Ein Baum T der Höhe $h+1$ mit Wurzelbeschriftung a hat die Form



mit Bäumen $T_1, \ldots, T_m$ der Höhe jeweils $\leq h$,
d.h. es gilt

- $\mathrm{dom}(T) = \{\varepsilon\} \cup \{iv \mid 1 \leq i \leq m, v \in v(T_i)\}$,
- $T(\varepsilon) = a$, $T(1v) = T_1(v), \ldots, T(mv) = T_m(v)$.

Definition 6.34

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine Grammatik.

1. Ein Ableitungsbaum zu $\mathcal{G}$ (englisch: Parse Tree) ist ein $(N \cup \Sigma)$-Baum T, für den gilt:
   - $T(\varepsilon) \in N$,
   - hat v die Nachfolger $v1, \ldots, vm$ in $\mathrm{dom}(T)$, so ist $T(v) \to T(v1) \ldots T(vm)$ eine Regel in P.

2. Ein Ableitungsbaum T zu $\mathcal{G}$ ist vollständig, wenn $T(\varepsilon) = S$ und $\beta(T) \in \Sigma^*$.

Satz 6.35

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik. Dann gilt für alle $A \in N$ und $\alpha \in (N \cup \Sigma)^*$:

$$A \overset{*}{\to} \alpha \iff \text{es gibt einen Ableitungsbaum } T \text{ zu } \mathcal{G} \text{ mit } T(\varepsilon) = A \text{ und } \beta(T) = \alpha.$$

Wir nennen T auch einen Ableitungsbaum für α aus A in $\mathcal{G}$

Korollar 6.36

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik. Dann gilt für alle $w \in \Sigma^*$:

$$w \in L(\mathcal{G}) \iff \text{es gibt einen vollständigen Ableitungsbaum } T \text{ zu } \mathcal{G} \text{ mit } \beta(T) = w.$$

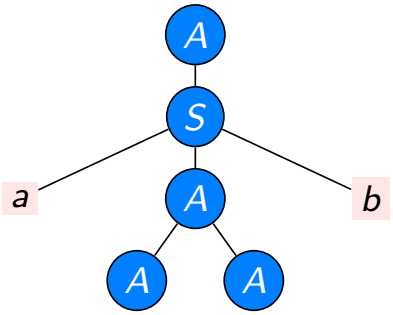
Wir nennen T auch einen Ableitungsbaum für w in $\mathcal{G}$

Beispiele 6.37

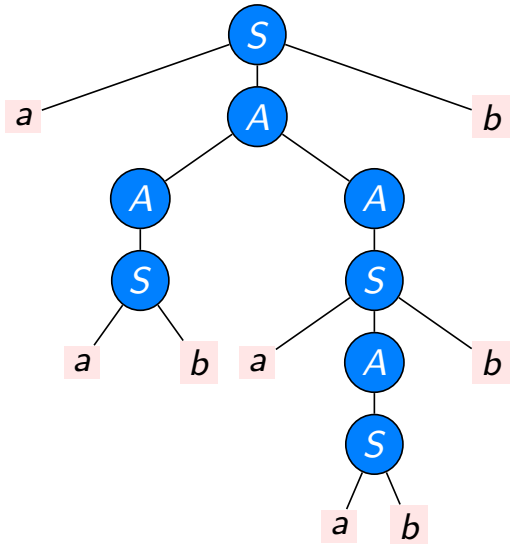
Wir betrachten wieder die Grammatik

$$S \to aAb \mid ab$$
$$A \to S \mid AA$$

Ableitungsbaum für *aAAb* aus *A*:



Ableitungsbaum für *aabaabbb*:



Beweis von Satz 6.35 I

Satz 6.35

Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik und $A \in N$ und $\alpha \in (N \cup \Sigma)^*$. Dann gilt

$$A \xrightarrow{*} \alpha \iff \text{es gibt einen Ableitungsbaum } T \text{ zu } \mathcal{G} \text{ mit } T(\varepsilon) = A \text{ und } \beta(T) = \alpha.$$

Beweis.„$\Longrightarrow$“: Induktion über die Länge n einer Ableitung von α aus A in $\mathcal{G}$.

Induktionsschritt $n = 0$:

Gelte $A \xrightarrow{0} \alpha$. Dann gilt $\alpha = A$, und

Beweis von Satz 6.35 II

ist ein Ableitungsbaum für α aus A.

Induktionsschritt $0, \ldots, n \to n+1$:

Sei

$$\alpha_0 = A \to \alpha_1 \to \ldots \to \alpha_{n+1} = \alpha$$

eine Ableitung von α aus A.

Wir zerlegen α_1 in Terminalwörter und Nichtterminale:

$$\alpha_1 = w_0 B_1 w_1 B_2 \ldots w_{k-1} B_k w_k,$$

für ein $k \geq 0$, $w_0, \ldots, w_k \in \Sigma^*$ und $B_1, \ldots, B_k \in N$. Die w_is können das leere Wort sein.

Nach dem Zerlegungslemma 6.10 lässt sich α dann zerlegen als

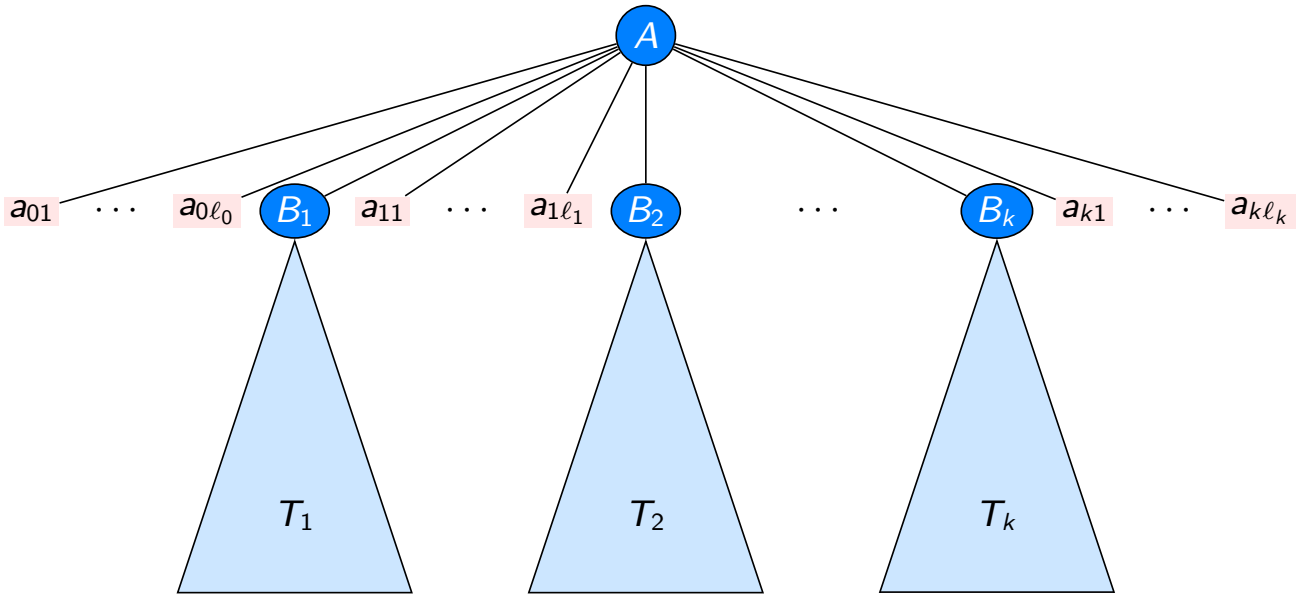
$$\alpha = w_0 \beta_1 w_1 \beta_2 \ldots w_{k-1} \beta_k w_k,$$

so dass $B_i \xrightarrow{\leq n} \beta_i$ für $1 \leq i \leq k$.

Beweis von Satz 6.35 III

Nach Induktionsannahme gibt es für $1 \leq i \leq k$ einen Ableitungsbaum T_i für β_i aus B_i.

Nehmen wir an, für $0 \leq i \leq k$ ist $w_i = a_{i1} \dots a_{i\ell_i}$. Dann ist



Beweis von Satz 6.35 IV

ein Ableitungsbaum für α aus A.

„$\Longleftarrow$“: Induktion über die Höhe h eines Ableitungsbaums T für α aus A.

Induktionsanfang $h = 0$:

Ein Ableitungsbaum der Höhe 0 von α aus A hat die Gestalt

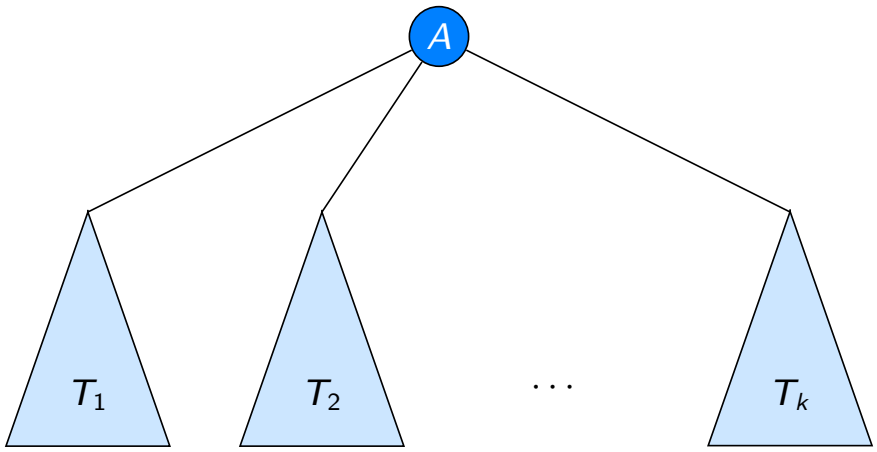
.

Dann gilt $\alpha = A$, und es gilt $A \xrightarrow{0} \alpha$.

Induktionsschritt $0, \dots, h \to h + 1$:

Sei T ein Ableitungsbaum der Höhe $h + 1$ von α aus A. Wir zerlegen T an der Wurzel in Bäume $T_1, \dots, T_k$ der Höhe $\leq h$:

Beweis von Satz 6.35 V



Dann gilt

$$\alpha = \beta(T) = \beta(T_1) \dots \beta(T_k).$$

Außerdem ist

$$A \to T_1(\varepsilon) \dots T_k(\varepsilon)$$

eine Regel von $\mathcal{G}$.

Beweis von Satz 6.35 VI

Nach Induktionsannahme gilt für $1 \leq i \leq k$

$$T_i(\varepsilon) \xrightarrow{*} \beta(T_i).$$

Nach dem Kombinationslemma 6.9 gilt dann

$$A \to T_1(\varepsilon) \dots T_k(\varepsilon) \xrightarrow{*} \beta(T_1) \dots \beta(T_k) = \alpha.$$

□

Wörter mit einkodierten Ableitungsbäumen

Idee

Wir wollen Ableitungsbäume „linearisieren“ und sie wieder als Wörter kodieren.

Genau das geschieht in XML-Dokumenten.

Methode

Für jedes Nichtterminalsymbol A führen wir neue Terminalsymbole

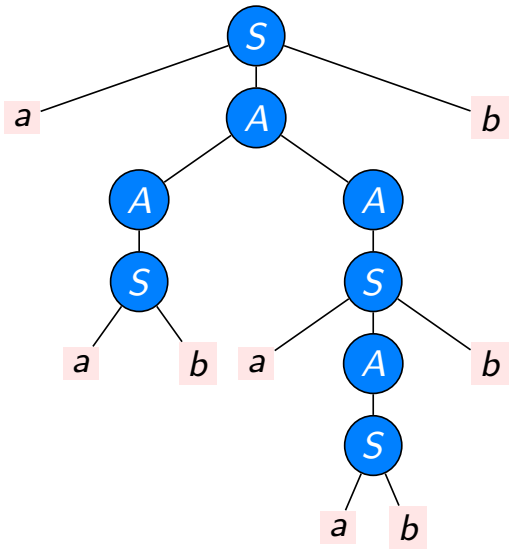
$$\langle_A \qquad \text{und} \qquad \rangle_A$$

ein.

Ableitungsbaum T für ein Wort w wird durch ein Wort $x(T)$ kodiert, das aus w durch Einfügen dieser neuen Symbole entsteht.

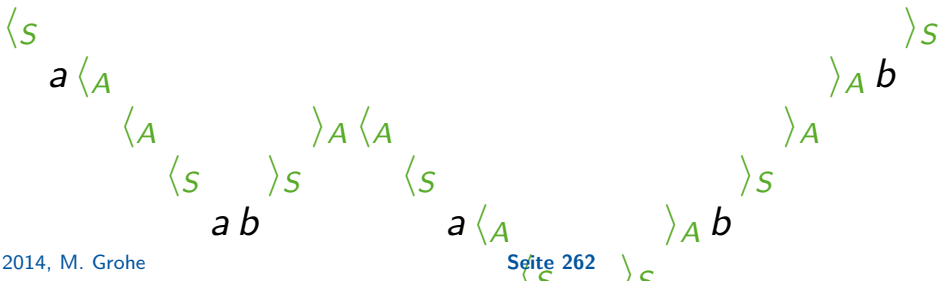
Beispiel

Ableitungsbaum T



Wort $x(T)$ mit einkodiertem Ableitungsbaum

$$\langle_S\ a\ \langle_A\ \langle_A\ \langle_S\ a\ b\ \rangle_S\ \rangle_A\ \langle_A\ \langle_S\ a\ \langle_A\ \langle_S\ a\ b\ \rangle_S\ \rangle_A\ b\ \rangle_S\ \rangle_A\ \rangle_A\ b\ \rangle_S$$



Definition 6.38

Sei $\mathcal{G} = (N, \Sigma, P, S)$.
Für jeden Ableitungsbaum T zu $\mathcal{G}$ definieren wir rekursiv über die Höhe h von T ein Wort

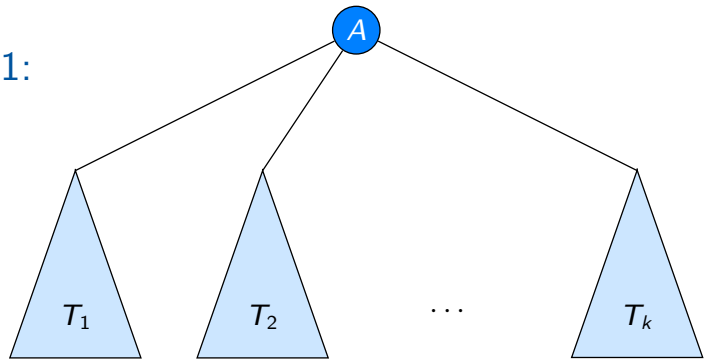
$$x(T) \in \Big(\Sigma \cup \big\{\langle_A, \rangle_A \mid A \in N\big\}\Big)^*.$$

Rekursionsanfang $h = 0$:

$$x(T) := \begin{cases} a & \text{falls } T(\varepsilon) = a \in \Sigma, \\ \langle_A \rangle_A & \text{falls } T(\varepsilon) = A \in N. \end{cases}$$

Rekursionsschritt $0, \ldots, h \to h+1$:
Wir zerlegen T an der Wurzel.
Sei $T(\varepsilon) = A$, und seien
$T_1, \ldots, T_k$ die Teilbäume.

Wir setzen
$x(T) := \langle_A x(T_1) \ldots x(T_k) \rangle_A$.



XML-Dokumente

XML-Dokumente sind Wörter der Form $x(T)$ mit modifizierter Klammernotation:

- `<A>` für $\langle_A$,
- `</A>` für $\rangle_A$.

Beispiel 6.39

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Studienprofil [...]>
<Studienprofil>
    <Studiengang> INFORMATIK </Studiengang>
    <Fachsemester> 3 </Fachsemester>
    <Vorlesungen>
        <Titel> PROGRAMMIERUNG </Titel>
        <Titel> TECHNISCHE INFORMATIK </Titel>
        <Titel> DATENSTRUKTUREN </Titel>
    </Vorlesungen>
</Studienprofil>
```

Document-Type Definitionen (DTD's)

Eine Document-Type Definition (DTD) ist ein Regelsystem zur Erzeugung einer Menge von XML-Dokumenten, also eine Grammatik zum Aufbau von Terminalwörtern zusammen mit zugehörigen Ableitungsbäumen.

Beispiel 6.40

DTD dargestellt als verallgemeinerte kontextfreie Grammatik.

Studienprofil → Studiengang · Fachsemester · Vorlesungen
Studiengang → Terminalwort
Fachsemester → Terminalwort
Vorlesungen → Titel*
Titel → Terminalwort

Beispiel 6.40 (Forts.)

DTD dargestellt als verallgemeinerte kontextfreie Grammatik

Studienprofil → Studiengang · Fachsemester · Vorlesungen
Studiengang → Terminalwort
Fachsemester → Terminalwort
Vorlesungen → Titel*
Titel → Terminalwort

Darstellung in Standardsyntax

```
<!DOCTYPE Studienprofil [
   <!ELEMENT Stpr (Studgang, Fachsem, Vorlesg)>
   <!ELEMENT Studgang (# PCDATA)>
   <!ELEMENT Fachsem (# PCDATA)>
   <!ELEMENT Vorlesg (Titel*)>
   <!ELEMENT Titel (# PCDATA)>
]>
```

Hierbei steht # PCDATA für "parsed character data", das heißt, für Wörter über dem Terminalalphabet.

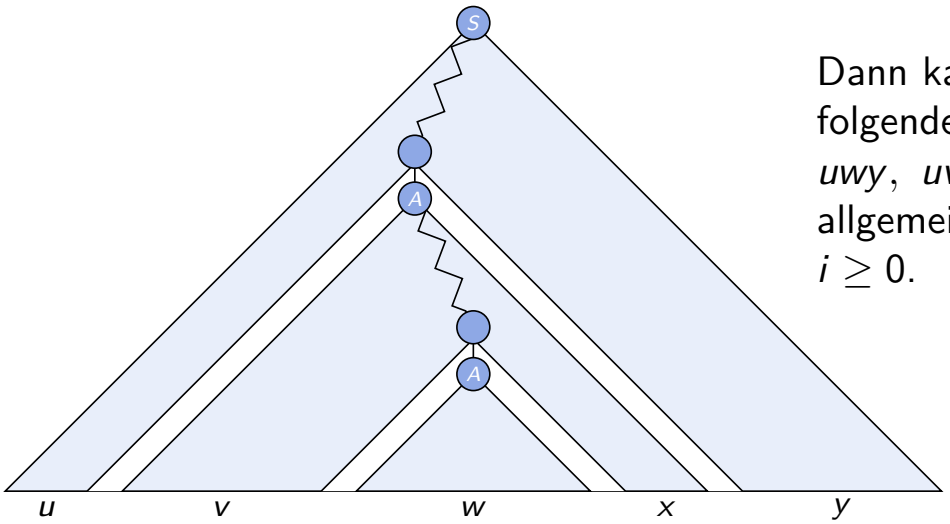
Abschnitt 6.4

Das Pumping-Lemma für kontextfreie Sprachen

Idee

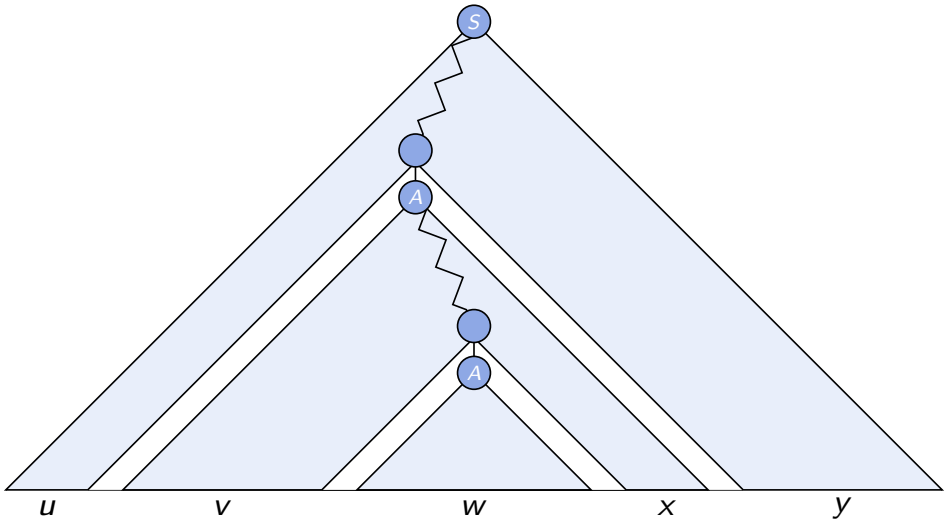
Sei $\mathcal{G}$ eine kontextfreie Grammatik.

Ist $z \in L(\mathcal{G})$ hinreichend lang, so hat z einen Ableitungsbaum so großer Höhe, dass sich auf einem langen Ast ein Nichtterminalsymbol wiederholt.



Dann kann man in $\mathcal{G}$ auch folgende Wörter ableiten: uwy, $uvvwxxy$ und allgemein uv^iwx^iy für $i \geq 0$.

Idee (Forts.)



Pumping-Lemma

Lemma 6.41 (Pumping-Lemma für kontextfreie Sprachen)

Sei $L \subseteq \Sigma^*$ kontextfrei. Dann gibt es eine Zahl $n \geq 1$, so dass jedes Wort $z \in L$ mit $|z| \geq n$ zerlegbar ist in Wörter $u, v, w, x, y \in \Sigma^*$ mit

$$z = uvwxy,$$

die folgende Eigenschaften haben:

(i) $vx \neq \varepsilon$,

(ii) $|vwx| \leq n$,

(iii) $uv^k wx^k y \in L$ *für alle* $k \geq 0$.

Beweis des Pumping-Lemmas I

Lemma 6.42

Ein höchstens d-verzweigter Baum der Höhe h hat höchstens d^h Blätter.

Beweis. Induktion über h.

Induktionsanfang $h = 0$:
Ein Baum der Höhe 0 hat $1 = d^0$ Blatt.

Induktionsschritt $h \to h+1$:
Sei T ein Baum der Höhe $h+1$. T lässt sich an der Wurzel in Bäume $T_1, \ldots, T_\ell$ der Höhe $\leq h$ für ein $\ell \leq d$ zerlegen.

$$\begin{array}{lll} \text{Induktionsannahme} & \Longrightarrow & (\text{Zahl der Blätter von } T_i) \leq d^h \\ & \Longrightarrow & (\text{Zahl der Blätter von } T) \leq \ell \cdot d^h \leq d^{h+1}. \end{array}$$

□

Beweis des Pumping-Lemmas. Sei $\mathcal{G} = (N, \Sigma, P, S)$ eine kontextfreie Grammatik in CNF für L (bzw. $L \setminus \{\varepsilon\}$).

Beweis des Pumping-Lemmas II

Wir setzen

$$n := 2^{|N|+1}.$$

Sei $z \in L$ mit $|z| > n$ und T ein Ableitungsbaum für z in G.

T ist Baum mit Verzweigungsgrad höchstens 2 (wegen CNF) mit $|z| \geq n > 2^{|N|}$ Blättern. Nach Lemma 6.42 hat T also Höhe

$$h \geq |N| + 1.$$

Sei $p_0 = \varepsilon, p_1, p_2, \ldots, p_h$ ein Ast von T der Länge h. Dabei sind die Knoten $p_0, \ldots, p_{h-1}$ innere Knoten des Baumes und deswegen mit Nichtterminalsymbolen beschriftet.

Weil $h > |N|$, gibt es i, j, so dass

$$h - 1 - |N| \leq i < j \leq h - 1$$

und $T(p_i) = T(p_j) \in N$.

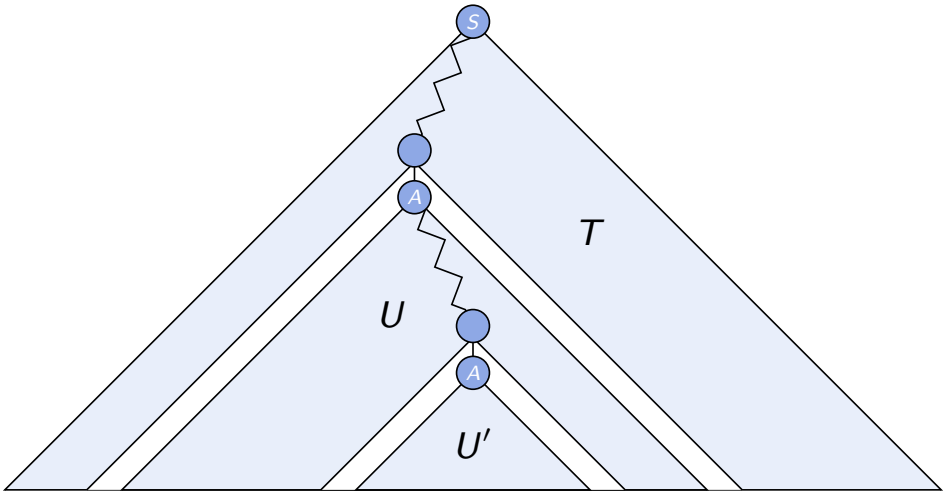
Sei U der Teilbaum von T mit Wurzel p_i.

Beweis des Pumping-Lemmas III

Formal:

$$\mathrm{dom}(U) := \{q \in \{1,2\}^* \mid p_i q \in \mathrm{dom}(T)\}$$

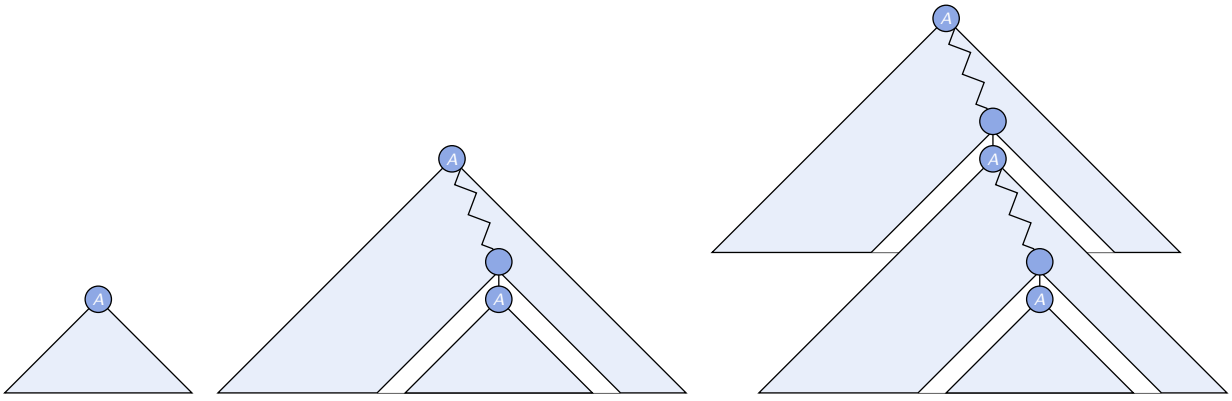
und $U(q) := T(p_i q)$.
Ähnlich sei U' der Teilbaum von T mit Wurzel p_j.



Beweis des Pumping-Lemmas IV

Sei $U_0 := U'$, $U_1 := U$, und für $i \geq 1$ sei U_{i+1} der Baum, der aus U entsteht, wenn man U' durch U_i ersetzt.

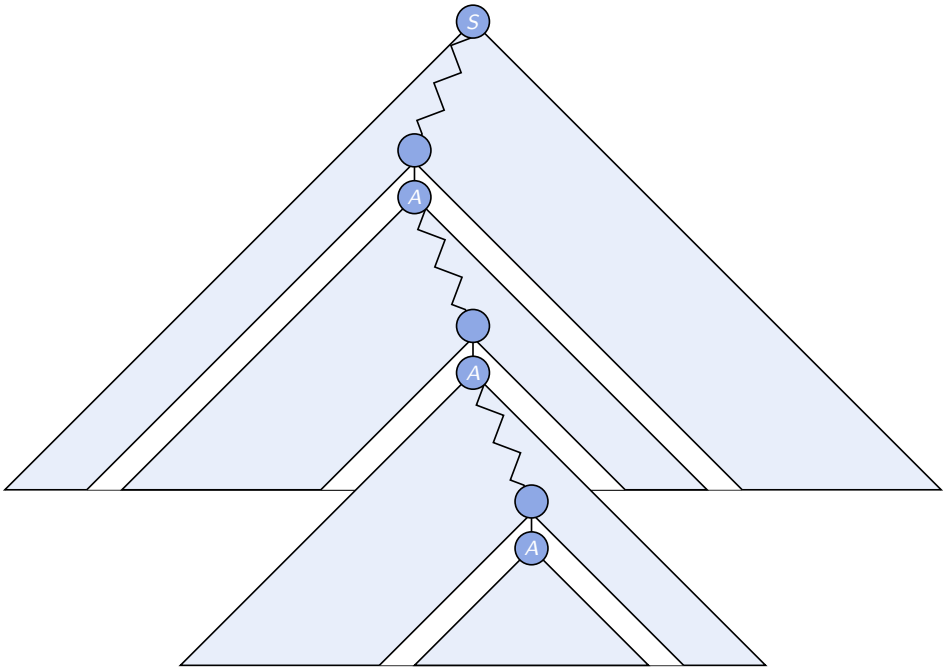
Beispiel: Die Bäume $U_0.U_1, U_2$.



Für $i \geq 0$ sei T_i der Baum, der aus T entsteht, indem man den Teilbaum U durch U_i ersetzt.

Beispiel: Der Baum T_2.

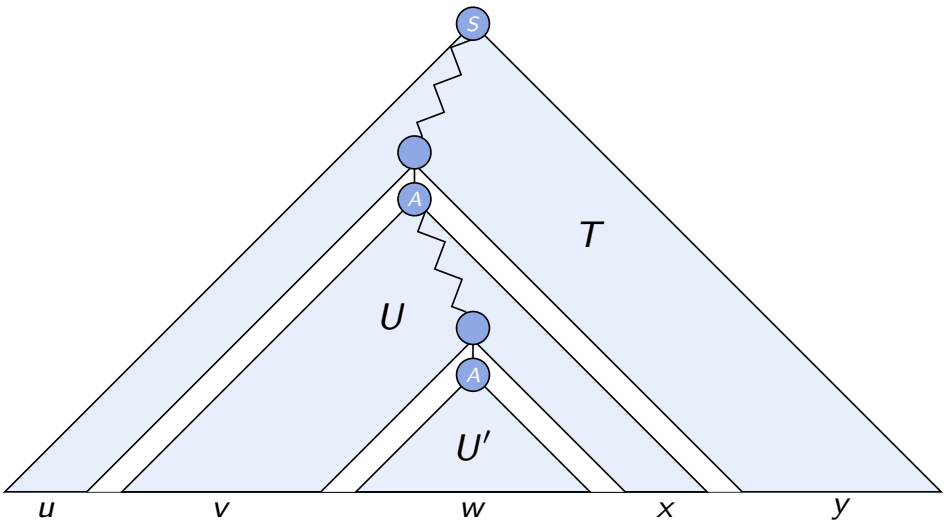
Beweis des Pumping-Lemmas V



Beweis des Pumping-Lemmas VI

Dann ist T_i Ableitungsbaum zu $\mathcal{G}$.

Sei $w := \beta(U')$. Wähle v, x so, dass $\beta(U) = vwx$.
Wähle u, y so, dass $z = \beta(T) = uvwxy$.



- Weil $i < j$ und damit $U \neq U'$ gilt $vx \neq \varepsilon$.

Beweis des Pumping-Lemmas VII

- Weil $i \geq h - 1 - |N|$ ist die Höhe von U höchstens $|N| + 1$. Deswegen gilt $|vwx| = |\beta(U)| \leq 2^{|N|+1} = n$.
- Für alle $k \geq 0$ gilt $uv^k wx^k y = \beta(T_i) \in L$.

□

Beispiel 6.43 I

Die Sprache $L = \{a^n b^n c^n \mid n \geq 0\} \subseteq \{a, b, c\}^*$ ist nicht kontextfrei.

Beweis.

Angenommen, L ist kontextfrei.

Wähle n zu L gemäß Pumping-Lemma und betrachte $z = a^n b^n c^n \in L$.

Pumping-Lemma liefert Zerlegung

$$z = uvwxy$$

mit $vx \neq \varepsilon$ und $|vwx| \leq n$ und $uwy \in L$.

Weil $|vwx| \leq n$ liegt das Infix vwx von $z = a^n b^n c^n$ ganz im Präfix $a^n b^n$ oder ganz im Suffix $b^n c^n$.

Fall 1: vwx liegt ganz im Präfix $a^n b^n$.

Dann sind alle cs aus z in y. Weil $vx \neq \varepsilon$ hat das Wort uwy mindestens ein a oder ein b weniger als z, aber genauso viele cs.

Beispiel 6.43 II

Also gilt $uwy \notin L$. Widerspruch!

Fall 2: vwx liegt ganz im Suffix $b^n c^n$.

Dann sind alle as aus z in u. Weil $vx \neq \varepsilon$ hat das Wort uwy mindestens ein b oder ein c weniger als z, aber genauso viele as.

Also gilt $uwy \notin L$. Widerspruch! □

Beispiel 6.44 I

Die Sprache $L_W = \{ww \mid w \in \{a, b\}^*\}$ aller „Wiederholungswörter“ über dem Alphabet $\{a, b\}$ ist nicht kontextfrei.

Beweis.

Angenommen, L_W ist kontextfrei.

Wähle n zu L_W gemäß Pumping-Lemma und betrachte $z = a^n b^n a^n b^n \in L$.

Pumping-Lemma liefert Zerlegung

$$z = uvwxy$$

mit $vx \neq \varepsilon$ und $|vwx| \leq n$ und $uwy \in L$.

Weil $|vwx| \leq n$ liegt das Infix vwx von $z = a^n b^n a^n b^n$ ganz im Präfix $a^n b^n$ oder ganz im Infix $b^n a^n$ oder ganz im Suffix $a^n b^n$.

Fall 1: vwx liegt ganz im Präfix $a^n b^n$.

Dann gilt $uwy = a^\ell b^m a^n b^n$, wobei $\ell + m < 2n$. Also gilt $uwy \notin L$.
Widerspruch!

Fall 2: vwx liegt ganz im Suffix $a^n b^n$.

Analog zu Fall 1.

Fall 3: vwx liegt ganz im Infix $b^n a^n$, aber weder im Präfix $a^n b^n$ noch im Suffix $a^n b^n$.

Dann gilt $uwy = a^n b^\ell a^m b^n$ für $1 \leq \ell, m < n$. Also $uwy \notin L$.

Widerspruch! □

Abschnitt 6.5
Abschlusseigenschaften kontextfreier Sprachen

Satz 6.45

Die Klasse der kontextfreien Sprachen ist unter Vereinigung, Verkettung und Iteration abgeschlossen.

Beweis I

Vereinigung: Übung.

Verkettung: Seien $L_1, L_2 \subseteq \Sigma^*$ kontextfreie Sprachen. Wir wollen zeigen, dass $L_1 L_2$ kontextfrei ist.

Seien dazu $\mathcal{G}_1 = (N_1, \Sigma, P_1, S_1)$ und $\mathcal{G}_2 = (N_2, \Sigma, P_2, S_2)$ kontextfreie Grammatiken mit $L(\mathcal{G}_1) = L_1$ und $L(\mathcal{G}_2) = L_2$. OBdA gelte $N_1 \cap N_2 = \emptyset$.

Wir definieren eine Grammatik $\mathcal{G} = (N, \Sigma, P, S)$ wie folgt:

- $S \notin N_1 \cup N_2$ ist ein neues Nichtterminalsymbol;
- $N := N_1 \cup N_2 \cup \{S\}$;
- $P := P_1 \cup P_2 \cup \{S \to S_1 S_2\}$.

Beweis II

Behauptung: $L(\mathcal{G}) = L_1 L_2$.

Beweis.
„$\subseteq$“: Sei $w \in L(\mathcal{G})$. Dann gilt $S \stackrel{*}{\to}_{\mathcal{G}} w$. Weil $S \to S_1 S_2$ die einzige Regel für S ist, gilt

$$S \to_{\mathcal{G}} S_1 S_2 \stackrel{*}{\to}_{\mathcal{G}} w.$$

Nach dem Zerlegungslemma 6.10 gibt es dann $w_1, w_2 \in \Sigma^*$, so dass $w = w_1 w_2$ mit

$$S_1 \stackrel{*}{\to}_{\mathcal{G}} w_1 \quad \text{und} \quad S_2 \stackrel{*}{\to}_{\mathcal{G}} w_2.$$

Dann gilt auch

$$S_1 \stackrel{*}{\to}_{\mathcal{G}_1} w_1 \quad \text{und} \quad S_2 \stackrel{*}{\to}_{\mathcal{G}_2} w_2.$$

(Man kann per Induktion über die Länge der Ableitung von w_i aus S_i in $\mathcal{G}$ beweisen, dass es sich schon um eine Ableitung in $\mathcal{G}_i$ handeln muss.)

Beweis III

Also $w_1 \in L(\mathcal{G}_1) = L_1$ und $w_2 \in L(\mathcal{G}_2) = L_2$ und damit $w = w_1 w_2 \in L_1 L_2$.

„$\supseteq$“: Sei $w \in L_1 L_2$, etwa $w = w_1 w_2$ mit $w_1 \in L_1$ und $w_2 \in L_2$. Dann gilt

$$S_1 \stackrel{*}{\to}_{\mathcal{G}_1} w_1 \quad \text{und} \quad S_2 \stackrel{*}{\to}_{\mathcal{G}_2} w_2.$$

und damit auch

$$S_1 \stackrel{*}{\to}_{\mathcal{G}} w_1 \quad \text{und} \quad S_2 \stackrel{*}{\to}_{\mathcal{G}} w_2,$$

weil $P_1, P_2 \subseteq P$. Weil $S \to_{\mathcal{G}} S_1 S_2$ gilt nach dem Kombinationslemma 6.9

$$S \to_{\mathcal{G}} S_1 S_2 \stackrel{*}{\to}_{\mathcal{G}} w_1 w_2 = w.$$

Also $w \in L(\mathcal{G})$.

Iteration: Übung.

Beispiel 6.46

Betrachte die Sprachen

$$L_1 = \{a^m b^m c^n \mid m, n \geq 0\},$$
$$L_2 = \{a^m b^n c^n \mid m, n \geq 0\}.$$

Beide Sprachen sind kontextfrei:

- L_1 wird erzeugt von der kontextfreien Grammatik

$$S \to AC, \quad A \to aAb \mid \varepsilon, \quad C \to cC \mid \varepsilon.$$

- L_2 wird erzeugt von der kontextfreien Grammatik

$$S \to AB, \quad A \to aA \mid \varepsilon, \quad B \to bBc \mid \varepsilon.$$

Aber es gilt

$$L_1 \cap L_2 = \{a^n b^n c^n \mid n \in \mathbb{N}\},$$

und nach Beispiel 6.43 ist diese Sprache nicht kontextfrei.

Bemerkung 6.47

Weil die Klasse der kontextfreien Sprachen abgeschlossen ist unter Vereinigung, aber nicht unter Durchschnitt, kann sie nicht unter Komplement abgeschlossen sein, denn Durchschnitt lässt sich mit Hilfe der de Morgan'schen Regel aus Vereinigung und Komplement darstellen.

Beispiel 6.48

Nach Beispiel 6.44 ist die Sprache $L_W = \{ww \mid w \in \{a, b\}^*\}$ nicht kontextfrei.

Folgende kontextfreie Grammatik erzeugt $\overline{L_W}$ (Beweis als Übung).

$$S \to AB \mid BA \mid A \mid B$$
$$A \to CAC \mid a$$
$$B \to CBC \mid b$$
$$C \to a \mid b.$$

Also ist $\overline{L_W}$ kontextfrei, aber $\overline{\overline{L_W}} = L_W$ nicht.