Reguläre Grammatiken

Definition:

Sei $G = (N, \Sigma, P, S)$ eine KFG, in der die Regelmenge P wie folgt eingeschränkt ist:

1. **Alle Regeln in P haben die Form (a) oder die Form (b):**

   (a) $A \to xB$, mit $x \in \Sigma^*$ und $B \in N$.

   (b) $A \to x$, mit $x \in \Sigma^*$.

   In diesem Falle heißt G eine rechtslineare Grammatik.

2. **Alle Regeln in P haben die Form (a) oder die Form (b):**

   (a) $A \to Bx$, mit $x \in \Sigma^*$ und $B \in N$.

   (b) $A \to x$, mit $x \in \Sigma^*$.

   In diesem Falle heißt G eine linkslineare Grammatik.

3. **Eine Grammatik heißt regulär oder Chomsky-Typ 3 Grammatik, wenn sie entweder rechtslinear oder linkslinear ist.**

Reguläre Grammatiken: Beispiele

Die Grammatiken G_1, G_2, G_3 erzeugen die Menge der Binärworte über dem Alphabet {0,1}:

- **G_1 : P={S → 0 | 1 | 0S | 1S} ist rechtslinear**
- **G_2 : P={S → 0 | 1 | S0 | S1} ist linkslinear**
- **G_3 : P={S → Z | ZS, Z → 0 | 1} ist nicht regulär**

Es gibt auch Sprachen, für die es keine reguläre Grammatik gibt:

- **Es gibt keine reguläre Grammatik, die die Sprache $L = \{a^n b^n \mid n \geq 1\}$ erzeugt (wird erzeugt von kontextfreier Grammatik mit Regeln P={S → aSb | ab}).**

Reguläre Grammatik: Beispiel Identifier

Betrachten wir folgende Sprache gegeben durch

$$L = \{ w \mid w = bX, X \in \{b, z\}^* \}$$

wobei ´b´ für einen beliebigen Buchstaben und ´z´ für eine beliebige Ziffer steht, i.e. ein Identifier der mit einem Buchstaben beginnen muss gefolgt von einer beliebigen Folge von Buchstaben und Ziffern.

Die reguläre (rechtslineare) Grammatik G ist gegeben durch:

G = ({S, A}, {b, z}, P, S) mit
P = { $S \rightarrow bA$, $A \rightarrow bA \mid zA \mid \varepsilon$ }

Das Wort bzzb läßt sich wie folgt in G ableiten:

$$S \Rightarrow bA \Rightarrow bzA \Rightarrow bzzA \Rightarrow bzzbA \Rightarrow bzzb$$

Ableitungen in einer rechtslinearen Grammatik

Sei $G = (N, \Sigma, P, S)$ rechtslinear. Alle Ableitungen in G haben folgende Struktur (für linkslineare Grammatik analog):

- **Fall 1: Es gibt eine Regel $S \to x$ mit $x \in \Sigma^*$. Dann ist $x \in L(G)$ und $S \Rightarrow x$ die zugehörige Ableitung.**

- **Fall 2: Es wird eine Folge von $n \geq 2$ Ableitungsschritten durchgeführt. Jede Satzform mit Ausnahme der letzten enthält genau ein Nichtterminalsymbol als letztes Zeichen, d.h.**

$$S \Rightarrow x_0A_1 \Rightarrow x_0x_1A_2 \Rightarrow \ \ldots \Rightarrow x_0x_1 \ldots x_{n-1}A_n \Rightarrow x_0x_1 \ldots x_{n-1}x_n.$$

wobei

- $A_i \in N$, $1 \leq i \leq n$
- $x_i \in \Sigma^*$, $0 \leq i \leq n$
- $S \to x_0A_1 \in P$
- $A_i \to x_iA_{i+1} \in P$ für alle i, $1 \leq i \leq n$, und $A_n \to x_n \in P$

Reguläre Ausdrücke

Regulärer Ausdruck:

- eine weitere Möglichkeit Sprachen zu definieren
- gleichmächtig wie reguläre Grammatiken
- definieren reguläre Sprachen

Vorteil regulärer Ausdrücke:

- deklarative Beschreibungsweise, die leicht nachvollziehbar ist
- Eingabesprache von vielen Tools wie Editoren, awk, grep, Scanner Generatoren, etc.

Operationen auf Mengen

Zur Definition von regulären Ausdrücken benötigen wir folgende Definitionen.

Im folgenden sei Σ ein Alphabet.

Defintion: **Seien $L_1, L_2 \subseteq \Sigma^*$. Die Konkatenation von L_1 und L_2, geschrieben in der Form $L_1.L_2$ oder L_1L_2, ist wie folgt definiert:**

$$L_1L_2 := \{x_1x_2 \mid x_1 \in L_1 , x_2 \in L_2 \}$$

Die Zeichenketten in L_1L_2 werden gebildet, indem man an eine aus L_1 gewählte Zeichenkette eine Zeichenkette aus L_2 anhängt und das für alle möglichen Kombinationen.

Beispiel: **Sei $L_1 = \{10,1\}$ und $L_2 = \{011,11,\varepsilon\}$.**

Dann ist $L_1L_2 = \{10011, 1011, 10, 111, 1\}$

Defintion: Sei $L \subseteq \Sigma^*$.

(1) $L^0 = \{\varepsilon\}$ (unabhängig von L), $L^1 = L$

(2) $L^i = LL^{i-1}$ für alle $i \geq 1$ (Konkatenation von i Kopien von L)

(3) $L^* = \bigcup_{i=0}^{\infty} L^i$ Kleene Hülle (oder einfach Hülle)

(4) $L^+ = \bigcup_{i=1}^{\infty} L^i$ positive Hülle

Beispiel: Sei L = {0,11}.

$L^0 = \{\varepsilon\}$, $L^1 = \{0,11\}$, $L^2 = \{00,011,110,1111\}$

$L^* = \{\varepsilon, 0, 11, 00, 011, 110, 1111, \ldots\}$

Definition: Reguläre Ausdrücke

Rekursive Definition von regulären Ausdrücken über Alphabet Σ mit den entsprechenden repräsentierten Mengen:

(1) $\emptyset$ ist ein regulärer Ausdruck und bezeichnet die leere Menge { }.
(2) ε ist ein regulärer Ausdruck und bezeichnet die Menge $\{\varepsilon\}$.
(3) Für jedes $a \in \Sigma$ gilt: **a** ist ein regulärer Ausdruck und bezeichnet die Menge {a}, i.e. L(**a**)={a}.

(4) Seien α, β reguläre Ausdrücke mit den dazugehörenden Mengen $\mathbf{L}(\alpha)$ und $\mathbf{L}(\beta)$, dann gilt:

(a) $\alpha \mid \beta$ ist ein regulärer Ausdruck und bezeichnet $\mathbf{L}(\alpha) \cup \mathbf{L}(\beta)$ (Vereinigung)
(alternative Schreibweise: $\alpha + \beta$)
(b) $\alpha\,\beta$ ist ein regulärer Ausdruck und bezeichnet $\mathbf{L}(\alpha)\,\mathbf{L}(\beta)$ (Konkatenation)

(c) α^* ist ein regulärer Ausdruck und bezeichnet $(\mathbf{L}(\alpha))^*$ (Stern)

(d) (α) ist ein regulärer Ausdruck und bezeichnet $\mathbf{L}(\alpha)$. (Klammern)

Reguläre Ausdrücke:

Priorität der Operatoren:

1. **Stern** *(höchste Priorität)*
2. **Konkatenation**
3. **Vereinigung**

Klammern können verwendet werden, um gewünschte Prioritäten festzulegen; redundante Klammern können gesetzt werden.

Beispiel: **01*** | **0** steht für: (**0** (**1**)*) | **0**

Das vereinfachende hochgestellte **+** ist erlaubt:

- Der Ausdruck $\alpha\alpha^*$ entspricht dem Ausdruck α^+ .

Beispiele für reguläre Ausdrücke und Mengen:

$$\Sigma = \{0, 1\}$$

regulärer Ausdruck	bezeichnete Menge
0	$\{0\}$
$(0 \mid 1)^*$	$\{0,1\}^*$ oder $\{ x \mid x \in \Sigma^* \}$
$0(0\mid1)^*1$	$\{0\}\,\{0,1\}^*\,\{1\}$ bzw. $\{ 0x1 \mid x \in \Sigma^* \}$
$(O\mid1)^*011$	$\{ x011 \mid x \in \Sigma^* \}$
0^*1^*	$\{ 0^i1^j \mid i, j \geq 0 \}$
00^*11^*	$\{ 0^i1^j \mid i, j \geq 1 \}$
0^+1^+	$\{ 0^i1^j \mid i, j \geq 1 \}$

Beispiel:

$L(a^*(a|b)) = L(a^*)\ L(a|b) =$
$\{\varepsilon, a, aa, aaa, \ldots\}\ \{a,b\} =$
$\{a, aa, aaa, \ldots, b, ab, aab, \ldots\}$

Beispiel: $r = (a|b)^*(a|bb)$

Intuitive Analyse: (a|b)* – beliebige Zeichenketten von a und b
(a|bb) – entweder a oder bb
Deshalb ist L(r) die Menge aller Zeichenketten
von {a,b} inkl. ε terminiert mit a oder bb.

$L(r) = \{a, bb, aa, abb, ba, bbb, \ldots\}$

Beispiel: $r = (aa)^*(bb)^*b$

$L(r) = \{ a^{2m}b^{2n+1} \mid m \geq 0, n \geq 0 \}$

Beispiel:

Gegeben sei der reguläre Ausdruck 0(10)*. Die dadurch definierte Sprache wird von folgender rechtslinearen Grammatik erzeugt:

G = ({S,A}, {0,1}, P, S) mit

P = { S $\rightarrow$ 0A , A $\rightarrow$ 10A | ε }

bzw. von folgender linkslinearen Grammatik:

G = ({S}, {0,1}, P, S) mit

P = { S $\rightarrow$ S10 | 0 }