Regulare Ausdrücke und Rechtslineare Grammatiken

Reguläre Ausrücke

- Alphabet $Z = \{|, (,), *, \emptyset\}$
- Alphabet A, welches keine Elemente aus Z enthält

Ein Regulärer Ausdruck über A ist eine Zeichenfolge über dem Alphabet $A \cup Z$, definiert als:

- $\emptyset$ ist ein regulärer Ausdruck
- Alle Elemente in A sind reguläre Ausrücke
- Wenn R, R_1 und R_2 regulärer Ausdrücke sind, dann sind auch reguläre Ausrücke:
  - $(R*)$
  - $(R_1|R_2)$
  - (R_1R_2)

(sonst ist nichts ein Regulärer Ausdruck)

Beschreibung durch Kontextfreie Grammatik

Die Sprache aller syntaktisch korrekten regulären Ausrücke kann auch mit einer Kontextfreien Grammatik beschrieben werden:

$G = (\{R\}, \{|, (,), *, \emptyset\} \cup A, R, P)$

$P = \{R \longrightarrow \emptyset, R \longrightarrow (R \,|\, R), R \longrightarrow (RR), R \longrightarrow (R*)\} \cup \{R \longrightarrow x \,|\, x \in A\}$

Klammerregeln

Das Weglassen von Klammern ist erlaubt, und es gilt “Stern vor Punkt” und “Punkt vor Strichrechnung”.

beschriebene Sprache

Reguläre Ausdrücke werden genutzt, um formale Sprachen zu definieren. $\langle R \rangle$ ist die von einem Regulären Ausdruck R definierte Sprache und ist definiert als:

- $\langle \emptyset \rangle = \{\}$ (leere Menge)
- Für $x \in A$ ist $\langle x \rangle = \{x\}$
- Seien R_1 und R_2 reguläre Ausdräcke, dann gilt:
  - $\langle R_1 | R_2 \rangle = \langle R_1 \rangle \cup \langle R_2 \rangle$
  - $\langle R_1 R_2 \rangle = \langle R_1 \rangle \cdot \langle R_2 \rangle$
- Sei R ein regulärer Ausdruck, dann ist $\langle R* \rangle = \langle R \rangle^*$

(die gleiche Formale Sprache kann durch verschieden reguläre Ausdrücke beschrieben werden.)

Beispiele

ohne Klammerregeln:

- $\emptyset$
- (ab)
- $(\emptyset \,|\, b)$
- $((ab)(aa))$
- $(((((ab)b)*)*)|(\emptyset*))$

mit Klammerregeln:

- ab
- $a**$
- $(abb)** \,|\, \emptyset*$
- $(bababaaaaaaaab)*(aaaaabaaa \,|\, abababaab)(abab)*$

Rechtslineare Grammatiken

eine Rechtslineare Grammatik ist eine kontextfreie Grammatik mit gewissen einschränkungen, sodass die erzeugten Sprachen immer regulär sind. Es kann genau jede reguläre Sprache mit einer Rechtslinearen Grammatik erzeugt werden.

Eine kontextfreie Grammatik $G = (N, T, S, P)$ ist eine rechtslineare Grammatik, wenn jede Produktion eine dieser Formen hat:

- $X \longrightarrow w$
- $X \longrightarrow wY$

mit $w \in T^*$ und $X, Y \in N$

("Jede Produktion darf maximal ein neues Nichtterminalsymbol einführen, und das nur ganz am Ende")

Zusammenhänge von regulären Sprachen

Für jede formale Sprache sind äquivalent:

1. L kann von einem endlichen Akzeptor erkannt werden.$

2. L kann durch einen regulären Ausdruck beschrieben werden.
3. L kann von einer rechtslinearen Grammatik erzeugt werden.

Sprachen mit diesen Eigenschaft werden auch *reguläre Sprachen* genannt. Außerdem ist jede Reguläre Sprache automatisch eine kontextfreie Sprache, da rechtslineare Grammatiken kontextfreie Grammatiken sind.

alternative benennung

- Kontextfreie Grammatiken = *Typ-2-Grammatiken*
- Rechtslineare Grammatiken = *Typ-3-Grammatiken*
- (in späteren Vorlesungen noch mehr)

wenn eine formale Sprache L von einer *Typ-i-Grammatik* erzeugt wird, nennt man die Sprache auch eine *Typ-i-Sprache*.

Kantorowitsch-Bäume

Reguläre Ausrücke können auch als *Kantorowitsch-Bäume* (Regex-Bäume) dargestellt werden.

Beispiel zum regulären Ausdruck $((b \mid \emptyset)a)(b*)$:

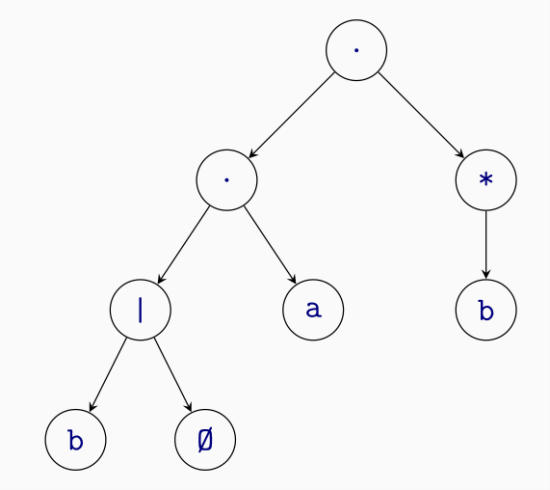

Abbildung 1: Regex-Graph Beispiel