Grundbegriffe der Informatik

Einheit 6: formale Sprachen

Thomas Worsch

Karlsruher Institut für Technologie, Fakultät für Informatik

Wintersemester 2009/2010

Überblick

Sprachen

- natürliche Sprache:
  - Aussprache
  - Stil, z. B. Wortwahl und Satzbau
  - Welche Formulierungen sind syntaktisch korrekt?
  - Ist und syntaktische welcher korrekt nicht?
- Informatik:
  - Sprachen, die nicht natürlich sind:
    - Programmiersprachen
    - Aufbau von Emails, WWW-Seiten, ...
    - Eingabedateien für ...
  - Syntax
    - Wie spezifiziert man, was korrekt ist?
    - Wie überprüft man, ob etwas korrekt ist?
  - Semantik
    - Wie definiert man, was syntaktisch korrekte Gebilde bedeuten?
    - darum kümmern wir uns später

Überblick

- Alphabet A gegeben
- Eine formale Sprache (über einem Alphabet A) ist eine Teilmenge $L \subseteq A^*$.
- im Zusammenhang mit syntaktischer Korrektheit:
  - formale Sprache L der syntaktisch korrekten Gebilde
  - syntaktisch falsche Gebilde gehören *nicht* zu L

Beispiele:

- $A = \{\texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{-}\}$.
  formale Sprache der Dezimaldarstellungen ganzer Zahlen
  - enthält z.B. `1`, `-22` und `192837465`,
  - aber nicht `2-3---41`.
- formale Sprache der syntaktisch korrekten Java-Programme
  - enthält alle Java-Programme
  - enthält zum Beispiel *nicht*: `[2] class int)(`

Überblick

Produkt oder Konkatenation formaler Sprachen

- kennen schon Konkatenation von Wörtern
- Produkt der Sprachen L_1 und L_2:

$$L_1 \cdot L_2 = \{w_1 w_2 \mid w_1 \in L_1 \wedge w_2 \in L_2\}$$

- **Lemma.** Für jede formale Sprache L ist

$$L \cdot \{\varepsilon\} = L = \{\varepsilon\} \cdot L \,.$$

Produkt oder Konkatenation formaler Sprachen

- kennen schon Konkatenation von Wörtern
- Produkt der Sprachen L_1 und L_2:

$$L_1 \cdot L_2 = \{w_1 w_2 \mid w_1 \in L_1 \land w_2 \in L_2\}$$

- **Lemma.** Für jede formale Sprache L ist

$$L \cdot \{\varepsilon\} = L = \{\varepsilon\} \cdot L \,.$$

Beweis des Lemmas.
Einfaches Nachrechnen:

$$
\begin{aligned}
L \cdot \{\varepsilon\} &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in \{\varepsilon\}\} \\
&= \{w_1 w_2 \mid w_1 \in L \wedge w_2 = \varepsilon\} \\
&= \{w_1 \varepsilon \mid w_1 \in L\} \\
&= \{w_1 \mid w_1 \in L\} \\
&= L
\end{aligned}
$$

Analog zeigt man $L = \{\varepsilon\} \cdot L$.

Beweis des Lemmas.
Einfaches Nachrechnen:

$$\begin{aligned} L \cdot \{\varepsilon\} &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in \{\varepsilon\}\} \\ &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 = \varepsilon\} \\ &= \{w_1 \varepsilon \mid w_1 \in L\} \\ &= \{w_1 \mid w_1 \in L\} \\ &= L \end{aligned}$$

Analog zeigt man $L = \{\varepsilon\} \cdot L$.

Beweis des Lemmas.
Einfaches Nachrechnen:

$$\begin{aligned} L \cdot \{\varepsilon\} &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in \{\varepsilon\}\} \\ &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 = \varepsilon\} \\ &= \{w_1 \varepsilon \mid w_1 \in L\} \\ &= \{w_1 \mid w_1 \in L\} \\ &= L \end{aligned}$$

Analog zeigt man $L = \{\varepsilon\} \cdot L$.

Beweis des Lemmas.
Einfaches Nachrechnen:

$$\begin{aligned} L \cdot \{\varepsilon\} &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in \{\varepsilon\}\} \\ &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 = \varepsilon\} \\ &= \{w_1 \varepsilon \mid w_1 \in L\} \\ &= \{w_1 \mid w_1 \in L\} \\ &= L \end{aligned}$$

Analog zeigt man $L = \{\varepsilon\} \cdot L$.

Beweis des Lemmas.
Einfaches Nachrechnen:

$$\begin{aligned} L \cdot \{\varepsilon\} &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in \{\varepsilon\}\} \\ &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 = \varepsilon\} \\ &= \{w_1 \varepsilon \mid w_1 \in L\} \\ &= \{w_1 \mid w_1 \in L\} \\ &= L \end{aligned}$$

Analog zeigt man $L = \{\varepsilon\} \cdot L$.

Beweis des Lemmas.
Einfaches Nachrechnen:

$$\begin{aligned}
L \cdot \{\varepsilon\} &= \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in \{\varepsilon\}\} \\
&= \{w_1 w_2 \mid w_1 \in L \wedge w_2 = \varepsilon\} \\
&= \{w_1 \varepsilon \mid w_1 \in L\} \\
&= \{w_1 \mid w_1 \in L\} \\
&= L
\end{aligned}$$

Analog zeigt man $L = \{\varepsilon\} \cdot L$.

Potenzen von Sprachen

- wir wollen Potenzen L^k formaler Sprachen definieren
- „Problem“: Was soll L^0 sein?
- Definiere:

$$L^0 = \{\varepsilon\}$$
$$\forall k \in \mathbb{N}_0 : \; L^{k+1} = L \cdot L^k$$

- Einfaches Nachrechnen ergibt z. B.:

$$L^1 = L$$
$$L^2 = L \cdot L$$
$$L^3 = L \cdot L \cdot L$$

- Genau genommen: $L^3 = L \cdot (L \cdot L)$, aber:
  Konkatenation von Sprachen ist eine assoziative Operation.

Potenzen von Sprachen

- wir wollen Potenzen L^k formaler Sprachen definieren
- „Problem“: Was soll L^0 sein?
- Definiere:

$$L^0 = \{\varepsilon\}$$
$$\forall k \in \mathbb{N}_0 : \; L^{k+1} = L \cdot L^k$$

- Einfaches Nachrechnen ergibt z. B.:

$$L^1 = L$$
$$L^2 = L \cdot L$$
$$L^3 = L \cdot L \cdot L$$

- Genau genommen: $L^3 = L \cdot (L \cdot L)$, aber:
  Konkatenation von Sprachen ist eine assoziative Operation.

Potenzen von Sprachen

- wir wollen Potenzen L^k formaler Sprachen definieren
- „Problem“: Was soll L^0 sein?
- Definiere:

$$L^0 = \{\varepsilon\}$$
$$\forall k \in \mathbb{N}_0 : \; L^{k+1} = L \cdot L^k$$

- Einfaches Nachrechnen ergibt z. B.:

$$L^1 = L$$
$$L^2 = L \cdot L$$
$$L^3 = L \cdot L \cdot L$$

- Genau genommen: $L^3 = L \cdot (L \cdot L)$, aber:
Konkatenation von Sprachen ist eine assoziative Operation.

- wir wollen Potenzen L^k formaler Sprachen definieren
- „Problem“: Was soll L^0 sein?
- Definiere:

$$L^0 = \{\varepsilon\}$$
$$\forall k \in \mathbb{N}_0 : \; L^{k+1} = L \cdot L^k$$

- Einfaches Nachrechnen ergibt z. B.:

$$L^1 = L$$
$$L^2 = L \cdot L$$
$$L^3 = L \cdot L \cdot L$$

- Genau genommen: $L^3 = L \cdot (L \cdot L)$, aber:
  Konkatenation von Sprachen ist eine assoziative Operation.

Beispiele für Produkt von Sprachen (1)

- $L = \{\mathtt{aa}, \mathtt{b}\}$
- Dann ist

$$
\begin{aligned}
L^0 &= \{\varepsilon\} \\
L^1 &= \{\mathtt{aa}, \mathtt{b}\} \\
L^2 &= \{\mathtt{aa}, \mathtt{b}\} \cdot \{\mathtt{aa}, \mathtt{b}\} = \{\mathtt{aa} \cdot \mathtt{aa}, \mathtt{aa} \cdot \mathtt{b}, \mathtt{b} \cdot \mathtt{aa}, \mathtt{b} \cdot \mathtt{b}\} \\
&= \{\mathtt{aaaa}, \mathtt{aab}, \mathtt{baa}, \mathtt{bb}\} \\
L^3 &= \ \{\mathtt{aa} \cdot \mathtt{aa} \cdot \mathtt{aa},\ \mathtt{aa} \cdot \mathtt{aa} \cdot \mathtt{b},\ \mathtt{aa} \cdot \mathtt{b} \cdot \mathtt{aa},\ \mathtt{aa} \cdot \mathtt{b} \cdot \mathtt{b}, \\
&\qquad \mathtt{b} \cdot \mathtt{aa} \cdot \mathtt{aa},\ \mathtt{b} \cdot \mathtt{aa} \cdot \mathtt{b},\ \mathtt{b} \cdot \mathtt{b} \cdot \mathtt{aa},\ \mathtt{b} \cdot \mathtt{b} \cdot \mathtt{b}\} \\
&= \{\mathtt{aaaaaa}, \mathtt{aaaab}, \mathtt{aabaa}, \mathtt{aabb}, \mathtt{baaaa}, \mathtt{baab}, \mathtt{bbaa}, \mathtt{bbb}\}
\end{aligned}
$$

- Sei

$$L = \{\mathtt{a}^n\mathtt{b}^n \mid n \in \mathbb{N}_+\} \ ,$$

also sozusagen (immer diese Pünktchen ...)

$$L = \{\mathtt{ab}, \mathtt{aabb}, \mathtt{aaabbb}, \mathtt{aaaabbbb}, \dots\} \ .$$

- Was ist $L^2 = L \cdot L$?
  -

$$\begin{aligned} L^2 = & \quad \{\mathtt{ab}\cdot\mathtt{ab}, \mathtt{ab}\cdot\mathtt{aabb}, \mathtt{ab}\cdot\mathtt{aaabbb}, \dots\} \\ & \cup \{\mathtt{aabb}\cdot\mathtt{ab}, \mathtt{aabb}\cdot\mathtt{aabb}, \mathtt{aabb}\cdot\mathtt{aaabbb}, \dots\} \\ & \cup \{\mathtt{aaabbb}\cdot\mathtt{ab}, \mathtt{aaabbb}\cdot\mathtt{aabb}, \mathtt{aaabbb}\cdot\mathtt{aaabbb}, \dots\} \\ & \quad \vdots \end{aligned}$$

  - Mit anderen Worten

$$L^2 = \{\mathtt{a}^{n_1}\mathtt{b}^{n_1}\mathtt{a}^{n_2}\mathtt{b}^{n_2} \mid n_1 \in \mathbb{N}_+ \wedge n_2 \in \mathbb{N}_+\} \ .$$

  - Beachte: die Exponenten n_1 „vorne“ und n_2 „hinten“ heißen verschieden.

- Sei
$$L = \{\mathtt{a}^n\mathtt{b}^n \mid n \in \mathbb{N}_+\} ,$$
also sozusagen (immer diese Pünktchen ...)
$$L = \{\mathtt{ab}, \mathtt{aabb}, \mathtt{aaabbb}, \mathtt{aaaabbbb}, \dots\} .$$
- Was ist $L^2 = L \cdot L$?
  -
$$\begin{aligned} L^2 = \quad & \{\mathtt{ab} \cdot \mathtt{ab}, \mathtt{ab} \cdot \mathtt{aabb}, \mathtt{ab} \cdot \mathtt{aaabbb}, \dots\} \\ \cup\ & \{\mathtt{aabb} \cdot \mathtt{ab}, \mathtt{aabb} \cdot \mathtt{aabb}, \mathtt{aabb} \cdot \mathtt{aaabbb}, \dots\} \\ \cup\ & \{\mathtt{aaabbb} \cdot \mathtt{ab}, \mathtt{aaabbb} \cdot \mathtt{aabb}, \mathtt{aaabbb} \cdot \mathtt{aaabbb}, \dots\} \\ & \vdots \end{aligned}$$
  - Mit anderen Worten
$$L^2 = \{\mathtt{a}^{n_1}\mathtt{b}^{n_1}\mathtt{a}^{n_2}\mathtt{b}^{n_2} \mid n_1 \in \mathbb{N}_+ \wedge n_2 \in \mathbb{N}_+\} .$$
  - Beachte: die Exponenten n_1 „vorne“ und n_2 „hinten“ heißen verschieden.

Potenzen mehrfach definiert

- für Alphabet A und für $i \in \mathbb{N}_0$ hatten wir schon Potenzen A^i definiert.
- Jedes Alphabet A kann man als formale Sprache L_A auffassen (enthält alle Wörter der Länge 1)
- Man mache sich klar:
  A^i ist („im Wesentlichen") das Gleiche wie L_A^i.

Überblick

Konkatenationsabschluss von L

- bei Alphabeten schon gesehen: $A^* = \bigcup_{i=0}^{\infty} A^i$.
- der Konkatenationsabschluss L^* von L ist

$$L^* = \bigcup_{i=0}^{\infty} L^i$$

- der ε-freie Konkatenationsabschluss L^+ von L ist

$$L^+ = \bigcup_{i=1}^{\infty} L^i$$

- Man sieht:

$$L^* = L^0 \cup L^+ \ .$$

Beispiele für Konkatenationsabschluss

- Es sei wieder $L = \{\mathtt{a}^n\mathtt{b}^n \mid n \in \mathbb{N}_+\}$.
- haben schon gesehen:

$$L^2 = \{\mathtt{a}^{n_1}\mathtt{b}^{n_1}\mathtt{a}^{n_2}\mathtt{b}^{n_2} \mid n_1 \in \mathbb{N}_+ \land n_2 \in \mathbb{N}_+\} .$$

- analog

$$L^3 = \{\mathtt{a}^{n_1}\mathtt{b}^{n_1}\mathtt{a}^{n_2}\mathtt{b}^{n_2}\mathtt{a}^{n_3}\mathtt{b}^{n_3} \mid n_1 \in \mathbb{N}_+ \land n_2 \in \mathbb{N}_+ \land n_3 \in \mathbb{N}_+\} .$$

- wir erlauben uns Pünktchen . . . :

$$L^i = \{\mathtt{a}^{n_1}\mathtt{b}^{n_1} \cdots \mathtt{a}^{n_i}\mathtt{b}^{n_i} \mid n_1, \ldots, n_i \in \mathbb{N}_+\} .$$

- Dann kann man für L^+ notieren:

$$L^+ = \{\mathtt{a}^{n_1}\mathtt{b}^{n_1} \cdots \mathtt{a}^{n_i}\mathtt{b}^{n_i} \mid i \in \mathbb{N}_+ \land n_1, \ldots, n_i \in \mathbb{N}_+\} .$$

- Sie merken (hoffentlich): L^+ und L^* sind *präziser und kürzer* hinzuschreiben als viele Pünktchen.

- Die Bezeichnung ε-freier Konkatenationsabschluss für L^+ ist irreführend.
  - Wie steht es um das leere Wort bei L^+ und L^*?
  - Klar ist:

$$\varepsilon \in L^0 \subseteq L^*$$

  - Aber: $L = L^1 \subseteq L^+$,
    wenn also $\varepsilon \in L$, dann auch $\varepsilon \in L^+$.

- Beachte

$$\{\}^* = \{\varepsilon\}$$

Zwei Warnungen

- Die Bezeichnung ε-freier Konkatenationsabschluss für L^+ ist irreführend.
  - Wie steht es um das leere Wort bei L^+ und L^*?
  - Klar ist:

$$\varepsilon \in L^0 \subseteq L^*$$

  - Aber: $L = L^1 \subseteq L^+$,
    wenn also $\varepsilon \in L$, dann auch $\varepsilon \in L^+$.

- Beachte

$$\{\}^* = \{\varepsilon\}$$

Was ist wichtig

Das sollten Sie mitnehmen:

- was formale Sprachen sind,
- wie ihr Produkt definiert ist und
- wie Konkatenationsabschluss und ε-freier Konkatenationsabschluss definiert sind.

Das sollten Sie üben:

- Erkennen von Strukturen der Form L^*, L^+, L_1L_2
- Lesen von Ausdrücken der Form $(L_1^+L_2)^*$ usw.
- „Rechnen" mit formalen Sprachen