Formale Sprachen

Alphabete

Ein *Alphabet* ist eine endliche, nicht leere Menge von Symbolen. Üblicherweise wird ein Alphabet durch das Symbol Σ dargestellt. *Beispiele*

- $\Sigma = \{0, 1\}$ (binäres Alphabet)
- $\Sigma = \{\mathrm{a}, \mathrm{b}, \ldots, \mathrm{z}\}$ (lateinsche Kleinbuchstaben)
- Die Menge aller ASCII-Zeichen

Wörter

Ist Σ ein Alphabet und n eine natürliche Zahl, dann ist ein *Wort* w der Länge n eine *endliche* Folge $(a_1, a_2, \ldots, a_n)$ von Symbolen $a_i \in \Sigma$ $(i = 1, 2, \ldots, n)$. *Beispiele:*

- 011010 (Wort aus dem binären Alphabet)
- abcdda (Wort aus dem Kleinbuchstaben-Alphabet)

Das spezielle Wort, das aus keinem Symbol besteht, wird *das leere Wort* genannt und meist mit ε bezeichnet.

Die Länge eines Worts

Die Länge eines Worts w beschreibt man mit $|w|$ und die Häufigkeit, mit der das Zeichen x in w auftritt, mit $|w|_x$. *Beispiele:*

- $|01101| = 5$
- $|\varepsilon| = 0$
- $|01101|_1 = 3$
- $|\mathrm{ababb}|_\mathrm{c} = 0$

Potenzen eines Alphabets

Ist Σ ein Alphabet, so bezeichnet Σ^k die Menge aller Wörter der Länge k über Σ. Für $\Sigma = \{0, 1\}$ sind dies:

- $\Sigma^0 = \{\varepsilon\}$
- $\Sigma^1 = \{0, 1\}$
- $\Sigma^2 = \{00, 01, 10, 11\}$
- $\Sigma^3 = \{000, 001, 010, 100, 011, 101, 110, 111\}$
- ...

Beachte, dass es zwischen Σ (der Menge aller Zeichen) und Σ^1 (der Menge aller Wörter der Länge 1) einen subtilen formalen Unterschied gibt. (Vergleichbar mit dem Unterschied zwischen der Ziffer „3“ und der Zahl „3“)

Die Kleenesche Hülle (Kleene star)

Die Menge aller Wörter über einem Alphabet Σ wird mit Σ^* bezeichnet. Präziser:

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$$

In speziellen Situationen möchte man das leere Wort ausschliessen. Dann schreibt man:

$$\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$$

Die Bezeichnung geht auf einen der Begründer der theoretischen Informatik, Stephen Cole Kleene (1909–1994), zurück.

Konkatenation

Sind $x = (a_1, a_2, \dots, a_m)$ und $y = (b_1, b_2, \dots, b_n)$ zwei Wörter über einem Alphabet Σ der Längen m und n, so ist ihre *Konkatenation* (Aneinanderreihung) $x \circ y$ wie folgt definiert:

$$x \circ y = (a_1, a_2, \dots, a_m, b_1, b_2, \dots b_n)$$

Beispiele: $\Sigma = \{0, 1\}$, $x = 10110$, $y = 110$

- $x \circ y = 10110110$
- $y \circ x = 11010110$
- $y^3 = y \circ y \circ y = 110110110$

Anstelle von „$\circ$“ wird manchmal auch „ $\cdot$ “ verwendet.

Eigenschaften

- Die Konkatenation ist im Allgemeinen nicht kommutativ.

  (Gegenbeispiel: $10 \circ 11 = 1011 \neq 1101 = 11 \circ 10$)
- Es gilt: $|x \circ y| = |x| + |y| = m + n$.
- Für ein beliebiges Wort $w \in \Sigma^*$ gilt: $w \circ \varepsilon = \varepsilon \circ w = w$.
- Die Konkatenation ist assoziativ:

  Für alle u, v und $w \in \Sigma^*$ gilt: $(u \circ v) \circ w = u \circ (v \circ w)$

Da die Konkatenation häufig gebraucht wird und meist aus dem Kontext erkennbar ist, lässt man den Operator „$\circ$“ oft weg.

$xy \overset{\text{Def.}}{=} x \circ y$

Teilwörter

Es seien v und w Wörter über dem Alphabet Σ.

- v wird *Infix* (*Teilwort*) von w genannt, wenn es Wörter x und $y \in \Sigma^*$ gibt, so dass die Bedingung $x \circ v \circ y = w$ erfüllt ist.

- v wird *Präfix* von w genannt ($v \sqsubset w$), wenn es ein Wort $y \in \Sigma^*$ gibt, so dass die Bedingung $v \circ y = w$ erfüllt ist.

- v wird *Suffix* von w genannt ($v \sqsupset w$), wenn es ein Wort $x \in \Sigma^*$ gibt, so dass die Bedingung $x \circ v = w$ erfüllt ist.

Bemerkungen:

- Es ist auch $x = \varepsilon$ oder $y = \varepsilon$ möglich.
- Das leere Wort ε ist Infix, Präfix und Suffix von jedem Wort.

Sprachen

Ist Σ ein Alphabet, so wird eine Teilmenge $L \subset \Sigma^*$ als *(formale) Sprache* bezeichnet. Dazu einige Bemerkungen:

- Im Gegensatz zu natürlichen Sprachen haben die Wörter formaler Sprachen keine feste sprachliche Bedeutung.
- Es wird nicht verlangt, dass in einer Sprache alle Symbole aus Σ vorkommen. Auch $L_1 = \{1, 11, 111, 1111, \dots\}$ oder $L_2 = \{0, 00, 00000\}$ sind Sprachen über $\Sigma = \{0, 1\}$.
- Eine Sprache kann aus endlich oder unedlich vielen Wörtern bestehen. Hingegen muss das zugrunde liegegnde Alphabet *endlich* sein.
- Vorerst wird nicht verlangt, dass einer Sprache eine Art Regelmässigkeit („Grammatik“) zugrunde liegt.

Probleme

In der theoretischen Informatik bezeichnet der Begriff *Problem* die Frage, ob ein bestimmtes Wort $w \in \Sigma^*$ in einer gegebenen Sprache L enthalten ist.

Vieles von dem, was umgangssprachlich als „Problem“ oder „Aufgabe“ bezeichnet wird, lässt sich so darstellen. *Beispiel:*

Die Frage „Ist 833698888858693277417 eine Primzahl?“, lässt sich wie folgt als *Entscheidungsproblem* darstellen:

$\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$L = \{2, 3, 5, 7, 11, 13, \dots\}$ (Menge der Primzahlen)

$w = 833698888858693277417$

Gilt $w \in L$?

Vereinigung von Sprachen

Sind L_1 und L_2 zwei Sprachen, dann ist ihre *Vereinigung*

$$L_1 \cup L_2$$

die Sprache, die aus allen Wörtern besteht, die entweder in L_1 oder in L_2 oder in beiden Sprachen enthalten sind.

Beispiel: $L_1 = \{\varepsilon, 001, 10\}$, $L_2 = \{10, 111\}$

$\Rightarrow L_1 \cup L_2 = \{\varepsilon, 10, 001, 111\}$

Verkettung (Konkatenation)von Sprachen

Sind L_1 und L_2 zwei Sprachen, dann ist ihre *Verkettung* (*Konkatenation*)

$$L_1 \circ L_2 \quad \text{oder kürzer} \quad L_1 L_2$$

Die Sprache, die aus allen Wörtern $w_1 \circ w_2$ mit $w_1 \in L_1$ und $w_2 \in L_2$ besteht.

Beispiel: $L_1 = \{\varepsilon, \mathrm{a}\}$, $L_2 = \{\mathrm{b}, \mathrm{ba}\}$

- $L_1 L_2 = \{\mathrm{b}, \mathrm{ba}, \mathrm{ab}, \mathrm{aba}\}$
- $L_2 L_1 = \{\mathrm{b}, \mathrm{ba}, \mathrm{ba}, \mathrm{baa}\}$

Die Kleenesche Hülle

Ist L eine Sprache, so ist ihre *(Kleenesche) Hülle* die Sprache L^*, die aus der Vereinigung aller Verkettungen von L mit sich selbst besteht. Formal:

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \cup \ldots$$

wobei
$$\begin{aligned} L^0 &= \{\varepsilon\} \\ L^1 &= L \\ L^2 &= L \circ L \\ L^3 &= L \circ L \circ L \\ \ldots &= \ldots \end{aligned}$$

Beispiel: $L = \{\mathrm{a}, \mathrm{ab}\}$

- $L^0 = \{\varepsilon\}$
- $L^1 = \{\mathrm{a}, \mathrm{ab}\}$
- $L^2 = \{\mathrm{aa}, \mathrm{aab}, \mathrm{aba}, \mathrm{abab}\}$
- $L^3 = \{\mathrm{aaa}, \mathrm{aaab}, \mathrm{aaba}, \mathrm{aabab}, \mathrm{abaa}, \mathrm{abaab}, \mathrm{ababa}, \mathrm{ababab}\}$
- $L^4 = \ldots$

$L^* = \{\varepsilon, \mathrm{a}, \mathrm{ab}, \mathrm{aa}, \mathrm{aaa}, \mathrm{aab}, \mathrm{aba}, \mathrm{aaaa}, \mathrm{aaab}, \mathrm{aaba}, \mathrm{abaa}, \mathrm{abab}, \ldots\}$