

Theoretische Informatik

Teil 1

Alphabete, Wörter und Sprachen

Frühlingssemester 2019

L. Di Caro

D. Flumini

O. Stern



Definition (Alphabet)

Ein **Alphabet** ist eine **endliche, nichtleere** Menge von **Symbolen**.

Beispiele

- $\Sigma = \{a, b, c\}$ ist die Menge der drei Symbole a , b und c .
- $\Sigma = \{-, +, \cdot, :\}$ ist die Menge der Symbole für die Grundrechenarten.
- $\Sigma_{\text{Bool}} = \{0, 1\}$ ist das Boolesche Alphabet.
- $\Sigma_{\text{lat}} = \{a, b, c, \dots, z\}$ ist die Menge der lateinischen Kleinbuchstaben.
- \mathbb{N} ist kein Alphabet (unendliche Mächtigkeit)

Konventionen:

- Ein *Alphabet* wird häufig durch Σ und
- ein *Symbol* durch einen Kleinbuchstaben (vom Anfang des Alphabets: a, b, c, \dots) dargestellt.

Definition (Wort)

Ein **Wort** (**Zeichenreihe**, **String**) ist eine **endliche** Folge von Symbolen eines bestimmten Alphabets.

Beispiele

- abc ist ein Wort über dem Alphabet Σ_{lat} (oder über $\Sigma = \{a, b, c\}$).
- 100111 ist ein Wort über dem Alphabet $\{0, 1\}$.

Konventionen:

- Man sagt *über* dem Alphabet Σ .
- *Wörter* werden häufig durch Kleinbuchstaben (vom Ende des Alphabets: \dots, w, x, y, z) dargestellt.
- In einer Folge werden normalerweise die einzelnen Symbole durch Kommata getrennt. Wenn klar ist, was die einzelnen Elemente sind, lassen wir die Kommata der Einfachheit halber weg.

Definition (leeres Wort)

Das **leere Wort** ist ein Wort, das keine Symbole enthält.
Es wird durch das Symbol ε dargestellt und ist ein Wort über jedem Alphabet.

Definition (Länge eines Wortes)

Die **Länge eines Wortes** w ist die Länge des Wortes als Folge, also die Anzahl der Symbole der Folge.

Wir bezeichnen diese Länge mit $|w|$.

Beispiele

- $|abc| = 3$
- $|100111| = 6$
- $|\varepsilon| = 0$
- $|Informatik\ ist\ spannend| = 23$
(Leerzeichen sind auch Symbole!)

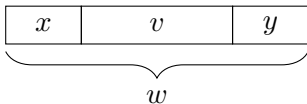
Definition (Teilwort)

Wir sagen, dass v ein **Teilwort** von w ist, wenn man w als

$$w = xvy$$

für beliebige Wörter x und y über Σ schreiben kann.

Ein **echtes Teilwort (Infix)** von w ist jedes Teilwort von w , das kürzer als w ist.



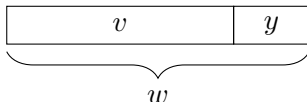
$[x \text{ oder } y \text{ nicht leer}]$

Beispiel

aba , aab oder $babaa$ sind echte Teilwörter von $babaab$.

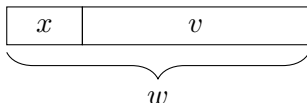
Definition (Präfix)

Ein Wort v ist ein **Präfix** von w , wenn $w = vy$ gilt für irgendein Wort y .



Definition (Suffix)

Ein Wort v ist ein **Suffix** von w , wenn $w = xv$ gilt für irgendein Wort x .



Definition (Menge aller Wörter der Länge k)

Die **Menge aller Wörter der Länge k** über einem Alphabet Σ wird mit Σ^k bezeichnet.

Beispiele

- Für $\Sigma = \{a, b, c\}$ ist $\Sigma^2 = \{aa, ab, ac, ba, bb, bc, ca, cb, cc\}$.
- Für $\{0, 1\}$ ist $\{0, 1\}^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$.

Anmerkung: Es gilt immer $\Sigma^0 = \{\varepsilon\}$ unabhängig von Σ .

Definition (Menge aller Wörter)

Die **Menge aller Wörter (Kleenesche Hülle)** über einem Alphabet Σ wird mit Σ^* bezeichnet.

$\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$ ist die **Menge aller nichtleeren Zeichenreihen** über einem Alphabet Σ .

Beispiel

Für $\{0, 1\}$ ist $\Sigma^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, \dots\}$.
Wörter aus $\{0, 1\}^*$ nennt man *Binärwörter*.

Eigenschaften:

- $\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$
- $\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$
- $\Sigma^* = \Sigma^+ \cup \Sigma^0 = \Sigma^+ \cup \{\varepsilon\}$

Definition (Konkatenation)

Seien x und y zwei beliebige Wörter. Dann steht $x \cdot y = xy$ für die **Konkatenation (Verkettung)** von x und y .

Beispiel

Seien $x = 01001$ und $y = 110$ zwei Wörter. Dann ist $xy = 01001110$ die Konkatenation der Wörter x und y .

Anmerkungen:

- Sei $x = a_1a_2 \dots a_i$ ein aus i Symbolen bestehendes Wort und sei $y = b_1b_2 \dots b_j$ ein aus j Symbolen bestehendes Wort. Dann hat das Wort $xy = a_1a_2 \dots a_ib_1b_2 \dots b_j$ die Länge $i + j$.
- Für jedes beliebige Wort w gilt $w\varepsilon = \varepsilon w = w$.

Definition (Wortpotenzen)

Sei x ein Wort über einem Alphabet Σ . Für alle $i \in \mathbb{N}$ sind **Wortpotenzen** wie folgt definiert:

$$x^0 = \varepsilon$$

$$x^1 = x$$

$$x^i = x \cdot x^{i-1}$$

Anmerkung: Mit Wortpotenzen kann man Wörter kürzer darstellen.

Beispiel

$$\begin{aligned}bbababababbbaaaabab &= b^2(ab)^4ba^4bab = b(ba)^4b^2a^3(ab)^2 \\abbabbabbabbabbabbabbabba &= a(bba)^9\end{aligned}$$

Definition (Sprache)

Eine Teilmenge $L \subseteq \Sigma^*$ von Wörtern über einem Alphabet Σ wird als **Sprache über Σ** bezeichnet.

Beispiele

- *Deutsch* ist eine Sprache über dem Alphabet der lateinischen Buchstaben, Leerzeichen, Kommata, Punkte, ...
- *Programmiersprachen* (wie *C*) sind Sprachen über dem Alphabet des ASCII-Zeichensatzes.
- $\{\epsilon, 10, 01, 1100, 1010, 1001, 0110, 0011, \dots\}$ ist die Sprache der Wörter über $\{0, 1\}$ mit der gleichen Anzahl von Nullen und Einsen.

Allgemein gilt:

- Wenn $\Sigma_1 \subseteq \Sigma_2$ gilt und L eine Sprache über Σ_1 ist, dann ist L auch eine Sprache über Σ_2 .
- Σ^* ist eine Sprache über jedem Alphabet Σ .
- $\{\} = \emptyset$ ist die *leere Sprache* für jedes Alphabet.
- $\{\varepsilon\}$ ist die Sprache, die aus dem leeren Wort ε besteht für jedes Alphabet Σ (Anmerkung: $\emptyset \neq \{\varepsilon\}$).

Anmerkungen:

- Sprachen können aus **unendlich** vielen Wörtern bestehen.
- Wörtern müssen aus einem festen, **endlichen** Alphabet gebildet werden.
- Wörter selber haben eine **endliche** Länge.

Mögliche Notationen von Sprachen:

- $L = \{\varepsilon, 10, 1100, 111000, \dots\}$
- L ist die Menge der Wörter über dem Alphabet $\{0, 1\}$, die aus n Einsen gefolgt von n Nullen besteht für eine natürliche Zahl n
- $L = \{w \mid w \text{ enthält } n \text{ Einsen gefolgt von } n \text{ Nullen für } n \in \mathbb{N}\}$
- $L = \{1^n 0^n \mid n \in \mathbb{N}\}$

Alle diese Notationen definieren dieselbe Sprache. In der Vorlesung bevorzugen wir die zwei letzten Notationen.

Konvention:

- $L = \{w \in \{0, 1\}^* \mid |w|_0 = 3\}$
 L ist die Menge der Wörter über dem Alphabet $\{0, 1\}$, die das Symbol 0 genau 3-mal beinhalten.

Definition (Konkatenation von Sprachen)

Sind $A \subset \Sigma^*$ und $B \subset \Gamma^*$ beliebige Sprachen, dann wird die Menge

$$AB = \{uv \mid u \in A \text{ und } v \in B\}$$

als **Konkatenation** von A und B bezeichnet.

Anmerkungen:

- Die Sprache AB besteht aus den Wörtern, die man (ohne Überschneidung) in ein Präfix aus A und ein Suffix aus B aufteilen kann.
- Ist A eine Sprache über Σ und ist B eine Sprache über Γ , dann ist AB eine Sprache über dem Alphabet $\Sigma \cup \Gamma$.

Beispiel

Die Sprachen A und B sind wie folgt gegeben:

- A enthält alle Binärwörter, die mit 0 enden.
- B enthält alle Binärwörter, die mit 0 beginnen.

Welche der folgenden Wörter sind Elemente von AB ?

ε ✗

0 ✗

01010 ✗

0000 ✓

00 ✓

1100110011 ✓

Wie kann man die Elemente von AB einfach beschreiben?

$AB =$ Binärwörter die 00 enthalten

Definition

Die **Kleenesche Hülle** A^* einer Sprache A ist durch

$$\{\varepsilon\} \cup A \cup AA \cup AAA \cup \dots$$

definiert.

Anmerkungen:

- Fasst man ein Alphabet Σ als Sprache über Σ auf, dann entspricht die Kleenesche Hülle von Σ gerade der Menge aller Wörter über Σ .
- Für alle Sprachen gilt $(A^*)^* = A^*$.

Beispiel

Welche Wörter gehören zur Sprache $\{aa, ab, ba, bb\}^*$?

ε ✓

ababa ✗

abbaba ✓

abaaabb ✗

abbaabba ✓

aaaaa ✗

Wie kann man die Elemente von $\{aa, ab, ba, bb\}^*$ einfach beschreiben?

Wörter aus $\{a, b\}^*$ mit einer **geraden Anzahl** Zeichen

Definition (Entscheidungsproblem)

Sei eine Sprache L über einem Alphabet Σ gegeben. Das **Entscheidungsproblem** (Σ, L) ist die folgende Berechnungsaufgabe:

Input: Eine Sprache L und ein Wort $x \in \Sigma^*$

Output: JA, falls $x \in L$, und
NEIN, falls $x \notin L$

Bedeutung dieser Definition: Modellierung von vielen alltäglichen Berechnungsproblemen in einer formalen Sprache.

Beispiel (Primzahltest)

Gegeben:

- Alphabet $\{0, 1\}$
- Sprache $L_p = \{w \mid w \text{ ist eine Primzahl}\}$
- Wörter x aus Σ^*

Der Test, ob x eine Primzahl darstellt, ist äquivalent zu der Entscheidung, ob $x \in L_p$ gilt.

Anmerkung: Die Entscheidung, ob ein Wort eine Primzahl ist oder nicht, ist für einige Wörter einfach: Alle Wörter $x = \dots 0$ stellen zum Beispiel gerade Zahlen dar, können also nicht prim sein (ausser die Zahl 2). Für viele andere Wörter ist diese Entscheidung erheblich aufwendiger.