

# Theoretische Informatik

Manuel Strenge

## Aplhabete

**Mächtig was ist die Decke der Menge: Unendlich=sehr mächtig**

Ein Alphabet ist eine endliche, nichtleere Menge von Symbolen

$\{:,2\} \rightarrow$  alphabet

$\{1,2,3\} \rightarrow$  alphabet

$\{1,2,3,\dots\} \rightarrow$  kein da nicht endlich

$\{a,\dots,b\} \rightarrow$  alphabet

$\{a,a,a\} \rightarrow$  ja

- $\Sigma = \{a, b, c\}$  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)

## Wort

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

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

## Leeres Wort

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

## Wörter

Die Länge eines Wortes  $w$  ist die Länge des Wortes als Folge, also die Anzahl der Symbole der Folge. Wir bezeichnen diese Länge mit  $|w|$ .

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

### Definition (Häufigkeit eines Symbols in einem Wort)

$|w|_x$  bezeichnet die absolute Häufigkeit eines Symbols  $x$  in einem Wortes  $w$ .

- $|abc|_a = 1$
- $|100111|_1 = 4$
- $|\varepsilon|_0 = 0$
- $|Informatik\ ist\ spannend|_n = 4$

### Definition (Spiegelung eines Wort)

Mit  $w^R$  wird das Spiegelwort zu  $w$  bezeichnet.

$$w^R = (x_1, x_2, \dots, x_n)^R = x_n, \dots, x_2, x_1$$

Es gilt  $|w| = |w^R|$  und  $|w|_x = |w^R|_x$  für alle  $x \in \Sigma$ . Wenn  $w = w^R$  gilt, dann bezeichnet man  $w$  als Palindrom.

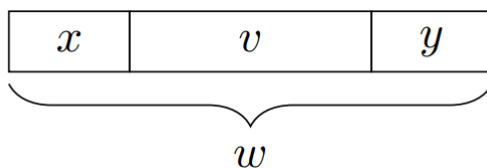
- $(abc)^R = cba$
- $(100111)^R = 111001$
- $\varepsilon^R = \varepsilon$
- $(Informatik\ ist\ spannend)^R = dnennaps\ tsi\ kitamrof\ nI$

### Definition (Teilwort)

Wir sagen, dass  $v$  ein Teilwort (Infix) von  $w$  ist, wenn man  $w$  als

$$w = xvy$$

für beliebige Wörter  $x$  und  $y$  über  $\Sigma$  schreiben kann



### Definition (echtes Teilwort)

Ein echtes Teilwort von  $w$  ist jedes Teilwort von  $w$ , das nicht identisch mit  $w$  ist (in diesem Falle ist  $x$  oder  $y$  nicht leer).

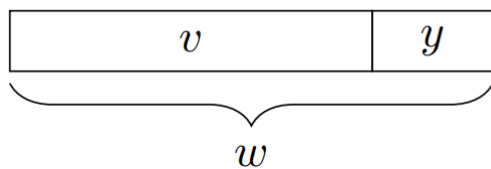
- $\varepsilon, a, b, ab, abb, bb, abba, bba$  und  $ba$  sind die Teilwörter von  $abba$ .
- $abba$  ist kein echtes Teilwort von  $abba$  (alle anderen ja).

In Programmiersprachen ist der Begriff substring gebräuchlich.

### Präfix

Ein Wort  $v$  ist ein Präfix von  $w$ , wenn

$$w = xy$$



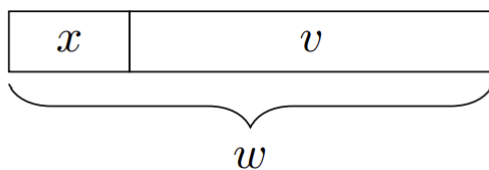
Ein echtes Präfix von  $w$  ist jedes Präfix von  $w$ , das nicht identisch mit  $w$  ist (in diesem Fall ist  $y$  leer).

- $\varepsilon, a, ab, abb$  und  $abba$  sind die Präfixe von  $abba$ .
- $abba$  ist kein echtes Präfix von  $abba$  (alle anderen ja).

### Definition (Suffix)

Ein Wort  $v$  ist ein Suffix von  $w$ , wenn

$$w = xv$$



Ein echtes Suffix von  $w$  ist jedes Suffix von  $w$ , das nicht identisch mit  $w$  ist (in diesem Fall ist  $x$  leer).

- $abba, bba, ba, a$  und  $\varepsilon$  sind die Suffixe von  $abba$ .
- $abba$  ist kein echtes Suffix von  $abba$  (alle anderen ja).

### Definition (Menge aller Wörter der Länge k)

Die Menge aller Wörter der Länge k über einem Alphabet  $\Sigma$  wird mit  $\Sigma^k$  bezeichnet.

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

### Definition (Menge aller Wörter (Zeichenreihen))

Die Menge aller Wörter (Kleenesche Hülle) über einem Alphabet  $\Sigma$  wird mit  $\Sigma^*$  bezeichnet.  $\Sigma^+ = \Sigma^* \varepsilon$  ist die Menge aller nichtleeren Wörter (positive Hülle) über einem Alphabet  $\Sigma$ .

Regex definitionen ursprung von hier.

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)

Definition (Konkatenation) Seien  $x$  und  $y$  zwei beliebige Wörter. Dann steht

$$x \circ y = xy := (x_1, x_2 \dots x_n, y_1, y_2 \dots y_m)$$

für die Konkatenation (Verkettung) von  $x$  und  $y$ .

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

### Definition (Wortpotenzen)

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

$$\begin{aligned} x^0 &:= \varepsilon \\ x^{n+1} &:= x^n \circ x = x^n x \end{aligned}$$

$$a^3 = a^2 a = a^1 aa = a^0 aaa = aaa$$

$$bbababababbbaaabab = b^2(ab)^4ba^4bab = b(ba)^4b^2a^3(ab)^2$$

$$abbabbabbabbabbabbabbabbabba = a(bba)^9$$

## Definition (Sprache)

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

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

### Anmerkungen:

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

## Definition (Konkatenation von Sprachen)

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

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

Die Sprachen  $A$  und  $B$  sind wie folgt gegeben:

- $A$  enthält alle Binärwörter, die mit 1 beginnen.
- $B$  enthält alle Binärwörter, die mit 0 enden.

Welche der folgenden Wörter sind Elemente von  $AB$ ?

$\varepsilon$ ✗	10 ✓	01010 ✗
1010 ✓	11 ✗	1100110010 ✓

Wie kann man die Elemente von  $AB$  einfach beschreiben?

$$AB = \text{Menge der geraden Binärzahlen ohne Null}$$

## Reguläre Ausdrücke

Reguläre Ausdrücke sind Wörter, die Sprachen beschreiben, also eine Möglichkeit (gewisse) Sprachen endlich zu repräsentieren.

- Die Syntax der regulären Ausdrücke befasst sich mit der Frage, welche Form diese Wörter haben.
- In der Semantik der regulären Ausdrücke wird erklärt, wie man reguläre Ausdrücke als Sprachen interpretiert.

Gegeben: Das Wort 101 über dem Alphabet  $\Sigma = \{0, 1, 2, \dots, 9\}$

- Die Syntax beschreibt, wie die Symbole des Alphabets zu einem Wort angeordnet bzw. aneinandergereiht werden.
- Aus der Semantik geht hervor, was diese Zeichenreihe bedeutet:  
Z.B. die Zahl 101 im Zehnersystem, die Zahl 5 im Dualsystem oder einfach nur eine Folge von Symbolen usw.

Ein regulärer Ausdruck, der die Sprache aller Binärwörter der Länge 4 beschreibt:

$$\underbrace{(0|1)}_{\text{0 oder 1}} \underbrace{(0|1)}_{\text{nochmals}} \underbrace{(0|1)}_{\text{dreimal}} \underbrace{(0|1)}_{\text{genug}}$$

Ein passender regulärer Ausdruck ist also

$$(0|1)(0|1)(0|1)(0|1)$$

Ein regulärer Ausdruck für die Sprache der Binärwörter, die das Teilwort 00 enthalten:

$$\underbrace{(0|1)^*}_{\text{0 oder 1 beliebig oft}} \underbrace{00}_{\text{das Teilwort}} \underbrace{(0|1)^*}_{\text{0 oder 1 beliebig oft}}$$

Ein passender regulärer Ausdruck ist also

$$(0|1)^*00(0|1)^*$$

### Definition (Reguläre Ausdrücke)

Es sei  $\Sigma$  ein beliebiges Alphabet. Die Sprache  $RA_{\Sigma}$  der regulären Ausdrücke über  $\Sigma$  ist wie folgt definiert:

- $\emptyset, \epsilon \in RA_{\Sigma}$
- $\Sigma \subset RA_{\Sigma}$
- $R \in RA_{\Sigma} \Rightarrow (R^*) \in RA_{\Sigma}$
- $R, S \in RA_{\Sigma} \Rightarrow (RS) \in RA_{\Sigma}$
- $R, S \in RA_{\Sigma} \Rightarrow (R|S) \in RA_{\Sigma}$

### Erläuterungen zur Definition

- Die Sonderzeichen  $\epsilon$  und  $\emptyset$  sind reguläre Ausdrücke.
- Jedes Symbol aus dem Alphabet  $\Sigma$  ist auch ein regulärer Ausdruck über  $\Sigma$ .

- Ist  $R$  ein regulärer Ausdruck über  $\Sigma$ , dann ist auch  $(R^*)$  ein regulärer Ausdruck über  $\Sigma$ .
- Sind  $R$  und  $S$  reguläre Ausdrücke über  $\Sigma$ , dann sind auch  $(RS)$  und  $(R|S)$  Ausdrücke über  $\Sigma$ .

## Eigenschaften und Konventionen:

- Die **Menge**  $RA_\Sigma$  **der regulären Ausdrücke** über dem Alphabet  $\Sigma$  ist **eine Sprache** über dem Alphabet  $\{\emptyset, \epsilon, *, (, ), |\} \cup \Sigma$ .

Elemente von  $RA_\Sigma$  sind z.B.  $\emptyset, \epsilon, a, a^*, a|b, aba, ..$

- Der Lesbarkeit halber werden “überflüssige” Klammern weggelassen.
- Damit reguläre Ausdrücke auch mit (teilweise) weggelassenen Klammern eindeutig lesbar bleiben, gilt folgende Rangfolge der Operatoren:
  - a) “\*” vor “Konkatenation” und
  - b) “Konkatenation” vor “|”.

Der Ausdruck  $ab^*|c$  wird beispielsweise als  $((a(b^*))|c)$  gelesen.

Einige reguläre Ausdrücke über dem Alphabet  $\{a, b\}$ .

- $a^*b$
- $(aa)^*b^*aba$
- $(a|(ab))^*$
- $(ab)|(ba)$
- $a(b(ba))|b$