Theoretische Informatik 3. Semester

Prof. Dr. Martin Plümicke

17. Oktober 2023

Inhaltsverzeichnis

2 Grundlagen der Theorie der formalen Sprache

Zunächst wollen wir die beiden Begriffe *formale Sprache* und *natürliche Sprache* gegeneinander abgrenzen.

Formale Sprache:

- erlaubt es ihre Regeln (mathematisch) extakt zu definieren.
- Bsp. Java, C++, C - Shell, perl, PIZZA

Natürliche Sprache: Deutsch, Englisch

- über Jahrhunderte gewachsen
- Rechtschreibung
- Grammatik (nicht exakt, viele Ausnahmen)

Im Folgenden sei Σ die Menge von Symbolen (Buchstaben). Man nennt Σ das Alphabet.

Definition 2.1 (Menge aller Wörter)**.** Die kleinste Menge Σ^* die folgende Bedingungen erfüllt, heißt *Menge aller Wörter* über Σ:

1. $\varepsilon \in \Sigma^*$ (ε = leeres Wort)
2. wenn $w \in \Sigma^*$ und $a \in \Sigma$, dann ist $aw \in \Sigma^*$

Die *Länge* eines Wertes w bezeichnen wir mit $|w|$

Beispiel.

1. $\Sigma = \{0, 1\} \Rightarrow \Sigma^* = \{\varepsilon, 0, 1, 10, 11, 110, 111, \dots\}$
2. $\Sigma = \{a, b\}$, $M = \{\varepsilon, x, y, a, ax, bx, \dots\}$
   Frage: Ist M die Menge aller Wörter über Σ
   Nein: alle Elemente mit x und y müssen entfernt werden, da M zwar die zwei Bedingungen erfüllt, aber nicht die kleinste Menge ist, die diese Bedingungen erfüllt.

Definition 2.2 (Sprache)**.** Jede Teilmenge $L \subseteq \Sigma^*$ heißt Sprache.

Beispiel. $\Sigma = \{a, b\}$

- $\Sigma \subseteq \Sigma^*$
- $\Sigma^* \subseteq \Sigma^*$
- $\{aa, bb\} \subseteq \Sigma^*$ (endliche Sprache)
- $\{w \mid w \in \Sigma^*, w$ enthält an der 2. Stelle kein $a\}$ (unendliche Sprache)

Definition 2.3 (Semi–Thue–System)**.** Ein Paar (Σ, Π) heißt *Semi–Thue–System*, wenn Σ ein Alphabet ist und $\Pi \subseteq \Sigma^* \times \Sigma^*$ eine *Relation* mit folgenden Eigenschaften ist:
$\Pi = \{(l_i \to r_i) \mid 1 \leqslant i \leqslant n, n \geqslant 0, l_i, r_i \in \Sigma^*\}$

Beispiel. $\Sigma = \{a, b\}$

$\Pi \subseteq \Sigma^* \times \Sigma^* = \{(a \to a), (a \to b), (b \to b), (\varepsilon \to bb)\}$

Definition 2.4 (Ableitung)**.** Sei (Σ, Π) ein Semi–Thue–System, dann nennt man

$$\left\{w \xrightarrow{1} w' \mid w = ulv, w' = urv, l \to r \in \Pi\right\}$$

die Menge der *direkten Ableitungen*.

$$\left\{w \xrightarrow{*} w' \mid x_i \xrightarrow{1} x_{i+1}, w = x_0, w' = x_n, 0 \leq i \leq n\right\}$$

heißt *Menge der Ableitungen*.

Beispiel. $\Sigma = \{a, b\}$ $\Pi = \{(a \to ab), (b \to bb)\}$

- $w = \underbrace{abb}_{u}\underbrace{a}_{l}\underbrace{b}_{v} \xrightarrow{1} \underbrace{abb}_{u}\underbrace{ab}_{r}\underbrace{b}_{v} = w'$
- $w = \underbrace{abbab}_{x_0} \xrightarrow{1} \underbrace{abbabb}_{x_1} \xrightarrow{1} \underbrace{abbabbb}_{x_2} = w' \Rightarrow w \xrightarrow{*} w'$

Definition 2.5 (erzeugte Sprache)**.** Sei (Σ, Π) ein Semi–Thue–System und $l \in \Sigma^*$ so heißt $\mathcal{L}(\Pi, l) = \left\{w \mid l \xrightarrow{*} w\right\}$ die von l *erzeugte Sprache*

Beispiel. $\Sigma = \{A, \ldots Z, a, \ldots z, \{, \}, ;, \sqcup\}$
$\Pi = \{S \to \varepsilon, S \to CS, C \to class\sqcup N\sqcup\{\}; N \to \varepsilon, N \to Na, \ldots, N \to Nz\}$
$\mathcal{L}(\Pi, S) = \{\varepsilon, CS, CCS, \ldots,$
$class\sqcup N\sqcup\{\};,$
$class\sqcup Nz\sqcup\{\};,$
$class\sqcup Ntz\sqcup\{\}; ,\ldots;,$
$class\sqcup Nfritz\sqcup\{\};,$
$class\sqcup fritz\sqcup\{\}; ,\ldots\}$

Als Problem ergibt sich, dass in der erzeugten Sprache eigentlich nur Strings der Form

$$class\sqcup fritz\sqcup\{\}; class\sqcup armin\sqcup\{\};$$

liegen sollten. Alle anderen Zwischenableitungen sind keine *Mini-Java-Programme.*

Wir führen eine Trennung im Alphabet ein:

1. Zeichen, die in der *Zielsprache* enthalten sind: *Terminale* (dürfen nicht weiter abgeleitet werden)
2. Zeichen, die in der *Zielsprache* nicht vorkommen: *Nichtterminale* (müssen weiter abgeleitet werden)

Diese Trennung macht es nötig, die Begriffe Semi–Thue–System und erzeugte Sprache neu zu definieren.

Definition 2.6 (Grammatik)**.** Eine *Grammatik* $G = (N, T, \Pi, S)$ ist gegeben durch:

- N, das Alphabet der *Nichtterminale*
- T, das Alphabet der *Terminale*
- $(N \uplus T, \Pi)$ ist eine Semi–Thue–System, wobei $N \uplus T$ die disjunkte Vereinigung ist $(N \cap T = \varnothing)$
- $S \in N$ ist das Startsymbol

Definition 2.7 (erzeugte Sprache)**.** Sei $G = (N, T, \Pi, S)$ eine Grammatik. Die von G erzeugte Sprache ist definiert durch $\mathcal{L}(G) = \left\{w \in T^* \mid S \xrightarrow{*} w\right\}$

Beispiel.

- $G = (N, T, \Pi, S)$
  $N = \{S, Z, R\}$
  $T = \{a, b\}$
  $\Pi = \{S \to aZ, S \to bZ, Z \to aR \mid a, R \to \varepsilon \mid aR \mid bR\}$
  $\mathcal{L}(G) = \{aa, aaa, aaaa, aab, aabb, \ldots\}$
  Die Sprache ist dadurch gegeben, dass immer an der 2.Stelle ein 'a' steht. Sonst stehen beliebige 'a' oder 'b'.

- $G = (N, T, \Pi, S)$ mit
  $T = \{c, l, a, s\}$,
  $N = \{S, A, B, C, D\}$,
  $\Pi = \{S \to cA, A \to lB, B \to aC, C \to sD, D \to s\}$
  $\mathcal{L}(G) = \{class\}$

Definition 2.8 (Chomsky–Hierarchie)**.** Sei $G = (N, T, \Pi, S)$ eine Gramatik.

Sprachtyp	Gramatik	Eigenschaft
0	allgemein	keine Einschränkung
1a	kontextsensitiv	Jede Produktion in Π hat die Form $(uAv \to u\beta v)$ mit $A \in N,\ u, v \in (N \uplus T)^*,\ \beta \in (N \uplus T)^* \setminus \{\varepsilon\}$ Ausnahme: $S \to \varepsilon$ ist erlaubt[1]
1b	nichtverkürzende Grammatik	Für jede Produktion in $p \to q \in \Pi$ gilt $\lvert p \rvert \leq \lvert q \rvert$. Ausnahme: $S \to \varepsilon$ ist erlaubt[1]
2	kontextfrei	Jede Produktion in Π hat die Form $(A \to \beta)$ mit $A \in N, \beta \in (N \uplus T)^* \setminus \{\varepsilon\}$ Ausnahme: $S \to \varepsilon$ ist erlaubt[1]
3	regulär	Jede Produktion hat in Π hat die Form $A \to aB$, $A \to a$, mit $a \in T; A, B \in N$. Ausnahme: $S \to \varepsilon$ ist erlaubt[1]

Beispiel. Grammatik für Klassennamen, Variablennamen, Methodennamen, ... (*Identifier*)

1. Kontextfreie Grammatik

   $G = (N, T, \Pi, S')$ mit

   $T = \{a, \ldots z, A \ldots Z, 0 \ldots 9, \_\}$

[1] Das Startsymbol S darf in dem Fall auf keiner rechten Seite vorkommen.

$N = \{Buchstabe, Ziffer, BezRest, S'\}$

$$
\begin{aligned}
\Pi = \{S' &\to Buchstabe \mid Buchstabe\ BezRest \\
BezRest &\to Buchstabe\ BezRest \mid Ziffer\ BezRest \mid Buchstabe \mid Ziffer \\
Buchstabe &\to a \mid \ldots \mid z \mid A \mid \ldots \mid Z, \\
Ziffer &\to 0 \mid \ldots \mid 9 \mid \_\}
\end{aligned}
$$

2. Reguläre Grammatik

$G = (N, T, \Pi, S')$ mit
$T = \{a, \ldots z, A \ldots Z, 0 \ldots 9, \_\}$
$N = \{S', A'\}$

$$
\begin{aligned}
\Pi = \{S' &\to aA' \mid \ldots \mid zA' \mid AA' \mid \ldots \mid ZA' \mid a \mid \ldots \mid z \mid A \mid \ldots \mid Z \\
A' &\to aA' \mid \ldots \mid zA' \mid AA' \mid \ldots \mid ZA' \mid 0A' \mid \ldots \mid 9A' \mid \_A' \\
&\quad \mid a \mid \ldots \mid z \mid A \mid \ldots \mid Z \mid 0 \mid \ldots \mid 9 \mid \_\}
\end{aligned}
$$