Formale Sprachen

Was ist eine *Formale Sprache* eigentlich?

Um diese Frage präzise beantworten zu können, benötigen wir den Begriff Alphabet:

> Als *Alphabet* bezeichnen wir eine endliche nichtleere Menge, deren Elemente *Buchstaben* genannt werden.

Normalerweise legt man sich auf ein bestimmtes Alphabet fest, das man mit dem griechischen Buchstaben Σ bezeichnet.

Typische Alphabete sind $\{a, b\}$ oder $\{A, B, C, \dots, Z, a, b, \dots, z\}$, aber auch $\{0, 1\}$, $\{0, 1, 2, 3, 4, 5, 6\}$ oder $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Diese Reihe lässt sich beliebig fortsetzen.

Das freie Monoid über Σ

Ein Monoid ist eine Menge mit einer assoziativen Verknüpfung und einem neutralen Element. Die Menge aller endlichen Zeichenketten, deren Zeichen Elemente von Σ sind, bilden mit der *Konkatenation* als Operation ein Monoid, das wir mit Σ^* bezeichnen, das *freie Monoid* über Σ. Neutrales Element dieses Monoids ist das *leere Wort* ε (manchmal auch λ genannt).

Beachte: Σ^* ist immer eine *abzählbar unendliche* Menge.

> Definition: Eine *formale Sprache* L (über Σ) ist eine Teilmenge von Σ^*, also $L \subseteq \Sigma^*$.

Frage: Wieviele formale Sprachen über Σ gibt es?

Antwort: Überabzählbar viele!

Grammatiken: Definition

Formal ist eine Grammatik ein Quadrupel:

$$G = (V, \Sigma, P, S)$$

Hierbei ist:

- V eine endliche nichtleere Menge, deren Elemente wir *Variablen* nennen.
- Σ eine endliche nichtleere Menge, das *Alphabet*.
- $P \subseteq (V \cup \Sigma)^+ \times (V \cup \Sigma)^*$ (endlich), die *Regelmenge*.

  Elemente von P heißen auch *Produktionen*.
- $S \in V$, das *Startsymbol*.

Wichtige Begriffe

Satzformen sind Elemente von $(V \cup \Sigma)^*$.

Die Grammatik G beschreibt eine Sprache, für deren Definition wir die Relation $\Rightarrow_G$ auf der Menge der Satzformen benötigen.

Die Übergangsrelation $\Rightarrow_G$ auf der Menge $(V \cup \Sigma)^*$ ist durch die Regelmenge P wie folgt festgelegt:

Wenn $(u, v) \in P$ und $w_1, w_2 \in (V \cup \Sigma)^*$ gilt, dann folgt

$$w_1 u w_2 \Rightarrow_G w_1 v w_2$$

Nun können wir $L(G)$ definieren:

$$\boxed{L(G) = \{w \in \Sigma^* \mid S \Rightarrow_G^* w\}}$$

Mit $\Rightarrow_G^*$ bezeichnen wir hierbei die reflexive transitive Hülle von $\Rightarrow_G$.

Korrekt geklammerte arithmetische Ausdrücke

Wir definieren eine Grammatik

$$G = (\{E, T, F\}, \{(,), a, +, *\}, P, E)$$

Hierbei steht E für *Expression* (Ausdruck), T für Term, sowie F für Faktor. Die Regelmenge P sei die folgende:

$$\begin{array}{llll} P = \{ & E \to T, & E \to E + T, & T \to F, \\ & T \to T * F, & F \to a, & F \to (E) \quad \} \end{array}$$

wobei $u \to v$ generell als andere Schreibweise für (u, v) anzusehen ist.

Wir geben *Ableitungen* von Satzformen in dieser Grammatik an:

$$E \Rightarrow_G T \Rightarrow_G F \Rightarrow_G a$$

$$E \Rightarrow_G T \Rightarrow_G T * F \Rightarrow_G F * F \Rightarrow_G a * F \Rightarrow_G a * (E)$$

Es gilt $a \in L(G)$ (warum?), aber $a * (E) \notin L(G)$ (warum nicht?).

Können wir $a * (a + a) \in L(G)$ zeigen?

Ein einfaches Beispiel

Betrachte die folgende Grammatik:

$$G = (\{S\}, \{a, b\}, \{S \to aSa, S \to bSb, S \to \varepsilon\}, S)$$

Offenbar gehören Wörter wie *abba* oder *bbbaabbb* zu $L(G)$, *aaabbb* oder *bab* dagegen nicht. Tatsächlich besteht $L(G)$ genau aus allen Wörtern der Form xy, wo $x = x_1x_2 \dots x_m$ mit $x_i \in \{a, b\}$ und $y = x_m \dots x_2x_1$ gilt. Anders gesagt: $L(G)$ besteht aus allen *Palindromen* gerader Länge über dem Alphabet $\{a, b\}$.

Man kann das auch formal beweisen. Versuchen Sie es bitte!

Eine ähnlich einfache Grammatik kann verwendet werden, um die Sprache $\{a^nb^n \mid n \geq 1\}$ zu beschreiben.

Deutlich mehr Mühe macht die Beschreibung der Sprache $\{a^nb^nc^n \mid n \geq 1\}$. Schauen Sie sich das im Buch genau an!

Die Chomsky-Hierarchie

Noam Chomsky definierte vier Klassen von Grammatiken, wobei er sich in erster Linie an den Eigenschaften der beteiligten Regeln, also der Elemente von P in der Grammatik $G = (V, \Sigma, P, S)$ orientierte. Dabei erhält man:

Typ 0: Allgemeine *Phrasenstrukturgrammatiken*

Typ 1: Sogenannte *kontextsensitive* Grammatiken

Typ 2: Die *kontextfreien* Grammatiken

Typ 3: Die *regulären* oder *rechtslinearen* Grammatiken

Man erhält eine echte *Hierarchie:* Alle Grammatiken sind vom Typ 0, aber nicht alle vom Typ 1. Die Grammatiken vom Typ 2 bilden eine echte Teilmenge in der Menge der Typ 1-Grammatiken, und die vom Typ 3 wiederum eine echte Teilmenge derer vom Typ 2.

Formale Definition

Definition:

Jede Grammatik G ist vom Typ 0.

Wenn jedes Paar $(u, v) \in P$ die Bedingung $|u| \leq |v|$ erfüllt, dann ist G vom Typ 1.

Wenn G vom Typ 1 ist und für jedes Paar $(u, v) \in P$ die Bedingung $u \in V$ gilt, dann ist G vom Typ 2.

Wenn G vom Typ 2 ist und für jedes Paar $(u, v) \in P$ die Bedingung $v \in \Sigma \cup \Sigma V$ erfüllt ist, dann ist G vom Typ 3.

Grundlage war hier generell eine Grammatik $G = (V, \Sigma, P, S)$.