Kontextfreie Grammatiken

Bisher haben wir verschiedene Automatenmodelle kennengelernt. Diesen Automaten können Wörter vorgelegt werden, die von den Automaten gelesen und dann akzeptiert oder abgelehnt werden. In der kommenden Woche wollen wir uns mit *Grammatiken* beschäftigen. Diese akzeptieren keine Wörter, sondern *generieren* sie. Bei einer Grammatik beginnen wir mit einem *Startsymbol* (üblicherweise S) und benutzen dann *Regeln*, um sogenannte *Satzformen* und letztendlich Wörter abzuleiten. Eine Regel hat z.B. die Form $S \to aS$ mit der Bedeutung, dass wir das S in einer Satzform nehmen und durch aS ersetzen können. Ein Beispiel: Angenommen wir haben die Regeln $S \to aS$ und $S \to a$. Das Startsymbol sei S. Wir benutzen nun zuerst die Regel $S \to aS$. Dies führt zu $S \Rightarrow aS$ (man beachte, dass wir den Pfeil $\to$ für Regeln benutzen, den Pfeil $\Rightarrow$ für Ableitungen). Wir haben nun die Satzform aS erreicht. Hier können wir wieder auf das S die Regel $S \to aS$ anwenden. Dies führt dann (insgesamt) zu $S \Rightarrow aS \Rightarrow aaS$. Wenn wir wollen, könnten wir nun wieder die Regel $S \to aS$ benutzen. Wir entscheiden uns aber diesmal für die andere Regel $S \to a$ und gelangen so zu dem Wort aaa. Das Wort aaa ist nun in der von der Grammatik generierten Sprache. Kommt in einer Zeichenkette noch ein Nonterminal vor, so sprechen wir meist von einer Satzform und erst, wenn nur noch Terminalzeichen auftreten, von einem Wort. Ziel ist es immer, zu Wörtern zu gelangen, die nur noch aus Terminalzeichen bestehen. Diese sind dann in der von der Grammatik generierten Sprache.

Detaillierter besteht eine Grammatik G aus einer Menge von *Nonterminalen* V_N (dies sind i.A. die Großbuchstaben), einer Menge aus *Terminalen* (i.A. die Kleinbuchstaben; die Terminalmenge entspricht zudem der Menge der Eingabesymbole bei einem Automaten), dem *Startsymbol* (ein besonders ausgezeichnetes Nonterminal) und einer Menge von *Regeln* oder *Produktionen*. Die Regeln sind ein Tupel (u, v) (dargestellt als $u \to v$), wobei in u mindestens ein Nonterminal auftreten muss. In einer *Ableitung* beginnt man nun mit dem Startsymbol und wendet dann jeweils eine der Regeln an. Eine Regel kann dabei dann angewendet werden, wenn die linke Seite der Regel (das u oben) in dem bisher abgeleiteten Wort auftritt. Ist dies der Fall, wird das u durch das v (die rechte Seite der Regel) ersetzt. An obigem, einführenden Beispiel konnte man auch bereits sehen, dass im Falls mehrerer anwendbarer Regeln eine beliebige gewählt werden darf.

Verschiedene Grammatik*typen* (vergleichbar mit verschiedenen Automatenmodellen) zeichnen sich dadurch aus, dass die Produktionen (die Regeln) eingeschränkt werden. Wir wollen uns nachfolgend auf *kontextfreie Grammatiken*

konzentrieren. Bei diesen wird verlangt, dass die linke Seite einer Produktion stets aus genau einem Nonterminal besteht (dieses Nonterminal kann dann unabhängig von dem Kontext, in dem es steht, ersetzt werden; daher der Name).

Wir wollen dies an zwei typischen Beispielen erläutern. Als Abkürzung für die Produktionen führen wir noch folgende Notation ein: Haben wir mehrere Produktionen mit gleicher linker Seite, so fassen wir die rechten Seiten zusammen und trennen sie mit einem "|". Haben wir bspw. die drei Regeln $S \to aS$, $S \to SbS$ und $S \to \lambda$, so haben diese alle die gleiche linke Seite S. Wir fassen diese drei Regeln daher zu $S \to aS \mid SbS \mid \lambda$ zusammen. Nun zu den zwei Beispielen.

1. Sei G_1 eine Grammatik mit dem Terminalalphabet $V_T = \{a, b\}$, dem einzelnen Nonterminal $V_N = \{S\}$, wobei S auch das Startsymbol sei, und den Produktionen $P = \{S \to aSb \mid \lambda\}$. Eine Ableitung beginnt nun mit S. Wir können nun wiederholt die Produktion $S \to aSb$ anwenden. Dies führt zu
$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow \ldots \Rightarrow a^iSb^i.$$
Hier wurde die Produktion i mal angewandt. Da hier nun noch das Nonterminal auftritt, ist die Satzform a^iSb^i noch nicht in der von G_1 generierten Sprache. Mit der Produktion $S \to \lambda$ kann das S aber entfernt (oder gelöscht) werden. Wir erhalten so a^ib^i. Allgemein lässt sich die von G_1 generierte Sprache, notiert als $L(G_1)$, mit $\{a^nb^n \mid n \in \mathbb{N}\}$ gleichsetzen. Wie bei Automaten ist aber $L(G_1) = \{a^nb^n \mid n \in \mathbb{N}\}$ noch zu zeigen, wobei üblicherweise wieder zwei Mengeninklusionen zu zeigen sind!

2. Sei G_2 eine Grammatik mit dem Terminalalphabet $V_T = \{a, b\}$, den Nonterminalen $V_N = \{S, A, B\}$, wobei S das Startsymbol sei, und den Produktionen $P = \{S \to AB, A \to aA \mid \lambda, B \to bB \mid \lambda\}$. Da S das Startsymbol ist, wird jede Ableitung mit $S \Rightarrow AB$ beginnen. Dann kann aus A durch wiederholte Anwendung der Produktion $A \to aA$ eine Satzform a^iA generiert werden. Ebenso aus B durch die Produktion $B \to bB$ eine Satzform b^jB. Beide können dann durch $A \to \lambda$ bzw. $B \to \lambda$ zu dem Wort a^i bzw. b^j abgeleitet werden. Da man die letztgenannten Produktionen auch sofort benutzen kann, ist aus A jedes Wort aus a^* ableitbar und aus B jedes Wort aus b^*. Insgesamt ist so in der Grammatik G_2 jedes Wort aus a^*b^* ableitbar.

Bei dem ersten Beispiel haben wir ein typisches "nach innen Wandern" eines Nonterminals, wobei der linke und rechte Rand wächst. Man kann auf diese Weise ein gleichmässiges Wachstum auf der linken und rechten Seite erreichen bzw. diese miteinander in Beziehung setzen. Das zweite Beispiel zeigt ein typisches "lineares Wachsen". Eine Zeichenkette aus a^* (oder b^*) wird hier Stück für Stück aufgebaut. Beide Techniken können auch gut kombiniert werden, um z.B. eine Grammatik für $\{a^nb^nc^m \mid n, m \in \mathbb{N}\}$ zu erstellen. Es sei aber noch einmal darauf hingewiesen, dass stets ein Beweis von $L(G) = M$ nötig ist, wenn man eine Grammatik G für eine gegebene Wortmenge M konstruiert.

In der Vorlesung werden wir Grammatiken formal einführen und uns dann hauptsächlich auf kontextfreie Grammatiken konzentrieren. Wir werden zeigen, dass diese genau die gleiche Sprachfamilie akzeptieren wie die Kellerautomaten. D.h. zu jeder von einem Kellerautomaten akzeptierten Sprache kann eine kontextfreie Grammatik konstruiert werden, die diese Sprache generiert. Und andersherum: Zu jeder Sprache, die von einer kontextfreien Grammatik generiert werden kann, kann auch ein Kellerautomat konstruiert werden, der diese Sprache akzeptiert. Außerdem werden wir eine noch stärkerer Einschränkung von Grammatiken einführen, sogenannte reguläre Grammatiken. Diese generieren dann genau die regulären Sprachen, d.h. jene Sprachen, die von einem DFA akzeptiert werden können. Es ist dabei spannend zu sehen, wie diese recht unterschiedlichen Formalismen (Grammatiken, die Wörter generieren, und Automaten, die Wörter akzeptieren) ineinander umgewandelt werden können. Man kann dann je nach Problemstellung situativ den passenderen Formalismus wählen.

Ein weiteres wichtiges Thema in der Vorlesung wird das *Wortproblem* sein. Das Wortproblem meint folgende Fragestellung: Kann zu einem gegebenen Wort w und einer gegebenen Grammatik G berechnet werden, ob w von G generiert werden kann? Im Kontext von Automaten ist dieses Problem oft recht einfach zu entscheiden, da man dem Automaten das Wort einfach vorlegen, die Übergänge entlang gehen und dann prüfen kann, ob man letztendlich in einen Endzustand gelangt ist. Bei Grammatiken ist dies aber nicht so einfach. Man müsste ja unter Umständen alle bei der aktuellen Satzform anwendbaren Produktionen ausprobieren und bei den dabei entstehenden Satzformen so weiter fortfahren. Davon abgesehen, dass man dabei in Kreise geraten kann (z.B. wenn man die Produktion $A \to B$ anwendet und dann die Produktion $B \to A$), wäre dies auch recht ineffizient. Wir werden ein schnelles Verfahren kennenlernen, um das Wortproblem zu lösen. Dazu benötigen wir allerdings quasi als Vorstufe ein Verfahren, um eine kontextfreie Grammatik in eine sogenannte *Normalform* zu überführen. Bei dieser Normalform treten nur zwei Arten von Regeln auf: Entweder hat eine Regel die Form $A \to BC$, wobei A, B und C Nonterminale sind oder die Form $A \to a$, wobei a ein Terminal ist.

Zuletzt werden wir ein weiteres Pumping Lemma kennenlernen. Bisher kannten wir das Pumping Lemma für reguläre Sprachen. Dieses können wir benutzen, um zu zeigen, dass eine Sprache *nicht* regulär ist. Wir werden nun noch das Pumping Lemma für kontextfreie Sprachen kennenlernen. Dieses kann dann benutzt werden, um zu zeigen, dass eine Sprache nicht kontextfrei ist.

(Selbsttest auf der nächsten Seite.)

Selbsttest (Lösung auf Seite 6)

1. Wie funktioniert eine Ableitung in einer Grammatik?
2. Wie sieht eine Regel in einer kontextfreien Grammatik aus?
3. Wenn man zwei Regeln mit gleicher linker Seite hat, wie wird ausgewählt welche Regel benutzt wird?
4. Kann man zwei Regeln mit gleicher rechten Seite haben?
5. Kann es mehrere Möglichkeiten geben, um das gleiche Wort abzuleiten?
6. Wozu kann das Pumping Lemma benutzt werden?

Die mathematische Seite

Wir definieren kontextfreie Grammatiken, Ableitungen und die von einer Grammatik generierte Sprache.

Definition 1. *Eine* Grammatik *ist ein Quadrupel* $G = (V_N, V_T, P, S)$ *mit*

1. *Dem endlichen Alphabet von* Nonterminalen V_N.
2. *Dem endlichen Alphabet von* Terminalen V_T *mit* $V_T \cap V_N = \emptyset$. *Das* Gesamtalphabet *wird mit* $V := V_T \cup V_N$ *bezeichnet.*
3. *Der endlichen Menge von* Produktionen *(oder* Regeln*)* $P \subseteq (V^* \backslash V_T^*) \times V^*$.
4. *Dem* Startsymbol $S \in V_N$.

Eine Grammatik ist eine kontextfreie Grammatik *(CFG), wenn die endliche Menge der Produktionen eingeschränkt ist auf* $P \subseteq V_N \times V^*$.

Eine kontextfreie Produktion (A, λ) *wird als* λ-Produktion *bezeichnet. Besitzt eine CFG keine* λ*-Produktionen, so heißt sie* λ-frei.

Eine Regel $(u, v) \in P$ wird üblicherweise als $u \to v$ notiert. Man beachte, dass bei den Regeln einer Grammatik auf der linken Seite stets ein Nonterminal vorkommen muss. Bei einer kontextfreien Grammatik darf auf der linken Seite sogar immer nur genau ein Nonterminal (und nichts weiteres mehr) stehen.

Definition 2. *Die* einschrittige Ableitung *eines Wortes* v *aus einem Wort* u *mittels einer Produktion einer Grammatik* G *wird notiert als* $u \underset{G}{\Longrightarrow} v$. *Dabei ist die Relation* $\underset{G}{\Longrightarrow} \subseteq V^* \times V^*$ *für alle* $u, v \in V^*$ *definert durch:*

$$u \underset{G}{\Longrightarrow} v \text{ gdw. } \exists u_1, u_2 \in V^* \exists (w_l, w_r) \in P : u = u_1 w_l u_2 \text{ und } v = u_1 w_r u_2$$

Ist der Kontext klar, wird das tief gestellte G *weggelassen. Ferner bedienen wir uns wieder der reflexiven, transitiven Hülle* $\overset{*}{\underset{G}{\Longrightarrow}}$ *für mehrschrittige Ableitungen.*

In einer kontextfreien Grammatik wird die Ableitung genauso definiert wie oben. Allerdings ist bei der gewählten Produktion (w_l, w_r) dann w_l stets nur genau ein Nonterminal.

Definition 3. *Sei* $G = (V_N, V_T, P, S)$ *eine Grammatik. Die von* G generierte *oder* erzeugte *Sprache ist*

$$L(G) := \{w \in V_T^* \mid S \overset{*}{\underset{G}{\Longrightarrow}} w\}$$

Definition 4. *Die* Familie der kontextfreien Sprachen *ist eine Familie von Sprachen. Für jede dieser Sprachen gibt es eine kontextfreie Grammatik, die sie generiert. Abgekürzt wird diese Sprachfamilie mit CF.*

Lösungen zum Selbsttest auf Seite 4

1. **F:** Wie funktioniert eine Ableitung in einer Grammatik?

   A: Beginnend mit dem Startsymbol werden wiederholt Regeln angewendet, um letztendlich zu einer Satzform zu kommen, die nur aus Nonterminalen besteht (also zu einem Wort). Eine Regel $u \to v$ kann dabei angewendet werden, wenn in dem bisher abgeleiteten Wort w das u als Teilwort auftritt. Das u wird dann aus w herausgelöscht und durch v ersetzt. Galt also $w = w_1 \cdot u \cdot w_2$, so hat man nach Anwendung der Regel das Wort $w_1 \cdot v \cdot w_2$.

2. **F:** Wie sieht eine Regel in einer kontextfreien Grammatik aus?

   A: Die linke Seite einer solchen Regel besteht aus genau einem Nonterminal. Bspw. sind $A \to abX$, $A \to aXbA$ und $A \to \lambda$ Regeln. Die rechte Seite unterliegt insb. keinerlei Einschränkungen. Auf der rechten Seiten darf das leere Wort oder eine beliebige Mischung aus Nonterminalen und Terminalen stehen. Auf der linken Seite darf aber immer nur genau ein Nonterminal stehen.

3. **F:** Wenn man zwei Regeln mit gleicher linken Seite hat, wie wird ausgewählt welche Regel benutzt wird?

   A: Hier hätte man die freie Wahl. Beides ist erlaubt. Führt dies letztendlich zu unterschiedlichen generierten Wörtern, so sind beide in der generierten Sprache der Grammatik. Hat man bspw. als Regeln $S \to a$ und $S \to b$, so kann man eine der beide Regeln auswählen und somit sowohl a als auch b generieren.

4. **F:** Kann man zwei Regeln mit gleicher rechten Seite haben?

   A: Ja! Dies braucht man bisweilen auch. Z.B. kann es nötig sein $A \to a$ als auch $B \to a$ zu haben. Hat man noch weitere Produktionen für A und B, so sind dies vielleicht die "Abbruchproduktionen", um das Nonterminal weg zu kriegen und zu einem Terminalwort zu kommen.

5. **F:** Kann es mehrere Möglichkeiten geben, um das gleiche Wort abzuleiten?

   A: Ja, auch das ist möglich. Dies tritt sogar recht häufig auf. Hat man z.B. bei einer kontextfreien Grammatik eine Produktion $S \to AB$, so kann man im Anschluss ja mit dem A oder dem B fortfahren.

6. **F:** Wozu kann das Pumping Lemma benutzt werden?

   A: Zur Wiederholung: Das Pumping Lemma für *reguläre Sprachen* kann benutzt werden, um für eine Sprache zu zeigen, dass diese *nicht regulär* ist. Es gibt aber auch nicht reguläre Sprachen, die die Bedingungen im Pumping Lemma erfüllen! Das Pumping Lemma für *kontextfreie Sprachen* kann nun benutzt werden, um für eine Sprache zu zeigen, dass diese *nicht kontextfrei* ist. Es gibt aber auch hier Sprachen, die die Bedingungen dieses Pumping Lemmas erfüllen und dennoch nicht kontextfrei sind.