Die Chomsky-Hierarchie

Ziel:

Wir wollen Regelsysteme entwerfen, mit denen sich genau die Wörter einer vorgegebenen Sprache erzeugen lassen. Diese werden auch Grammatiken genannt.

Eine Grammatik besteht aus vier Komponenten:

- T (oder Σ) - Terminalzeichen, oder auch das endliche Alphabet, über dem die zu erzeugende Sprache definiert ist
- V - Variablen, oder auch die zu T disjunkte Menge an Hilfszeichen
- $S \in V$ - dem Startsymbol
- P - endliche Menge von Produktionen oder Ableitungsregeln in Form von Paaren (l,r) wobei $l \in (V \cup T)^+$ und $r \in (V \cup T)^*$ ist. Wenn l in einem Wort z ein Teilwort ist, so darf dieses durch r aus der entsprechenden Produktion ersetzt werden.

Notation:

Wir benutzen die Form $w \rightarrow z$, wenn sich z durch **einen** Produktionsschritt aus w ableiten lässt. Bei endlich vielen Produktionsschritten verwenden wir $w \rightarrow^* z$. Daraus folgt, dass sich die zu einer Grammatik gehörende Sprache L(G) durch $\forall z \in T^*$ mit $S \rightarrow^* z$ definieren lässt.

Beispielgrammatik für arithmetische Ausdrücke:

T={(,),a,+,*}, V={S} und P={

S→(S) + (S), (S) * (S), a, a+a, a*a

}

Das Wortproblem:
Hierbei geht es darum, zu entscheiden, ob ein Wort $w \in T^*$, für eine gegebene Grammatik G, in der Sprache L(G) enthalten ist. Um einen Kompromiss zwischen der hohen Komplexität des Problems bei unbeschränkten Grammatiken und zu wenig mächtigen Programmiersprachen bei stärkerer Einschränkung zu schaffen, gibt es folgende 4 Grammatikklassen in der Chomsky-Hierarchie:

- Typ Chomsky-0: nicht weiter eingeschränkte Grammatiken
- Typ Chomsky-1: alle Produktionen haben die Form $u \rightarrow v$ mit $u \in V^+, v \in ((V \cup T) - \{S\})^+$ und $|u| \leq |v|$ oder $S \rightarrow \varepsilon$. Sie werden auch kontextsensitiv genannt.
- Typ Chomsky-2: alle Produktionen haben die Form $A \rightarrow v$ mit $A \in V$, $v \in (V \cup T)^*$. Sie heißen auch kontextfreie Grammatiken.
- Typ Chomsky-3: alle Produktionen haben die Form $A \rightarrow v$ mit $A \in V$ und $v = \varepsilon$ oder $v = aB$ mit $a \in T$ und $B \in V$. Sie heißen reguläre oder auch rechtslineare Grammatiken.

Dabei gilt: $L_3 \subseteq L_2 \subseteq L_1 \subseteq L_0$

Chomsky-0 Grammatiken

Satz: *Falls L rekursiv aufzählbar ist, gibt es eine Chomsky-0 Grammatik mit L(G)=L.*

Satz: *Wenn L durch eine Chomsky-0 Grammatik beschrieben ist, gibt es eine NTM M, die genau L akzeptiert.*

Satz: *Wenn L durch eine NTM M Akzeptiert wird, ist L rekursiv aufzählbar.*

Damit sind folgende Aussagen äquivalent: **Eine Sprache**...

- gehört zu den **rekursiv aufzählbaren**.
- wird von **DTM's akzeptiert**.
- wird von **NTM's akzeptiert**.
- wird durch **Chomsky-0 Grammatiken** erzeugt.

Chomsky-3-Grammatiken

Satz: *Die Klasse der von endlichen Automaten akzeptierten Sprachen und die Klasse der Chomsky-3-Grammatiken stimmen überein.*

Definition:*Die Menge der regulären Ausdrücke über einem endlichen Alphabet* Σ *wird rekursiv durch folgende Regeln definiert:*

- Die Ausdrücke der **leeren Sprache** ($\emptyset$), des **leeren Wortes** (ε) und des **einbuchstabigen Wortes** a für $a \in \Sigma$ sind regulär.
- Wenn A_1 und A_2 regulären Ausdrücke sind, so sind auch deren **Vereinigung** $(A_1)+(A_2)$, **Konkatenation** $(A_1)*(A_2)$ und **Kleenescher Abschluss** $(A_1)^*$ reguläre Ausdrücke.
- Alle regulären Ausdrücke lassen sich durch endliche Anwendung der ersten beiden Punkte erzeugen.

Satz: *Genau die regulären Sprachen lassen sich durch reguläre Ausdrücke beschreiben.*

Satz: *Die Menge der regulären Ausdrücke (und damit der regulären Sprachen) ist gegen Substitution abgeschlossen. Das bedeutet, wenn für L und alle* $f(a) \subseteq \Delta^*$, $a \in \Sigma$, *die Substitution f regulär ist, so ist f(L) regulär.*

Kontextsensitive Grammatiken

Diese gehören, wie die kontextfreien, zu dem Kompromiss zwischen den zu stark eingeschränkten regulären Grammatiken und den zu komplexen Chomsky-0 Grammatiken. Sie lassen sich durch platzbeschränkte nichtdeterministische Turingmaschinen charakterisieren.

Definition: *DTAPE(s(n)) und NTAPE(s(n)) sind die Klassen der Sprachen, die von einer deterministischen bzw nichtdeterministischen Turingmaschine mit Platzbedarf s(n) akzeptiert werden.*
PSPACE ist die Vereinigung aller DTAPE(n^k), $k \in N$, und NPSPACE die Vereinigung aller NTAPE(n^k),$k \in N$.

Satz: *Die Klasse der kontextsensitiven Sprachen stimmt mit der Klasse NTAPE(n) überein.*

Korollar: *Für jede kontextsensitive Sprache L über Σ gibt es eine kontextsensitive Grammatik G = (V,Σ,S,P), bei der alle Regeln die Form $A \rightarrow C, A \rightarrow CD, AB \rightarrow CD, A \rightarrow a$ oder $S \rightarrow \varepsilon$ haben, wobei $A,B \in V$ und $C,D \in V-\{S\}$ sowie $a \in \Sigma$ ist.*

Quellen: *Theoretische Informatik von I. Wegener Kap.5*

Wikipedia unter den Begriffen Chomsky-Hierarchie und Optimalitätsprinzip von Bellman

sowie die Vorlesung Grundlagen der Theoretischen Informatik