Informatik III

WS 03/04

Prof. Dr. Dorothea Wagner

`dwagner@ira.uka.de`

Kapitel 5 : Grammatiken und die Chomsky-Hierarchie

5.1 Definition

*Eine **Grammatik** G besteht aus vier Komponenten:*

- *dem endlichen **Alphabet** Σ;*
- *einer endlichen Menge V mit $V \cap \Sigma = \emptyset$ von **Variablen**;*
- *dem **Startsymbol** $S \in V$;*
- *einer endlichen Menge von **Ableitungsregeln** R. Dabei ist eine Ableitungsregel ein Paar (ℓ, r), wobei $\ell \in (V \cup \Sigma)^+$ und $r \in (V \cup \Sigma)^*$ ist. Wir schreiben oft auch $\ell \to r$.*

Bedeutung: Wenn in einem Wort z das Wort ℓ Teilwort von z ist, so darf ℓ durch r in z ersetzt werden.

Notation: Wir schreiben $w \rightarrow z$, wenn w durch Anwendung einer Ableitungsregel in z verwandelt wird, und $w \xrightarrow{*} z$, wenn w durch eine Anwendung von mehreren Ableitungsregeln in z verwandelt wird.

Die von einer Grammatik G **erzeugte Sprache** $L(G)$ ist die Menge aller Wörter $z \in \Sigma^*$, für die $S \xrightarrow{*} z$ gilt.

5.2 Definition (Chomsky-Hierarchie)

1. *Grammatiken ohne weitere Einschränkungen heißen Grammatiken vom* ***Typ 0****.*

2. *Grammatiken, bei denen alle Ableitungsregeln die Form*
    - $u \longrightarrow v$ *mit* $u \in V^+, v \in ((V \cup \Sigma) \setminus \{S\})^+$ *und* $|u| \leq |v|$, *oder*
    - $S \longrightarrow \varepsilon$

   haben, heißen ***kontextsensitiv*** *oder Grammatiken vom* ***Typ 1****.*

3. *Grammatiken, bei denen alle Ableitungsregeln die Form $A \rightarrow v$ mit $A \in V$ und $v \in (V \cup \Sigma)^*$ haben, heißen **kontextfrei** oder Grammatiken vom **Typ 2**.*

4. *Grammatiken, bei denen alle Ableitungsregeln die Form $A \rightarrow v$ mit $A \in V$ und $v = \varepsilon$ oder $v = aB$ mit $a \in \Sigma, B \in V$ haben, heißen **rechtslinear** oder Grammatiken vom **Typ 3**.*

Bemerkung:

Bei kontextsensitiven Grammatiken kann die Ableitung $ABC \to AXYC$ erlaubt, aber $DBC \to DXYC$ verboten sein.

Chomsky-0-Grammatiken u. rek. aufzählbare Sprachen

5.3 Satz

Falls L rekursiv aufzählbar (semi-entscheidbar) ist, so gibt es eine Chomsky–0–Grammatik mit $L(G) = L$.

5.4 Satz

Die von Typ–0–Grammatiken G erzeugten Sprachen sind rekursiv aufzählbar.

Chomsky-3-Grammatiken u. reguläre Sprachen

5.5 Satz

Die Klasse der von endlichen Automaten akzeptierten Sprachen ist genau die Klasse der von Chomsky–3–Grammatiken erzeugten Sprachen.

5.6 Definition

$\mathcal{DTAPE}(s(n))$ und $\mathcal{NTAPE}(s(n))$ sind die Klassen der Sprachen, die von einer deterministischen beziehungsweise einer nichtdeterministischen Turing-Maschine mit Platzbedarf $s(n)$ (bei Eingabelänge n) akzeptiert werden können.

Chomsky-1-Grammatiken bzw. kontextsensitive Sprachen

5.7 Satz

Die Klasse der von Chomsky–1–Grammatiken erzeugten Sprachen stimmt mit der Klasse $\mathcal{NTAPE}(n)$ *überein.*

5.8 Satz

Das Cliquen-Problem gehört zu $\mathcal{DTAPE}(n)$.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.9 Definition

Eine kontextfreie Grammatik G heißt ***eindeutig****, wenn es für jedes Wort $w \in L(G)$ genau einen Syntaxbaum gibt. Eine kontextfreie Sprache L heißt* ***eindeutig****, wenn es eine eindeutige Grammatik G mit $L(G) = L$ gibt. Ansonsten heißt L* ***inhärent mehrdeutig****.*

5.10 Definition

Eine kontextfreie Grammatik ist in ***Chomsky–Normalform****, wenn alle Regeln von der Form:*

$$A \rightarrow BC \qquad \text{oder} \qquad A \rightarrow a$$

sind, mit $A, B, C \in V$ *und* $a \in \Sigma$*. Grammatiken in Chomsky–Normalform können also nicht das Wort* ε *erzeugen.*

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

Für kontextfreie Sprachen, die ε enthalten, läßt sich eine Grammatik leicht „ergänzen" durch die Regeln

$$S' \rightarrow \varepsilon \qquad \textit{und} \qquad S' \rightarrow S$$

wobei S' ein neues Startsymbol zur Erzeugung von ε ist.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.11 Satz

Jede kontextfreie Grammatik, die nicht das leere Wort erzeugt, kann in eine Grammatik in Chomsky–Normalform überführt werden.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.12 Satz

Es gibt einen Algorithmus (den Cocke–Younger–Kasami Algorithmus), der für eine kontextfreie Grammatik G *in Chomsky–Normalform und ein Wort* $w \in \Sigma^*$ *in Zeit* $\mathcal{O}(|R| \cdot n^3)$ *entscheidet, ob* $w \in L(G)$*, wobei* $n = |w|$ *und* $|R|$ *die Anzahl der Regeln von* G *ist.*

5.13 Satz (Pumping–Lemma)

Für jede kontextfreie Sprache L gibt es eine Konstante $n \in \mathbb{N}$, so dass sich jedes Wort $z \in L$ mit $|z| \geq n$ so als $z = uvwxy$ schreiben lässt, dass $|vx| \geq 1$, $|vwx| \leq n$ und für alle $i \geq 0$ das Wort $uv^iwx^iy \in L$ ist.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.14 Satz (Ogden's Lemma)

Für jede kontextfreie Sprache L gibt es eine Konstante $n \in \mathbb{N}$, so dass für jedes Wort $z \in L$ mit $|z| \geq n$ gilt: Wenn wir in z mindestens n Buchstaben markieren, so lässt sich z so als $z = uvwxy$ schreiben, dass von den mindestens n markierten Buchstaben mindestens einer zu vx gehört und höchstens n zu vwx gehören und für alle $i \geq 0$ das Wort $uv^iwx^iy \in L$ ist.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.15 Satz

Die Chomsky-Hierarchie ist echt, d.h.
$\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0$*, wobei* $\mathcal{L}_i, 0 \leq i \leq 3,$ *Klasse der durch Typ-*i*-Grammatiken erzeugten Sprachen.*

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.16 Definition

Sei G eine kontextfreie Grammatik. Eine Variable A heißt ***nutzlos****, falls es keine Ableitung $S \xrightarrow{*} w$ gibt, $w \in \Sigma^*$, in der A vorkommt.*

5.17 Satz

Für eine kontextfreie Grammatik kann die Menge der nutzlosen Variablen (in polynomialer Zeit) berechnet werden.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.18 Korollar

Für eine kontextfreie Grammatik G kann (in polynomialer Zeit) entschieden werden, ob $L(G) = \emptyset$ ist.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.19 Satz

Für eine kontextfreie Grammatik G *kann (in polynomialer Zeit) entschieden werden, ob* $L(G)$ *endlich ist.*

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.20 Satz

Die Klasse der kontextfreien Sprachen ist abgeschlossen bzgl. Vereinigung, Konkatenation und Kleenschem Abschluss.

Chomsky-2-Grammatiken bzw. kontextfreie Sprachen

5.21 Satz

Die Klasse der kontextfreien Sprachen ist nicht abgeschlossen bzgl. Komplementbildung und Durchschnitt.

5.22 Definition

Eine kontextfreie Grammatik ist in ***Greibach-Normalform****, wenn alle Ableitungsregeln von der Form*

$$A \longrightarrow a\alpha \textit{ mit } A \in V, a \in \Sigma \textit{ und } \alpha \in V^*$$

sind.

5.23 Satz

Für jede kontextfreie Grammatik G*, für die* $L(G)$ *das leere Wort nicht enthält, kann eine äquivalente kontextfreie Grammatik* G' *(d.h.* $L(G) = L(G'))$ *in Greibach-Normalform konstruiert werden.*

5.24 Definition

Ein (nichtdeterministischer) ***Kellerautomat*** *(NPDA bzw. PDA) besteht aus* $(Q, \Sigma, \Gamma, q_0, Z_0, \delta, F)$, *wobei*

- Q *endliche Zustandsmenge*
- Σ *endliches Eingabealphabet*
- Γ *endliches STACK-Alphabet*
- $q_0 \in Q$ *Anfangszustand*
- $Z_0 \in \Gamma$ *Initialisierung des STACK*

Kontextfreie Sprachen und Kellerautomaten

- $\delta : Q \times (\Sigma \cup \{\varepsilon\}) \times \Gamma \longrightarrow 2^{Q \times \Gamma^*}$ *, d.h.*
  $\delta(q, a, z) \subseteq \{(q, \gamma) : q \in Q, \gamma \in \Gamma^*\}$ *und*
  $\delta(q, \varepsilon, z) \subseteq \{(q, \gamma) : q \in Q, \gamma \in \Gamma^*\}$
- $F \subseteq Q$ *Menge der akzeptierenden Endzustände,*
  $F = \emptyset$ *ist möglich.*

Kontextfreie Sprachen und Kellerautomaten

Eine ***Konfiguration eines PDA*** *ist ein Tripel* (q, w, α) *mit* $q \in Q$ *und* $\alpha \in \Gamma^*$ *STACK-Inhalt und* $w \in \Sigma^*$ *Teil der Eingabe, der noch nicht gelesen wurde. Zu* $(q, w_1 \dots w_k, Z_1 \dots Z_m)$ *gibt es folgende* ***Nachfolgekonfigurationen****:*

$(q^{'}, w_2 \dots w_k, Z_1^{'} \dots Z_r^{'} Z_2 \dots Z_m)$
für alle $(q^{'}, Z_1^{'} \dots Z_r^{'}) \in \delta(q, w_1, Z_1)$ *und*
$(q^{'}, w_1 \dots w_k, Z_1^{'} \dots Z_r^{'} Z_2 \dots Z_m)$
für alle $(q^{'}, Z_1^{'} \dots Z_r^{'}) \in \delta(q, \varepsilon, Z_1)$.

Kontextfreie Sprachen und Kellerautomaten

*Ein PDA **akzeptiert** ein $w \in \Sigma^*$ **durch leeren Stack**, wenn es eine zulässige Folge von Konfigurationen aus der Anfangskonfiguration (q_0, w, Z_0) in eine Konfiguration $(q, \varepsilon, \varepsilon), q \in Q$, gibt.*

Ein PDA ***akzeptiert*** *ein* $w \in \Sigma^*$ ***durch einen akzeptierenden Endzustand****, wenn es eine zulässige Folge von Konfigurationen aus der Anfangskonfiguration* (q_0, w, Z_0) *in eine Konfiguration* (q, ε, γ) *mit* $q \in F$ *und* $\gamma \in \Gamma^*$ *gibt.*

Kontextfreie Sprachen und Kellerautomaten

*Ein PDA ist **deterministisch**, falls*

$$|\delta(q, a, Z)| + |\delta(q, \varepsilon, Z)| \leq 1$$

für alle $q \in Q, a \in \Sigma, Z \in \Gamma$.

5.25 Satz

Zu einem PDA, der eine Sprache L durch einen akzeptierenden Endzustand akzeptiert, kann ein PDA konstruiert werden, der L mit leerem STACK akzeptiert.

5.26 Satz

Zu einem PDA, der eine Sprache L mit leerem STACK akzeptiert, kann ein PDA konstruiert werden, der L durch akzeptierende Endzustände akzeptiert.

5.27 Satz

Für eine Grammatik G in Greibach-Normalform kann ein PDA konstruiert werden, der $L(G)$ mit leerem STACK akzeptiert.

5.28 Satz

Jede durch einen PDA (mit leerem STACK oder durch akzeptierende Endzustände) akzeptierte Sprache ist kontextfrei.

5.29 Korollar

Die Klasse der von nichtdeterministischen Kellerautomaten akzeptierten Sprachen ist gleich der Klasse der kontextfreien Sprachen.

5.30 Satz

Sei $L \subseteq \Delta^*$ eine kontextfreie Sprache, $h : \Sigma \to \Delta^*$ ein Homomorphismus. Dann ist auch die Sprache $h^{-1}(L) = \{w \in \Sigma^* | h(w) \in L\}$ kontextfrei.

Unentscheidbare Probleme für kontextfreie Grammatike

5.31 Satz

Es ist nicht entscheidbar, ob $L(G_1) \cap L(G_2) = \emptyset$ *für kontextfreie Grammatiken* G_1 *und* G_2*.*

Unentscheidbare Probleme für kontextfreie Grammatike

5.32 Satz

Es ist nicht entscheidbar, ob für eine kontextfreie Grammatik G die Sprache $L(G)$ eindeutig ist.

5.33 Definition

Die **Sprache $\mathrm{B_M}$ der korrekten Rechenwege einer TM** M *besteht aus allen Worten*

$w_1\#w_2^R\#w_3\#w_4^R\ldots w_n^R\#$, *falls* n *gerade und*
$w_1\#w_2^R\#w_3\#w_4^R\ldots w_n\#$, *falls* n *ungerade,*

wobei w_i *Konfigurationen für alle* $1\leq i\leq n$, w_1 *eine Anfangskonfiguration,* w_n *eine akzeptierende Konfiguration und* w_{i+1} *direkte Nachfolgekonfiguration zu* w_i *für alle* $1\leq i\leq n-1$ *ist.*

5.34 Lemma

Für alle Turingmaschinen M *ist* B_M *der Durchschnitt zweier Sprachen* $L(G_1)$ *und* $L(G_2)$*, für die* G_1 *und* G_2 *kontextfrei sind.*

5.35 Lemma

Sei M ein TM, die auf jeder Eingabe mindestens zwei Rechenschritte ausführt. Die Sprache B_M ist genau dann kontextfrei, wenn $L(M)$ endlich ist.

5.36 Lemma

Für jede TM M *ist das Komplement* B_M^c *von* B_M *kontextfrei.*

5.37 Satz

Seien G, G_1, G_2 kontextfreie Grammatiken über Σ. Es ist nicht entscheidbar, ob

- $L(G)^c$ *kontextfrei ist*
- $L(G_1) \cap L(G_2)$ *kontextfrei ist*
- $L(G) = \Sigma^*$
- $L(G_1) = L(G_2)$
- $L(G_1) \subseteq L(G_2)$.

Unentscheidbare Probleme für kontextfreie Grammatike

5.38 Satz
Die Sprache $L = \{a^i b^j c^k | i = j$ *oder* $j = k\}$ *ist inhärent mehrdeutig.*