Formale Sprachen und Automaten

Lischke

SS 2004

Vorwort

Dieses Dokument wurde als Skript für die auf der Titelseite genannte Vorlesung erstellt und wird jetzt im Rahmen des Projekts „Vorlesungsskripte der Fakultät für Mathematik und Informatik“ weiter betreut. Das Dokument wurde nach bestem Wissen und Gewissen angefertigt. Denoch garantiert weder der auf der Titelseite genannte Dozent, die Personen, die an dem Dokument mitgewirkt haben, noch die Mitglieder des Projekts für dessen Fehlerfreiheit. Für etwaige Fehler und dessen Folgen wird von keiner der genannten Personen eine Haftung übernommen. Es steht jeder Person frei, dieses Dokument zu lesen, zu verändern oder auf anderen Medien verfügbar zu machen, solange ein Verweis auf die Internetadresse des Projekts `http://uni-skripte.lug-jena.de/` *enthalten ist.*

Diese Ausgabe trägt die Versionsnummer 1701 und ist vom 21. Oktober 2008. Eine (mögliche) aktuellere Ausgabe ist auf der Webseite des Projekts verfügbar.

Jeder ist dazu aufgerufen, Verbesserungen, Erweiterungen und Fehlerkorrekturen für das Skript einzureichen bzw. zu melden oder diese selbst einzupflegen – einfach eine E-Mail an die Mailingliste `<uni-skripte@lug-jena.de>` *senden. Weitere Informationen sind unter der oben genannten Internetadresse verfügbar.*

Hiermit möchten wir allen Personen, die an diesem Skript mitgewirkt haben, vielmals danken:

- *Jörg Sommer* `<joerg@alea.gnuu.de>` *(2004)*
- *Jens Kubieziel* `<jens@kubieziel.de>` *(2006)*

Inhaltsverzeichnis

Auflistung der Theoreme

Sätze

Definitionen und Festlegungen

1 Einführung und Wiederholung

Definition 1
L heißt *(formale) Sprache* über dem Alphabet X (d. h. X ist endl. Menge) $:\Longleftrightarrow L \subseteq X^*$

$X^* \ldots$	Vereinigung aller Teilmengen
$e \ldots$	leeres Wort
$X^+ =$	$X^* \setminus \{e\}$

Definition 2
$G = [X, N, S, R]$ heißt *(generative) Grammatik* $:\Longleftrightarrow$

1. X endl. Menge (Menge terminaler Symbole)
2. N endl. Menge (Menge nichtterminaler oder Hilfsymbole), $X \cap N = \emptyset$
3. $S \in N$ Anfangs- oder Startsymbol
4. $R \subseteq (X \cup N)^* \times (X \cup N)^*$ (R endl. Regelmenge)

Die Paare $[p, q] \in R$ heißen Regel und werden oft auch als $p \longrightarrow q$ geschrieben.

Definition 3
w' heißt unmittelbare Ableitung von w bezgl. einer Grammatik $G = [X, N, S, R]$ $w \longrightarrow w :\Leftrightarrow$

$$w, w' \in (X \cup N)^* \wedge \exists p \exists q \exists p_1 \exists p_2 (p, q, p_1, p_2 \in (X \cup N)^* \wedge [p, q] \in R \wedge w = p_1 p p_2 \wedge w' = p_1 q p_2)$$

w' heißt Ableitung aus w bezgl. G, $w \longrightarrow *w' \Leftrightarrow$ oder es gibt endl. viele Wörter $w_0, w_1, \ldots w_r$ mit $w_0 = w, w_{i-1} \underset{G}{\rightarrow} w_i$ für $i = 1, \ldots, r$ und $w_r = w'$. Man nennt dies auch eine Ableitung der Länge r und schreibt dafär $w_0 \Rightarrow w_1 \Rightarrow \ldots \Rightarrow w_r$ oder $w_0 \overset{r}{\rightarrow} w_r$

Definition 4
L heißt die von der Grammatik $G[\ldots]$ erzeugte Sprache, $L = L(G) :\Leftrightarrow L = \{w : S \underset{G}{\overset{*}{\rightarrow}} w\} \cap X^*$

Definition 5
Zwei Grammatiken G_1 und G_2 heißen äquivalent $G_1 \sim G_2 :\Leftrightarrow L(G_1) = L(G_2)$

Definition 6
Eine Grammatik $G = [X, N, S, R]$ heißt *Normalfomgrammatik* $:\Leftrightarrow R \subseteq N^+ \times (X \cup N)^*$

Satz 1
Zu jeder Grammtik gibt es eine äquivalente Normalformgramatik.

Beweis:
Es sei $G = [X, N, S, R]$ belibige Grammatik und zunächst $[\alpha, \beta] \notin R$ für jedes Wort p

$X' := x' : x' \in X$ wobei $X' \cap X = X' \cap N = \emptyset$

R'...Menge aller Regeln, die aus der Regel der Menge R entstehen, in dem darin alle Buchstaben $x \in X$ durch die entspechenden Buchstaben $x' \in X'$ ersetzt werden.

$R'' := \{[x', x] : x \in X\}$ und $G' := [X, N \cup X', S', R' \cup R''] \longrightarrow G' \sim G$

Ist $[e, p] \in R$ für gewisse p, so sei R' wie oben mit Ausnahme aller Regeln $[e, p']$.

Sei $E \notin X' \cup N$ und $R_E := \{[B, BE] : B \in X \cup N\} \cup \{[B, EB] : B \in\in N\} \cup \{[E, p'] :$ $[e, p] \in R \wedge p'$ entsteht aus p durch Ersetzen aller x in X durch $x \in X'\}$,

$G_E := [N, N \cup X' \cup \{E\}, S, R' \cup R'' \cup R_E] \longrightarrow G_E \sim G.$ ∎

Avram Noam Chomsky

Definition 7
Eine Grammatik $G = [X, N, S, R]$ (ohne Einschränkung an die Regelmënge R) heißt Grammatik vom *Typ 0*

G heißt Grammatik vom *Typ 1* oder nicht verkürzende Grammtik $:\Leftrightarrow \forall p \forall q([p, q] \in R \rightarrow l(p) \leq l(q))$

G heißt Grammatik vom Typ 1' oder Kontextgrammtik $\Longleftrightarrow \forall p \forall q([p, q]) \in R \rightarrow \exists p_1 \exists p_2 \exists u \exists B(p_1, p_2, u \leq (X \cup N) * \wedge B \in N \wedge u \neq e \wedge p = p_1 B p_2 \wedge q = p_1 \cup p_2$

G heißt Grammtik vom Type 2 oder kontextfreie Grammtik $:\Leftrightarrow \forall p \forall q([p, q] \in R) \rightarrow p \in N \wedge q(X \cup N)^+$

G ehißt Grammtik vom Typ 3, Automaten- oder rechtslineare Grammtik $:\Leftrightarrow \forall p, q([p, q] \in R \rightarrow p \in N \wedge q \in X \cdot N \cup X)$

Eine Sprache L heißt vom Typ i, wenn sie von einer Grammtik vom Typ i erzeugt wird. $rund L_i$ ist die Klasse aller Sprachen vom Typ i ($i \in \{0, 1, 1', 2, 3\}$).

Definition 8
Ist L eine Sprache von Typ i, so heißt auch $L \cup \{e\}$ Sprache vom Typ i ($i \in 0, 1, 1', 2, 3$)

Ist $G = [X, N, S, R]$ eine Grammtik vom Typ i und $S' \notin X \cap N$, so heißt auch $G' := [X, N \cup S, S, \{[S', e], [S', S]\} \cup R]$ Grammatik vom type i ($i \in \{0, 1, 1', 2, 3\}$)

Satz 2
Jede Grammtik $G = [X, N, S, R]$ mit $R \subseteq N \times (X \cup N)^*$ gibt es eine äquivalente Grammatik $G = [X, N, S, R]$ mit $R \subseteq N \times (X^*N \cup X^*)$ gibt es eine äquivalente rechtslineare Grammatik.

BEWEIS:

1. $G = [X, N, S, R]$ ist Grammatik mit $R \subseteq N \times (X \cup N)^*$

   Falls keine Regel $[A, e] \in R$, so G kontextfrei gemäß D7, andernfalls konstruiere G' nach Algo

Algo:

- setze $W_0 := A : [A, e] \in R$ und $W_{i+1} := W_i \cup A : \exists q([A, q] \in R \wedge q \in W_i^*)$ solange bis $W_{i+1} = W_i$
- Dann $W := W_i, G' = [X, N, S, R']$ mit $R' := [p, q'] : q \neq e \wedge \exists q([p, q] \in R \wedge (q' = q \vee q$ entsteht a G' kontextfrei gemäß D7 und $A \in W \longleftrightarrow A \in N \wedge A \xrightarrow[G]{*} e$.

Behauptung: $L(G') = L(G)\{e\} \subseteq$: $B \longrightarrow w' \longrightarrow [B, w'] \in R' \longrightarrow$ ex. $w = b_1 \dots b_n$ mit $b_r \in X \cup N, [B, w] \in R \wedge w' = b_{i_1} \dots b_{i_k}$, wobei $b_{i_1} \dots b_{i_k}$ Teilfolge von $b_1 \dots b_n undb \in W$ für $b \in \{b_1, \dots, b_n\} \setminus \{b_{i_1}, \dots, b_{i_k}\} \to B \xrightarrow[G]{*} w'$.

$p \in L(G') \longrightarrow S \to^*_{G'} p \longrightarrow S \Rightarrow^*_G p \longrightarrow p \in L(G).L \notin L(G')$, da nicht verkürzend $\longrightarrow L(G') \subseteq L(G) \setminus \{e\}$

rückw $\subseteq$: $G' := [X, N, S, R \cup R'] \longrightarrow L(G) \subseteq L(G') \longrightarrow L(G) \setminus \{e\} \subseteq L(G') \setminus \{e\} \longrightarrow$ reicht z. z. $L(G'') \setminus \{e\} \subseteq L(G)$

Entsprechend ex. $w \in L(G) \setminus \{e\}, w \notin L(G')$.

Es sei w_0

$\Rightarrow_G w_1 \Rightarrow \dots \Rightarrow_G w_t$ eine Ableitung von w aus S bzgl. G'' kürzester Länge $w_0 \Rightarrow S, w_t = w$

Da $w \in L(G'') \setminus L(G')$ auf wenigstens einmal eine Regel aus $R \setminus R'$ angewandt worden sein, d. h. eine Regel $[A, e]$ erwa bei $w_i \Rightarrow_{G''} w_{i+1} \longrightarrow A \in W, w_i = w_{i+1}Aw_{i+2}, w_{i+1} = w_1w_2 \to ej < imitw_j = w_{j+1}Bw_{j2}, w_{j+1} = w_{j1}q_jAq_iw_{j2}, q_1q_2 \neq q \longrightarrow [B, q_1Aq_2] \in R', w_{j1}q_1 \Rightarrow^{i-(i+1)}_{G''} w_1, q_1w_{j2} \Rightarrow^r_{G''} w_{j2}, 0 \leq i \leq i-(j+1) \longrightarrow [B, q_1q_2] \in R' \longrightarrow w_q \longrightarrow^j_G w_1Bw_2 \longrightarrow_G w_{j1}q_1q_2w_{j2} \longrightarrow^{i-j(+1)}_{G''} w_1w_2 \longrightarrow^{t+(i+1)}_{G''}$ ist Ableitung von w aus S der Länge $t-1$ Widerspruch dazu, dass t eine stets kürzeste Länge ist. $\longrightarrow L(G'') \setminus \{e\} \subseteq L(G')$

Ist $q_1q_2 = e[B, A] \in R' \longrightarrow B \in W \longrightarrow \alpha j' < jnitw_j = w_{j'1}Cw_{j'2}w_{j'+1} = w_{j'+1} \dots\dots \longrightarrow$ Schluss analog mit BX statt AB

$e \notin L(G) \longrightarrow G \sim G'$

$e \in L(G) \longrightarrow$ ex. kontextfreie gram. für $L(G'') \cup e$ nach D7! ∎

Satz 3
$REG = L_3 \subset L_2 \subset L_1 = L_{1'} \subset REC \subset RE = L_0$

Definition 9
7" $CS := L_1 = L_{1'}$ heißt Klasse der kintextsensitiven oder kontextsprachen

$CF := L_2$ heißt Klasse der kontextfreien oder Chompsky-Spachen.

$REG := L_3$ heißt Klasse der regulären, Automaten- oder rechstlinearen Sprachen.

Satz 4
$REG \subset CF \subset CS \subset REC \subset RE = L_0$

Beweis:
$CF \subset noteq$

$L := \{a^n b^n c^n : n \in N \backslash \{0\}\} G = [a, b, c, S, A, B, S, [S, aSBA], [A, abA], [AB, BA], [bB, bb], [A, c]] \longrightarrow L(G) = L$ es ist leicht ersichtlich, dass G L erzeugt und G kontextsensitiv ist. ∎

Satz 5 (Pumping-Lemma; Bar-Hillit, Perles, Shamir)
Zu jeder kontextfreien Sprache L existiert eine Zahl $n \in N$, so dass jedes Wort $w \in L$ mit $l(w) > m$ in der Fom $w = w_1 w_2 w_3 w_4 w_5$ dargestellt werden kann, wobei $w_2 w_4 \neq e, l(w_2 w_3 w_4) < m$ mit $w_1 w_2 w_3 w_4 w_5 \in L$ für jedes $i \in N$.

Satz 6
Jede kontextfreie Sprache über einbuchstabigem Alphabet ist regulär.

Beweis:
Sei $L \in CE, L \subset a^*$.

Nach Satz 4 existiert $n \in N$, so dass für jedes Wort $a^l \in L$ mit $l > m$ folgt: $l = i + j$ mit $i \leq j \leq < mmita^{i+kj} \subseteq L$ für jedes $k \in N$

$L_0 := a^i : 1 \leq i \leq m \cap L$ Falls $L \setminus L_0 \neq \emptyset seil_1 := minl : a^l \in L \setminus L_0$

Nach Satz 4 ist $l_1 = i_1 + j_1$ mit $1 \leq j_1 \leq m, 0 \leq i_1 \leq m$ mit $a^{i_1 + kj_1} : k \in N \subseteq L$

$L_1 := a_{i_1 + kj_1} : k \in N$

Seien $L_1, L_2, \ldots, L_{i'}$ konstruiert mit $L \backslash \cup_{i=0}^{i'}, sol_{i'+1} := minl : a^l \in L \setminus \cup_{r=0}^{i'} L_r \longrightarrow l_{i+j} = i + jmit1 \leq j \leq m, 0 \leq i$ und $L_{i+1} := a^{i+kj} : k \in N \subseteq L$.

Nach endl. vielen Schritten muss $L \backslash \cup_{r=0}^{i} L_r = \emptyset$ eintreten $- \longrightarrow L = \cup_{r=0}^{i} L_r, L_0, \ldots, L_i \in REG \longrightarrow L \in REG$ ∎

$(A4) CF_1 \subset CS_1 \subset REC_1 \subset RE_1$