Astérisque

GÜNTER HOTZ

Der Satz Von Chomsky-Schützenberger und die schwerste kontextfreie sprache von S. Greibach

Astérisque, tome 38-39 (1976), p. 105-115

<http://www.numdam.org/item?id=AST_1976__38-39__105_0>

Société Mathématique de France
Astérisque 38-39 (1976) p.105-115

DER SATZ VON CHOMSKY-SCHÜTZENBERGER
UND DIE SCHWERSTE KONTEXTFREIE SPRACHE VON S. GREIBACH

von

Günter HOTZ

ABSTRACT

The theorem of S. Greibach about a hardest context-free language can be derived from a normal form of the theorem of Chomsky-Schützenberger which was proved in [Ho,2]. We don't prove the theorem of S. Greibach in exactly the original form, but we use the construction of a syntactical algebra as defined in [Ho,1]. In this form the theorem states the existence of a certain universal acceptor for context-free languages.

EINLEITUNG.-Der Satz von Chomsky-Schützenberger [Ch-Sch] und der Satz von Greibach [Gr] über eine schwerste kontext-freie Sprache sind Darstellungssätze kontext-freier Sprachen. In beiden kommt der Begriff der Grammatik nicht vor. Beide Sätze beweist man, indem man von kontext-freien Grammatiken ausgehend die Ableitungen in diesen Grammatiken kodiert.

Da es sich bei beiden Sätzen um fundamentale Theoreme der Theorie der formalen Sprachen handelt, ist es gerechtfertigt, zu untersuchen, ob sich

----

This theorem was announced and a proof for it was sketched at the conference on "Automatentheorie und Formale Sprachen" in Oberwolfach from November 23, 1975 to November 29, 1975.

beide Sätze auseinander ableiten lassen, ohne den Weg über die Konstruktion kontext-freier Grammatiken zu wählen.

Wir werden hier den Satz von Greibach aus dem Satz von Chomsky-Schützenberger herleiten. Die hierbei verwendete Technik ist aber darüber hinaus interessant, da sie eine neue Idee enthält, algorithmische Probleme in kontext-freien Sprachen zu kodieren. Als Beispiel seien die Entscheidungsprobleme $A \cdot B = C$ für Matrizen oder Polynome A, B, C genannt, die sich durch diese Technik als Wortprobleme kontext-freier Sprachen erkennen lassen [Ho, 1].

Der mathematische Hintergrund bildet die Verwendung von syntaktischen Algebren, die in [Ho, 1] eingeführt wurden. Die Technik besteht in einer geschickten Verwendung von Nullteilern, die bewirken, daß unerwünschte Wörter annulliert werden.

DER SATZ VON CHOMSKY-SCHÜTZENBERGER

Ist $L \subset T^*$ eine kontext-freie Sprache, dann gibt es zu L ein Alphabet $X_k = \{x_o, x_1, \ldots, x_{k-1}, x_o^{-1}, \ldots, x_{k-1}^{-1}\}$ und einen Monoidhomomorphismus $\varphi : X_k^* \to T^*$ mit folgenden Eigenschaften : Es gibt ein Standardereignis $R \subset X_k^*$ derart, daß

$$\varphi(R \cap D_k) = L$$

ist, worin D_k die Dycksprache mit den Klammerpaaren x_i, x_i^{-1} ist. Es kann dabei ohne Beschränkung der Allgemeinheit angenommen werden, daß die Menge a der Anfangszeichen von R, der Menge b der Endzeichen von R und die Menge r der Nachbarschaftsrelationen folgenden Voraussetzungen erfüllen :

1. $r \subset (X_k - b) \times (X_k - a)$.

2. Kein Zeichen von X_k ist überflüssig, d.h. zu jedem $y \in X_k$ gibt es $u, v \in X_k^*$ mit $u y v \in R$.

3. $a \subset \{x_o, \ldots, x_{k-1}\}$ und $b = \{y^{-1} | y \in a\}$.

4. $\varphi(w) \in T$ und $\underset{u,v}{\exists}\ u w v \in (R \cap D_k) \Rightarrow$ Länge $(w) \leqslant 3$.

Die Voraussetzungen 1., 2., und 3 sind wohl bekannt, wegen 4. sehe man [Ho,2].

DER SATZ VON S. GREIBACH

Man erhält die von S. Greibach angegebene schwerste kontext-freie Sprache auf die folgende Weise : Man zerlegt Wörter $w \in X_2^*$ in Produkte, die man durch Trennzeichen "|" und ";" wie folgt unterteilt :

$$|w_{11};\dots;w_{1k_1}|w_{21};\dots;w_{2k_2}|\dots|w_{r1};\dots;w_{rk_r}|.$$

Die Wörter $\tilde{w}$ über dem Alphabet $\{x_o, x_o^{-1}, x_1, x_1^{-1}, |, ;\}$, die man so erhält versieht man noch mit einem Anfangszeichen \$. Ein Wort $\$\tilde{w}$ liegt genau dann in der Sprache G, wenn es eine Folge $i_1,\dots,i_r$ mit $1 \leqslant i_\ell \leqslant k_\ell$ für $\ell = 1,\dots,r$ gibt mit :

$$w_{1i_1} \cdot w_{2i_2} \cdot \dots \cdot w_{ri_r} \in D_2.$$

Der Satz von S. Greibach besagt nun, daß es zu jeder kontext-freien Sprache $L \subset T^*$ einen Homomorphismus :

$$\mu : T^* \to \{x_o, x_o^{-1}, x_1, x_1^{-1}, |, ;, \$\}^*$$

gibt mit : $\qquad L = \mu^{-1}(G)$

für $1 \notin L$ und $\qquad L = \mu^{-1}(\{1\} \cup G)$

falls $1 \in L$.

DIE SYNTAKTISCHE ALGEBRA $\mathbb{B} < D_2 >$

Wir verwandeln ~~unser~~ Wort $\tilde{w}$ in ein Produkt von Summen, indem wir + für ; und . für | schreiben. Wir erhalten dann den Ausdruck :

$$A = (w_{11} + \dots + w_{1k_1}) \cdot \dots \cdot (w_{r1} + \dots + w_{rk_r}).$$

Rechnen wir diesen Ausdruck unter Verwendung der Rechenregel :

$$x_i x_i^{-1} = 1, \quad x_i x_j^{-1} = 0 \quad \text{für} \quad i \neq j, \; i,j = 0,\dots,k-1$$

aus, dann erhalten wir für $\$\tilde{w} \in G$

$$A = 1 + u.$$

Dies führt uns zu einer neuen Definition der Sprache G, die wir beim Beweis unseres angekündigten Satzes verwenden.

Wir definieren nun die syntaktische Algebra $\mathbb{B} < D_2 >$.

Sei $\mathbb{B} = \{0,1\}$ die boolesche Algebra aus zwei Elementen und sei $\mathcal{V}_2$ das syntaktische Monoid von D_2. Dann bilden wir Polynome mit Koeffizienten aus $\mathbb{B}$ und Monome aus $\mathcal{V}_2$, d.h. Summen

$$p = \sum_{m \in \mathcal{V}_2} \alpha_m \cdot m \quad \text{mit} \quad \alpha_m \in \mathbb{B}$$

und $\alpha_m \neq 0$ für nur endlich viele m.

Wir setzen :

$$(\Sigma \alpha_m \cdot m) + (\Sigma \beta_m \cdot m) = \Sigma(\alpha_m + \beta_m) \cdot m$$

und

$$(\sum_m \alpha_m \cdot m) \cdot (\sum_m \beta_m \cdot m) = \sum_m \Big(\sum_{u.v = m} \alpha_u \beta_v \Big) \cdot m .$$

Die so definierte Addition und Multiplikation sind assoziativ, die Multiplikation ist distributiv. Die so erhaltene Algebra bezeichnen wir als die syntaktische Algebra von D_2 über $\mathbb{B}$. In natürlicher Weise läßt sich $\mathbb{B}$ und $\mathcal{V}_2$ in $\mathbb{B} < D_2 >$ einbetten.

Wir bezeichnen die Menge :

$$E = 1 + \mathbb{B} < D_2 >$$

als die Menge der Einheiten von $\mathbb{B} < D_2 >$. Setzen wir den Homomorphismus μ und den kanonischen Homomorphismus, der Ausdrücken über $\mathbb{B} < D_2 >$ ihren Wert zuweist, zu einem Monoidhomomorphismus $\mu' : T^* \to \mathbb{B} < D_2 >$ zusammen, dann erhalten wir, wenn wir zunächst das Zeichen \$, das in $\mathbb{B} < D_2 >$ noch nicht vorhanden ist, vernachlässigen,

$$L = \mu'^{-1}(\$E) \text{ bzw. } L = \mu'^{-1}(\{1\} \cup \$E).$$

In dieser Form werden wir den Satz von S. Greibach beweisen.

EIN HOMOMORPHISMUS $\psi : X_k^* \to \mathbb{B}<D_2>$

Wir kodieren nun $D_k \cap R$ in $\mathbb{B}<D_2>$. Hierzu definieren wir uns einen Homomorphismus $\psi : X_k^* \to \mathbb{B}<D_2>$.

Sei :
$$2^{n-1} \leqslant k < 2^n$$

und :
$$\alpha(x_i^\epsilon) = x_{\alpha_0} x_{\alpha_1} \dots x_{\alpha_n} \qquad \text{für} \quad i = 0,\dots,k-1 \ ; \ \epsilon = 1,-1.$$

Hierin ist $\alpha_0,\alpha_1 \dots \alpha_{n-1}$ die Darstellung von i im Dualsystem, wobei alle Dualzahlen durch eventuelle Hinzufügung führender Nullen auf gleiche Länge gebracht sind. Es ist $\alpha_n = 1$ für $\epsilon = 1$ und $\alpha_n = 0$ für $\epsilon = -1$. α bildet X_k injektiv in $\mathcal{V}_2$ ab.

Nun setzen wir für $x \in X_k$:

$$\rho(x) = \begin{cases} \alpha(x)\, x_0 & \text{falls es ein } y \in X_k \text{ gibt mit } (xy) \in r \\ 1 & \text{sonst.} \end{cases}$$

Gibt es ein y mit $(x,y) \in r$, dann setzen wir :

$$\lambda(x) = v(x) \,.\, \sum_{(y,x) \in r} (\alpha(y))^{-1} \,.$$

Ist $(y,x) \notin r$ für alle $y \in X_k$, dann setzen wir :

$$\lambda(x) = v(x).$$

Hierin ist :

$$v(x) = \begin{cases} x_0^{-1} & \text{falls es ein } y \in X_k \text{ gibt mit } (y,x) \in r, \\ x_1^{-1} & \text{sonst.} \end{cases}$$

Nun definieren wir :

$$\psi(x_i^\epsilon) = \lambda(x_i^\epsilon)(\alpha(x_i))^\epsilon \ \rho(x_i^\epsilon) \qquad \text{für } x_i^\epsilon \in X_k$$

und setzen anschließend ψ zu einem Homomorphismus von X_k^* in $\mathbb{B}\langle D_2\rangle$ fort.

Wir betrachten :

$$\psi(x^\epsilon \,.\, y^\eta) = \lambda(x^\epsilon)(\alpha(x))^\epsilon \rho(x^\epsilon)\, \lambda(y^\eta)(\alpha(y))^\eta \,.\, \rho(y^\eta)$$

und zeigen

<u>LEMMA 1</u>.-

$$\rho(x)\,.\,\lambda(y) = \begin{cases} 0 & \text{für } (x,y) \notin r \text{ und } x \notin b \\ 1 & \text{für } (x,y) \in r \\ \lambda(y) & \text{für } (x,y) \in r,\ x \in b. \end{cases}$$

<u>Beweis</u>.- Für $x \in b$ ist $\rho(x) = 1$ und die Behauptung also erfüllt. Für $x \notin b$ gibt es y mit $(x,y) \in r$ und also ist dann $\rho(x) = \alpha(x)\,.\,x_o$. Ist $y \in a$, dann ist $\lambda(y) = x_1^{-1}$ und deshalb $\rho(x)\,.\,\lambda(y) = 0$. Ist $y \notin a$, dann ist $v(y) = x_o^{-1}$ und in diesem Falle :

$$\rho(x)\,.\,\lambda(y) = \alpha(x) \sum_{(z,y)\in r} (\alpha(z))^{-1} \,.$$

Nun ist :

$$\alpha(x)\,.\,(\alpha(z))^{-1} = \begin{cases} 0 & \text{für } x \neq z \\ 1 & \text{für } x = z. \end{cases} \tag{1}$$

Da unser Grundring eine boolesche Algebra ist, gilt also :

$$\alpha(x) \sum_{(z,y)\in r} (\alpha(z))^{-1} = \begin{cases} 0 & \text{für } (x,y) \notin r \\ 1 & \text{für } (x,y) \in r. \end{cases}$$

Damit ist unser Lemma bewiesen.

Hieraus ergibt sich nach nochmaliger Anwendung von (1)

$$\psi(x_i^\epsilon \,.\, x_j^\eta) = \begin{cases} 0 & \text{für } x_i^\epsilon \notin b,\ (x_i^\epsilon, y_j^\eta) \notin r \ \text{ oder } (x_i^\epsilon, x_j^\eta) \quad r,\ i \neq j,\ \epsilon = 1,\ \eta = -1 \\ \lambda(x_i)\,\rho(x_i^{-1}) \ \text{ für } (x_i, x_i^{-1})\, r\,,\ i = j,\ \epsilon = 1,\ \eta = -1\,; \\ \lambda(x_i^\epsilon)(\alpha(x_i))^{-1}\, \lambda(\chi_j^\eta)(\alpha(\chi_j))^\eta \rho(\chi_j^\eta) & \text{für } x_i^\epsilon \in b\,; \\ \lambda(x_i^\epsilon)(\alpha(x_i))^\epsilon\, (\alpha(\chi_j))^\eta \rho(\chi_j^\eta) & \text{sonst.} \end{cases}$$

Hieraus folgt :

LEMMA 2.- Aus $\psi(w) \neq 0$ folgt, daß es eine Zerlegung $w = w_1 w_2 \cdot \ldots \cdot w_\ell$ gibt und Worte $u_1, \ldots, u_\ell, v$, so daß:

$$u_i w_i \in R, \; u_\ell w_\ell v \in R \quad \text{für } i = 1, \ldots, \ell-1.$$

Beweis.- Aus $\psi(w) \neq 0$ folgt für $w = y_1 \cdot \ldots \cdot y_m$:

$$\psi(y_i \cdot y_{i+1}) \neq 0.$$

Aus der obigen Relation entnimmt man :

$$(y_i, y_{i+1}) \in r \quad \text{oder} \quad y_i \in b.$$

Wir zerlegen nun w so in ein Produkt $w_1 \cdot \ldots \cdot w_\ell$, daß letzte Zeichen von w_i aus b ist, aber kein anderes Zeichen von w_i für $i = 1, \ldots, \ell-1$. w_ℓ enthält kein Element aus b. Da in X_k keine überflüssigen Zeichen vorkommen, läßt sich jedes w_i nach links mittels eines $u_i \in X_k^*$ so fortsetzen, daß $u_i w_i \in R$ ist für $i = 1, \ldots, \ell-1$. Für w_ℓ findet man ein u_ℓ und v mit $u_\ell w_\ell v \in R$. Dies war zu zeigen.

LEMMA 3.- $$\psi(w) = X_1^{-1} \Longleftrightarrow w \in R \cap D_k.$$

Beweis.- Zunächst zeigen wir "$\Longrightarrow$". Sei $w = y_1^{\epsilon_1} \cdot \ldots \cdot y_m^{\epsilon_m}$ mit $y_i^{\epsilon_i} \in X_k$. Da $\psi(y_1^{\epsilon_1}) = \lambda(y_1^{\epsilon_1})u$ ist und da $\lambda(y_1^{\epsilon_1})$ nur Elemente mit negativen Exponenten enthält, können sich diese Faktoren nicht wegkürzen. Da $\psi(w) = x_1^{-1}$ ist, kann $\lambda(y_1^{\epsilon_1})$ also nur x_1^{-1} enthalten. Also ist $y_1 \in a$.

Nach Lemma 2 können wir also $w = u.v$ mit $u \in R$ bilden, oder aber w durch ein Wort v zu $w.v \in R$ verlängern. Wir betrachten beide Fälle getrennt.

1. Es gebe ein v mit $w.v \in R$.

   In diesem Falle haben wir :

$$\psi(w) = x_1^{-1}\, \alpha(y_1) \,.\, (\alpha(y_2))^{\epsilon_2} \ldots (\alpha(y_m))^{\epsilon_m}.$$

Da $\psi(w) = x_1^{-1}$ gilt :

$$\alpha(y_1) \, . \, (\alpha(y_2))^{\epsilon_2} \ldots (\alpha(y_m))^{\epsilon_m} = 1.$$

Nach (1) läßt sich das Kürzen dieses Wortes so vornehmen, daß man stets ganze α-Faktoren gegeneinander weghebt. Da α injektiv ist, kann man diese Kürzungen in w nachvollziehen. Also ist $w \in D_k$.
Nach Voraussetzung über R Kürzt sich y_1 nur gegen ein $y \in b$. Dies kann in diesem Falle nur $y_m^{\epsilon_m}$ sein. Also ist $w \in R$ und $v = 1$. Damit ist für den ersten Fall $w \in D_k \cap R$ gezeigt.

2. Es gebe eine Zerlegung $w = u \, . \, v$ mit $u \in R$. Wir haben :

$$\psi(u) = x_1^{-1} \; \alpha(y_1) \ldots (\alpha(y_\ell))^{-1} \quad \text{mit } y_\ell^{-1} \in b.$$

Weiter ist $y_i^{\epsilon_i} \notin b$ für $i < \ell$. Da $\psi(u)$ Teilwort eines Wortes ist, das sich auf 1 kürzt und da sich y_ℓ^{-1} nicht von rechts her kürzen läßt, gibt es ein $i \leqslant \ell$ mit $y_i^{\epsilon_i} = y$. Dann ist $\epsilon_i = 1$ und $y_\ell \in a$. Jedes Element von a, das im inneren eines Wortes w steht, annulliert $\psi(w)$. Also sit $\ell = 1$. Kürzen sich aber y_1 und y_ℓ^{-1} gegenseitig, dann hebt sich auch das zwischen beiden stehende Teilwort zu 1 weg. Also haben wir $\psi(u) = x_1^{-1}$. Nun folgt weiter $\psi(v) = 1$. Ist $v \neq 1$, dann ist $\psi(v) = \lambda(y_{\ell+1}^{\epsilon_{\ell+1}}).v'$. Da λ nur Faktoren mit negativen Exponenten enthält, ist dann $\psi(v) \neq 1$. Also ist $v = 1$ und $u = w$ und also auch in diesem Falle $w \in R \cap D_k$.

Die anderen Richtung des Lemma 3 ist einfach zu beweisen. Wir überlassen diesen Beweis dem Leser.

DER BEWEIS DES SATZES VON S. GREIBACH

Wir wollen nun den Satz von S. Greibach aus dem Satz von Schützenberger-Chomsky ableiten. Wir haben die durch das folgende Diagramm veranschaulichte Situation :

$$\begin{tikzcd}(R \cap D_k) \subset X_k^* \arrow[r, "\varphi"] \arrow[d, "\psi"'] & T^* \supset L \arrow[dl, dashed, "\mu"] \\ \mathbb{B}\langle D_2\rangle & \end{tikzcd}$$

Wir wollen nun einen Monoidhomomorphismus μ konstruieren, so daß :

$$L = \mu^{-1}(x_1^{-1} + \mathbb{B}\langle D_2\rangle)$$

ist. Sei :

$$Y_m = X_k \cup X_k^2 \cup \ldots \cup X_k^m$$

und :

$$\mu(t) = \sum_{w \in \varphi^{-1}(t) \cap Y_m} \psi(w) \qquad \text{für } t \in T.$$

<u>LEMMA 4</u>.- Aus $\mu(u) = x_1^{-1} + v$ folgt $u \in L$.

<u>Beweis</u>.- Aufgrund der Definition von μ gibt es $w \in X_k^*$ mit :

$$\psi(w) = x_1^{-1} + v.$$

Nun ist :

$$\psi(w) = \psi(w_1) . \psi(w') \qquad \text{mit } w_1 \in X_k$$

und nach Voraussetzung x_1^{-1} Summand in $\lambda(w_1)$. Dann ist aber $\lambda(w_1) = x_1^{-1}$ und also :

$$\psi(w) = x_1^{-1} . u'$$

und weiter $u' = 1$. Also haben wir $\psi(w) = x_1^{-1}$, woraus durch Anwendung von Lemma 3 $w \in D_k \cap R$ folgt. Dann ist aber :

$$u = \varphi(w) \in L$$

womit das Lemma bewiesen ist.

Es bleibt zu zeigen, daß auch die Umkehrung dieses Lemmas gilt. Hierzu verwenden wir die Voraussetzung (4) über unsere Darstellung $\varphi : (D_k \cap R) = L$, nämlich daß :

$$\varphi(w) \in T \text{ und } \exists_{u,v}\ uwv \in (R \cap D_k) \Longrightarrow \text{Länge } (w) \leqslant 3. \qquad (4)$$

Hieraus folgt nämlich :

ist : $$u = t_1 \ldots t_m \in L,\ t_i \in T$$

dann ist wegen (4) :

$$(\varphi^{-1}(u) \cap Y_2^m) \cap (R \cap D_k) \neq \emptyset$$

da Y_2^m alle wörter der Länge m und m + 1 enthält, die auf u abgebildet werden.

Also ist :

$$\mu(u) = \psi(w) + v$$

mit $w \in (R \cap D_k)$ d.h.

$$\mu(u) = x_1^{-1} + v.$$

LEMMA 5.- $$w \in L \Longrightarrow \mu(w) = x_1^{-1} + v.$$

Damit haben wir :

SATZ 1.- Zu jeder kontextfreien Sprache $L \subset T^*$ gibt es einen Homomorphismus

$$\mu : T^* \to \mathbb{B}\langle D_2 \rangle$$

mit :

$$L = \mu^{-1}(x_1^{-1} + \mathbb{B}\langle D_2 \rangle).$$

Dieser Satz ist ein Analogon zu dem genannten Satz von S. Greibach. Man kann diesen Satz auch als einen sequentiellen Akzeptor für kontextfreie Sprachen interpretieren. Die Kompexität der Operationen in $B\langle D_2\rangle$ schätzt also die Komplexität der kontextfreien Sprachen nach unten ab . Das Interesse an dieser Algebra wird noch dadurch gesteigert, daß sich die Matrixmultiplikation und Polynommultiplikation treu in $B\langle D_2\rangle$ einbetten lassen.

* * *

REFERENCES

[Ch-Sch] CHOMSKY N. and SCHÜTZENBERGER,M.P. : The algebraic theory of context-free languages, in P. Braffort and S. Hirschberg eds., Computer Programming and Formal Systems, North-Holland, Amsterdam, 1970, 116-161.

[Gr] GREIBACH S., The Hardest context-free languages, SIAM J. Computing 2, 1973, 304,310.

[Ho,1] HOTZ, G. :"Untere Schranken für das Analyseproblem kontext-freier Sprachen" Technischer Bericht des Fachbereichs Angewandte Mathematik und Informatik der Universität des Saarlandes, XI/1975.

[Ho,2] HOTZ, G. : Normal form transformations of context-free languages, submitted for publication.

Günter HOTZ
Universität des Saarlandes
Angewandte Mathematik
und Informatik
D-6600 SAARBRUCKEN 15 (R.F.A.)