Habe Sie Verbesserungs / Veränderungs - Wünsche ?

TK

The Power of the Greibach Normal Form

by
Günter Hotz and Thomas Kretschmer
Universität des Saarlandes
Fachbereich Informatik
D–6600 Saarbrücken

Abstract Applying the fact that for any context–free language there exists a grammar in Greibach normal form, one can give easy algebraic proofs of Shamir's theorem, of Chomsky–Schützenberger's theorem and of Greibach's theorem on the existence of a hardest context–free language. These proofs will make it clear that there is a close connection between these results.

Introduction If one looks in standard textbooks on formal language theory (e.g. [Ha]) one will notice that the theorems of Chomsky–Schützenberger [CS] and of Greibach on the existence of a hardest context–free language [Gr] are proven in an ad–hoc manner. Also there seems to be no connection between both results. The theorem of Shamir [Sh] is rarely mentioned. In this paper we want to take up some ideas of [Ho]. First we will use the existence of the Greibach normal form to show a weaker version of Shamir's result. In its turn this version implies the existence of the Greibach normal form, so that essentially these two results are equivalent. With the help of this version of Shamir's theorem we will be able to give short algebraic proofs of the above–mentioned results of Chomsky–Schützenberger and of Greibach.

Notations Let T^* denote the free monoid over the finite alphabet T with identity 1. Define $T^+ := T^* \setminus \{1\}$. We write $\mathcal{P}_f(T^*)$ for the set of finite subsets of T^*. We make $\mathcal{P}_f(T^*)$ to a monoid with identity $\{1\}$ by introducing the following operation :

$$U \cdot V := \{uv \mid u \in U \text{ and } v \in V\}$$

for U,V finite subsets of T^*. If $\theta : X \to Y$ is a mapping, we write $x\theta$ for the image of x under θ. If $\theta' : Y \to Z$ is another mapping we denote the composition of θ and θ' by $\theta\theta'$. For a subset $U \subseteq Y$, $U\theta^{-1}$ is defined by $U\theta^{-1} := \{x \in X \mid x\theta \in U\}$.
A context–free grammar is denoted by $G = (V,T,P,S)$ where V is the set of variables, T is the set of terminals, $S \in V$ is the start symbol and

$$P \subset V \times (V \cup T)^*$$

is the finite set of productions. If $u \in (V \cup T)^*$ derives to $v \in (V \cup T)^*$, we write $u \xrightarrow{*} v$. For leftmost derivations we use the symbol $\underset{lm}{\xrightarrow{*}}$. G generates the language $L(G) := \{w \in T^* \mid S \xrightarrow{*} w\}$. We say that G is in Greibach normal form if and only if $P \subset V \times VT^*$.
Now we want to define the Dyck set over a finite alphabet X. Let $\overline{X}$ be another alphabet such that $X \cap \overline{X} = \emptyset$ and such that there is a bijection $b : X \to \overline{X}$. We write $\overline{x}$ for xb $(x \in X)$. Then the Dyck set $D(X)$ over X is generated by $G = (\{S\}, X \cup \overline{X}, P, S)$ where

$$P = \{S \to SS, S \to 1\} \cup \{S \to xS\overline{x} \mid x \in X\}$$

(see e.g. [Be]). The elements of a Dyck set are called Dyck words. For $u, v \in X^*$ we define $\overline{uv} = \overline{v} \cdot \overline{u}$.
The polycyclic monoid $X^{(*)}$ is defined as

$$(X \cup \overline{X})^* / \{x\overline{x} = 1 \mid x \in X\}$$

Modulo these relations, each Dyck word is congruent to the empty word.

A weak version of Shamir's theorem Now we are able to give a weak but nonetheless very useful version of Shamir's theorem [Sh].

Theorem 1 *For any context-free language $L \subseteq T^+$ there is an alphabet V, an element $S \in V$ and a monoid homomorphism*

$$\varphi : T^* \to \mathcal{P}_f((V \cup \overline{V})^*)$$

such that

$$w \in L \iff w\varphi \cap \overline{S}D(V) \neq \emptyset$$

Moreover it holds for all $t \in T$: $t\varphi \subset \overline{V}V^*$.

Proof Let $G = (V, T, P, S)$ be a grammar in Greibach normal form for L. Define φ by

$$t\varphi := \{\overline{A}u^\sigma \mid (A, tu) \in P\}$$

for $t \in T$. Here, u^σ denotes u mirrored, e.g. $(abcb)^\sigma = bcba$. We prove by induction on the length $|w|$ of w :

$$\forall A \in V, w \in T^*, u \in V^* : \quad A \underset{lm}{\xrightarrow{*}} wu \iff \exists m \in w\varphi : m\overline{u}^\sigma \in \overline{A}D(V) \qquad (*)$$

$|w| = 0$: The claim is true because it is immediately clear that $u = A$ and $m = 1$.
$|w| > 0$: Let $w = w't$ where $t \in T$.
"$\Rightarrow$" : There must be a derivation

$$A \underset{lm}{\xrightarrow{*}} w'Bu' \to w'tu''u' = wu$$

where $(B, tu'') \in P$ and $u''u' = u$. Because of the induction hypothesis we find an $m' \in w'\varphi$ such that $m'\overline{Bu'}^{\sigma} \in \overline{A}D(V)$. Because $\overline{B}u''^{\sigma} \in t\varphi$, we know that $m := m'\overline{B}u''^{\sigma} \in w\varphi = w'\varphi \cdot t\varphi$. We have to check that $m\overline{u}^{\sigma} \in \overline{A}D(V)$. But this is clear because

$$m\overline{u}^{\sigma} = m'\overline{B}u''^{\sigma} \cdot \overline{u}^{\sigma} = m'\overline{B}u''^{\sigma}\overline{u''}^{\sigma}\overline{u'}^{\sigma} \in \overline{A}D(V) \Longleftrightarrow m'\overline{Bu'}^{\sigma} \in \overline{A}D(V)$$

"$\Leftarrow$" : Let $m \in w\varphi$ such that $m\overline{u}^{\sigma} \in \overline{A}D(V)$. Because $w\varphi = w'\varphi \cdot t\varphi$ there must be $m' \in w'\varphi$ and $m'' \in t\varphi$ such that $m'm'' = m$. From $m'' \in t\varphi$ one concludes that there is a production $(B, tu'') \in P$ such that $m'' = \overline{B}u''^{\sigma}$ $(**)$. Hence $m\overline{u}^{\sigma} = m'm''\overline{u}^{\sigma} = m'\overline{B}u''^{\sigma}\overline{u}^{\sigma} \in \overline{A}D(V)$ and thus there is $u' \in V$ such that $u = u''u'$. It follows that $m'\overline{Bu'}^{\sigma} = m'\overline{Bu'}^{\sigma} \in \overline{A}D(V)$. With the help of the induction hypothesis and $(**)$ we get

$$A \underset{lm}{\overset{*}{\rightarrow}} w'Bu' \rightarrow w'tu''u' = wu$$

. ■

In the following corollary we formulate the above theorem in the language of Shamir [Sh]. Thus it will become clear that our theorem is just a weak version of Shamir's result. Those readers that are not used to the theory of semi–rings are to refer to [SS]. First we have to introduce some more notations. For a semi–ring R and a monoid M, $R\langle M\rangle$ denotes the semi–ring of polynomials over M with coefficients in R. If s is a polynomial and $m \in M$, $\langle s, m\rangle$ is the coefficient of m in s, $\mathbf{B}$ is the Boolean semi–ring.

Corollary 1 *For any context–free language $L \subseteq T^+$ there is an alphabet V, an element $S \in V$ and a monoid homomorphism*

$$\tilde{\varphi} : T^* \rightarrow \mathbf{B}\left\langle V^{(*)}\right\rangle$$

such that

$$w \in L \Longleftrightarrow \langle w\tilde{\varphi}, \overline{S}\rangle \neq 0$$

In a second corollary we want to get rid of the alphabet V that depends on the language L and replace it by a fixed alphabet $Y = \{x, y, \overline{x}, \overline{y}\}$.

Corollary 2 *For any context–free language $L \subseteq T^+$ there is a monoid homomorphism*

$$\varphi_2 : T^* \rightarrow \mathcal{P}_f(Y^*)$$

such that

$$w \in L \Longleftrightarrow w\varphi_2 \cap \overline{x}D(x, y) \neq \emptyset$$

Proof Let $V = \{v_0, \ldots, v_n\}$ where $v_0 = S$. Define a homomorphism

$$\pi' : (V \cup \overline{V})^* \to Y^*$$

by

$$\begin{array}{rcl} v_i & \mapsto & xy^i \\ \overline{v_i} & \mapsto & \overline{y}^i\overline{x} \end{array}$$

and extend π' to a monoid homomorphism

$$\pi : \mathcal{P}_f((V \cup \overline{V})^*) \to \mathcal{P}_f(Y^*)$$

One defines $\varphi_2 := \varphi\pi$ and immediatly sees that

$$w\varphi \cap \overline{S}D(V) \neq \emptyset \Longleftrightarrow w\varphi_2 \cap \overline{x}D(x,y) \neq \emptyset$$

■

We want to show that Theorem 1 implies the existence of the Greibach normal form. Let $L \subseteq T^+$ be any context–free language and let be φ the corresponding homomorphism from Theorem 1. Then we define

$$P := \{(A, tu) \mid \overline{A}u \in t\varphi(t \in T)\}$$

Here one needs the fact that $t\varphi \subset \overline{V}V^*$. We set $G = (V, T, P, S)$. Then there is the same connection between G and φ as in the proof of Theorem 1 and therefore $(*)$ holds also in this situation. Of course $(*)$ implies that $L(G) = L$.

The Theorem of Chomsky–Schützenberger Now it is very easy to prove the theorem of Chomsky–Schützenberger [CS].

Theorem 2 *For any context–free language $L \subseteq T^+$ there is a regular set K, a Dyck set D and a monoid homomorphism μ such that*

$$L = (K \cap D)\mu$$

Proof Let $G = (V, T, P, S)$ be a grammar in Greibach normal form for L. Define $Z := V \cup \overline{V} \cup T \cup \overline{T}$. Now we define a monoid homorphism which essentially is φ from Theorem 1.

$$\begin{array}{c} g : T^* \to \mathcal{P}_f(Z^*) \\ tg \mapsto \{\overline{A}u^\sigma t\bar{t} \mid (A, tu) \in P\} \end{array}$$

for $t \in T$. The only difference to φ is that we have appended the term $t\bar{t}$ to each element in the image of t. Of course this term can be neglected if one is interested only in Dyck words. In other words we get for all words $w \in T^+$:

$$\begin{aligned} & wg \cap \overline{S}D(V \cup T) \neq \emptyset \\ \iff \quad & w\varphi \cap \overline{S}D(V) \neq \emptyset \\ \iff \quad & w \in L \end{aligned}$$

In contrast to φ we don't lose any information on w when applying g to w because of the terms $t\bar{t}$. We define μ in such a way that we get w back, i.e.

$$\forall m \in wg : m\mu = w$$

Hence we define

$$\mu : Z^* \to T^*$$

by

$$\begin{aligned} A, \overline{A} & \mapsto 1 \\ \bar{t} & \mapsto 1 \\ t & \mapsto t \end{aligned}$$

for $A \in V$ and $t \in T$. Finally we let

$$K := \{S\} \cdot \Big(\bigcup_{t \in T} tg\Big)^*$$

Claim : $L = (K \cap D(V \cup T))\mu$
Proof of the claim :
"$\subseteq$" : Let $w = t_1 \cdot \ldots \cdot t_n \in L$ $(t_i \in T)$. Then $wg \cap \overline{S}D(V \cup T) \neq \emptyset$, i.e., there is an $m \in wg$ such that $Sm \in D(V \cup T)$. There must exist $m_i \in t_i g$ satisfying $m = m_1 \cdot \ldots \cdot m_n$. It follows that

$$m \in \Big(\bigcup_{t \in T} tg\Big)^*$$

i.e., $Sm \in K$ and thus $Sm \in K \cap D(V \cup T)$. Because $m_i\mu = t_i$ we get $(Sm)\mu = t_1 \cdot \ldots \cdot t_n = w$.
"$\supseteq$" : Let $w \in (K \cap D(V \cup T))\mu$. Then there exists an $m \in \left(\bigcup_{t \in T} tg\right)^*$ such that $Sm \in D(V \cup T)$ and $(Sm)\mu = w$. It follows that there is a factorization of m of the form $m = m_1 \cdot \ldots \cdot m_n$ where $m_i \in t_i g$ $(t_i \in T)$ and $w = t_1 \cdot \ldots \cdot t_n$. Hence $m \in wg$. Because $m \in \overline{V}Z^*$ and $Sm \in D(V \cup T)$, we get $m \in \overline{S}D(V \cup T)$ implying that $w \in L$. ∎

The Theorem of Greibach We will see that the theorem of Greibach on the existence of a hardest context–free language is just a reformulation of Corollary 2.

Theorem 3 *There is a context–free language $L_{Gr} \subseteq T_{Gr}^*$ such that :*
For any context–free language $L \subseteq T^+$ there is a monoid homomorphism ν

$$\nu : T^* \to T_{Gr}^*$$

such that

$$L = L_{Gr}\nu^{-1}$$

Proof Define

$$\begin{aligned} T_{Gr} &:= \{x, y, \overline{x}, \overline{y}, [,], +\} \\ L_{Gr} &:= \{[v_1^1 + \ldots + v_{n_1}^1][v_1^2 + \ldots + v_{n_2}^2] \ldots [v_1^l + \ldots + v_{n_l}^l] \mid \\ &\qquad l \in \mathbf{N}, v_j^i \in \{x, y, \overline{x}, \overline{y}\}^* \text{ and } \exists j_1, \ldots j_l : \\ &\qquad v_{j_1}^1 \cdot \ldots \cdot v_{j_l}^l \in \overline{x} D(x, y)\} \end{aligned}$$

and

$$\nu : T^* \to T_{Gr}^*$$

by

$$t \mapsto [u_1 + \ldots u_n]$$

where $t \in T$ and $t\varphi_2 = \{u_1, \ldots, u_n\}$ (φ_2 comes from Corollary 2).
Then it is very easy to check that $L = L_{Gr}\nu^{-1}$. ■

References

[**Be**] J. BERSTEL, Transductions and Context–Free Languages, B.G. Teubner, Stuttgart 1979.

[**CS**] N. CHOMSKY AND M. SCHÜTZENBERGER, The algebraic theory of context–free languages, in : P. Braffort and D. Hirschberg (edts.), Computer programming and formal systems, North–Holland, Amsterdam 1963, 118–161.

[**Gr**] S. GREIBACH, The hardest context–free languages, *SIAM Journal of Computing* **2** (1973) 304–310.

[**Ho**] G. HOTZ, A Representation Theorem of Infinite Dimensional Algebras and Applications to Language Theory, *Journal of Computer and System Sciences* **33** (1986) 423–455.

[**SS**] A. SALOMAA AND M. SOITTOLA, Automata–Theoretic Aspects of Formal Power Series, Springer–Verlag, New York, 1978.

[**Sh**] E. SHAMIR, A Representation Theorem For Algebraic and Context–Free Power Series in Noncommuting Variables, *Information and Control* **11** (1967) 239–254.