Abstract. The purpose of this note is to provide a sketch of the proof of the "strongest" form of the Chomsky-Schützenberger Theorem.

On the Chomsky-Schützenberger Theorem

Ronald V. Book

Research Report #33

Revised January 1976

The preparation of this paper was supported in part by the National Science Foundation under Grants GJ-30409 and DCR-75-15945.

An important result in the theory of context-free languages is that known as the "Chomsky-Schützenberger Theorem." The best known version of this result can be stated as follows.

Theorem A. For every context-free language L, there exist an integer k, a regular set R, and a homomorphism h such that $L = h(D_k \cap R)$, where D_k is the Dyck set on k letters.

Equivalently, one can state that every context-free language is the image of a Dyck set under a finite-state transduction. Theorem A appeared first in Chomsky [1] and Chomsky and Schützenberger [2]. Proofs appear in secondary sources such as Ginsburg [3] and Salomaa [8].

A stronger (in fact, the "strongest" possible) version of Theorem A is known, although no proof appears in the literature. First, one can replace D_k with $h_2^{-1}(D_2)$ for a suitable homomorphism h_2. Second, the homomorphism h can be made length-preserving if h_2 and R are suitable chosen. This leads to a result which is the "strongest" form of the Chomsky-Schützenberger Theorem.

Theorem B. For every context-free language L, there exist a regular set R and homomorphisms h_1 and h_2, with h_1 length-preserving, such that $L = h_1(h_2^{-1}(D_2) \cap R)$, where D_2 is the Dyck set on two letters.

The purpose of this note is to provide a sketch of a proof of Theorem B using only the basic machinery of the theory of context-free languages. Before doing this we review some concepts and notation used in the proof.

For any $n \geq 1$, let Δ_n be a set of 2n distinct symbols, $\Delta_n = \{a_1,\ldots,a_n, \bar{a}_1,\ldots,\bar{a}_n\}$. The Dyck set D_n on n letters is the language L(G) where $G = (\Delta_n \cup \{S\}, \Delta_n, P, S)$ is the context-free grammar with the set of rewriting rules $P = \{S \rightarrow SS, S \rightarrow e\} \cup \{S \rightarrow a_i S \bar{a}_i \mid 1 \leq i \leq n\}$. Alternatively, let be the congruence on $\Delta_n{}^*$ determined by defining $a_1\bar{a}_i \sim e$ for each $i = 1,\ldots,n$. Then $D_n = \{w \in \Delta_n{}^* \mid w \sim e\}$.[1] For any $n \geq 1$, any two Dyck sets on n letters are isomorphic (as semigroups of free semigroups), so that one refers to the Dyck set on n letters. Intuitively, D_n is the set of all "balanced nested" strings of matching "parentheses" in $\Delta_n{}^*$. For any n, the congruence $\sim$ on Δ^* which determines D_n has the property that for every $w \in \Delta^*$, there is a unique minimum length string $\mu(w) \in \Delta^*$ such that $w \sim \mu(w)$, i.e., $w \sim \mu(w)$ and if $w \sim y$ and $y \neq \mu(w)$, then $|y| > |\mu(w)|$.[2] The function μ has the following properties:

i) $\mu(w) = e$ if and only if and only $w \in D_n$;

ii) for any $x,y \in \Delta^*$, $\mu(xy) = \mu(\mu(x)y)$;

iii) for any $x \in \Delta^*$ and any $y \in \{a_1,\ldots,a_n\}^*$, $\mu(xy) = \mu(x)y$.

For any $n \geq 1$, consider the homomorphism $h\colon \Delta_n{}^* \rightarrow \Delta_2{}^*$ determined by defining $h(a_i) = a_1{}^i a_2$ and $h(\bar{a}_i) = \bar{a}_2\bar{a}_1{}^i$ for each $i = 1,\ldots,n$. Now h is one-to-one but is not onto. It is easy to see that $h^{-1}(D_2) = \{w \in \Delta_n{}^* \mid h(w) \in D_2\} = D_n$. Thus, every Dyck set can be obtained from the Dyck set on two letters by applying an inverse homomorphism.

---

1. If Σ is a finite set of symbols, then Σ^* is the free monoid with identity e generated by Σ.

2. For any string x, the length of x is denoted by $|x|$.

Let $h: \Sigma^* \to \Delta^*$ be a homomorphism and let $L \subseteq \Sigma^*$. Suppose that there is an integer k such that for all $x,y,z \in \Sigma^*$, if $xyz \in L$ and $h(y) = e$, then $|y| \leq k$. Then we say that <u>h is k-limited on L</u>. If there exists k such that h is k-limited on L, then <u>h is e-limited on L</u>. If for all $a \in \Sigma$, $|h(a)| = 1$, then h is a <u>length-preserving</u> homomorphism.

A context-free grammar $G = (V,\Sigma,P,S)$ is in <u>Greibach Normal Form</u> (standard 2-form) if each production in P is of the form $Z \to a$ or $Z \to aY_1$ or $Z \to aY_1Y_2$ where $a \in \Sigma$ and $Z,Y_1,Y_2 \in V - \Sigma$.[3] It is well-known [7] that for every context-free language L there is a Greibach Normal Form grammar G such that $L(G) = L - \{e\}$.

Before proving Theorem B we prove a slightly weaker result.

<u>Theorem C</u>. For every context-free language L, there exist a regular set R and homomorphisms h_1 and h_2 such that $L = h_1(h_2^{-1}(D_2) \cap R)$ and h_1 is e-limited on $h_2^{-1}(D_2) \cap R$, where D_2 is the Dyck set on two letters.

<u>Proof</u>. For a context-free language L such that $e \notin L$, we show that there is an integer t, a homomorphism h_1, and a regular set R such that $L = h_1(D_t \cap R)$, $e \notin R$, and h_1 is e-limited on $D_t \cap R$. If h_2 is any homomorphism with the property that $h_2^{-1}(D_2) = D_t$, then we have $L = h_1(h_2^{-1}(D_2) \cap R)$ and h_1 is

---

3. In a context-free grammar $G = (V,\Sigma,P,S)$, V is the finite set of symbols, $\Sigma \subset V$ is the set of terminal symbols, $S \in V - \Sigma$ is the initial symbol, and $P \subseteq (V - \Sigma) \times V^*$ is the finite set of productions. A production is written as $Z \to u$ instead of (Z,u). Define a binary relation $\Rightarrow$ on V^* by $\alpha Z\beta \Rightarrow \alpha\gamma\beta$ if $\alpha,\beta,\gamma \in V^*$, $Z \in V - \Sigma$, and $Z \to \gamma \in P$. Let $\overset{*}{\Rightarrow}$ be the transitive reflexive closure of $\Rightarrow$. The language generated by G is $L(G) = \{w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w\}$.

e-limited on $h_2^{-1}(D_2)$ R. Since $e \in D_2$, $e \in h_2^{-1}(D_2)$. Since R is regular, $R \cup \{e\}$ is regular. Since h_1 is a homomorphism, $h_1(e) = e$. Thus, if $L = h_1(h_2^{-1}(D_2 \cap R)$ and h_1 is e-limited on $h_2^{-1}(D_2) \cap R$, then $L \cup \{e\} = h_1(h_2^{-1}(D_2) \cap (R \cup \{e\}))$ and h_1 is e-limited on $h_2^{-1}(D_2) \cap (R \cup \{e\})$. This yields Theorem C.

Let L be a context-free language such that $e \notin L$, and let $G = (V, \Sigma, P, S)$ be a Greibach Normal Form grammar such that $L(G) = L$. For each symbol $Z \in V$, let $\bar{Z}$ be a new symbol. Let $\Delta = V \cup \{\bar{Z} \mid Z \in V\}$. Let p and q be two new symbols, $p, q \notin \Delta$. Let $G_0 = (\{p,q\} \cup \Delta, \Delta, P_0, p)$ be the left linear grammar obtained by defining P_0 as follows:

i) $p \to Sq$ is in P_0;

ii) for each $Z \in V - \Sigma$, $a \in \Sigma$ such that $Z \to a$ is in P, $q \to a\bar{a}\bar{Z}q$ is in P_0;

iii) for each $Z, Y \in V - \Sigma$, $a \in \Sigma$ such that $Z \to aY$ is in P, $q \to a\bar{a}\bar{Z}Yq$ is in P_0;

iv) for each $Z, Y_1, Y_2 \in V - \Sigma$, $a \in \Sigma$ such that $Z \to aY_1Y_2$ is in P, $q \to a\bar{a}\bar{Z}Y_2Y_1q$ is in P_0;

v) $q \to e$ is in P_0.

Let R be the regular set $L(G_0)$. Let $\mu: \Delta^* \to \Delta^*$ be the function which assigns to each $w \in \Delta^*$, the unique minimum length string $\mu(w)$ obtained by applying the congruence on Δ^* determined by defining $a\bar{a} \sim Z\bar{Z} \sim e$ for each $a \in \Sigma$, $Z \in V - \Sigma$, i.e., $w \sim \mu(w)$ and if $w \sim y$ and $y \neq \mu(w)$, then $|y| > |\mu(w)|$.

Let t be one-half the number of symbols in Δ. We claim that $D_t \cap R$ is a set of "histories" of left-to-right derivations of strings in $L(G) = L$. Further, if $h_1: \Delta^* \to \Sigma^*$ is the homomorphism determined by defining $h_1(a) = a$ and $h_1(\bar{a}) = h_1(Z) = h_1(\bar{Z}) = e$ for $a \in \Sigma$, $Z \in V - \Sigma$, then we claim that

$h_1(D_t \cap R) = L$ and h_1 is k-limited on $D_t \cap R$ for $k = 4$.

By construction of G_0, it is immediate that h_1 is 4-limited on $L(G_0) = R$ and therefore on $D_t \cap R$.

Since G is a Greibach Normal Form grammar, for every $n \geq 1$, $a_1, \ldots, a_n \in \Sigma$, and $v \in (V - \Sigma)^*$, $S \overset{*}{\Rightarrow} a_1 \ldots a_n v$ in G if and only if there is a left-to-right derivation $S \overset{*}{\Rightarrow} a_1 \ldots a_n v$ with n steps in G.[4] Thus, to show that $h_1(D_t \cap R) = L$, it is sufficient to establish the following technical result.

Claim. For each $n \geq 1$, $a_1, \ldots, a_n \in \Sigma$, $v \in (V - \Sigma)^*$, there is a left-to-right derivation $S \overset{*}{\Rightarrow} a_n \ldots a_n v$ in G if and only if there exists $w \in \Delta^*$ such that $\mu(w) = v^R$, $h_1(w) = a_1 \ldots a_n$, and there is a derivation $p \overset{*}{\Rightarrow} wq$ with n+1 steps in G_0.

The proof of the claim is by induction on n and depends on the construction of G_0. We shall sketch the proof of the induction step and leave the details to the reader. Assume the result for some $n \geq 1$.

Suppose that for some $a_1, \ldots, a_{n+1} \in \Sigma$, $v \in (V - \Sigma)^*$, there is a left-to-right derivation $S \overset{*}{\Rightarrow} a_1 \ldots a_{n+1} v$ in G. Thus, for some $Z \in V - \Sigma$, $u \in (V - \Sigma)^*$, there is a left-to-right derivation $S \overset{*}{\Rightarrow} a_1 \ldots a_n Zu$ in G and there is a production $Z \to a_{n+1} x$ in P where $x \in (V - \Sigma)^*$ and $xu = v$. By the induction hypothesis, there exists $w_1 \in \Delta^*$ such that $\mu(w_1) = (Zu)^R = u^R Z$, $h_1(w_1) = a_1 \ldots a_n$, and there is a derivation $p \overset{*}{\Rightarrow} w_1 q$ with n+1 steps in G_0.

4. A derivation is left-to-right if in each step the leftmost nonterminal symbol is rewritten.

Since $\mu(w_1) = u^R Z$, $\mu(u^R Z) = u^R Z$. Since $Z \in V - \Sigma$, $\mu(u^R Z) = \mu(u^R)Z$. Thus, $\mu(u^R) = u^R$.

There are three possibilities for the form of the production $Z \to a_{n+1}x$:

$x = e$ so that $Z \to a_{n+1}$ is in P, $q \to a_{n+1}\bar{a}_{n+1}\bar{Z}q$ is in P_0, and $v = u$;

$x = Y$ for some $Y \in V - \Sigma$ so that $Z \to a_{n+1}Y$ is in P, and $q \to a_{n+1}\bar{a}_{n+1}\bar{Z}Yq$ is in P_0, and $v = Yu$;

$x = Y_1Y_2$ for some $Y_1, Y_2 \in V - \Sigma$ so that $Z \to a_{n+1}Y_1Y_2$ is in P, $q \to a_{n+1}\bar{a}_{n+1}\bar{Z}Y_2Y_1q$ is in P_0, and $v = Y_1Y_2u$.

In each case, the string $w = w_1a_{n+1}\bar{a}_{n+1}\bar{Z}x^R$ is the required string in Δ^*. To see this, note that $x^R \in (V - \Sigma)^*$ so that $\mu(w) = \mu(w_1a_{n+1}\bar{a}_{n+1}\bar{Z})x^R$, and that $\mu(w_1a_{n+1}\bar{a}_{n+1}\bar{Z}) = \mu(w_1\bar{Z}) = \mu(u^R Z\bar{Z}) = \mu(u^R) = u^R$, so that $\mu(w) = u^R x^R = (xu)^R = v^R$. Also, $h_1(w) = h_1(w_1)h_1(a_{n+1})h_1(\bar{a}_{n+1})h_1(\bar{Z})h_1(x^R) = a_1 \ldots a_n a_{n+1}$. Finally, since there is a derivation $p \overset{*}{\Rightarrow} w_1q$ with n+1 steps in G_0 and $q \to a_{n+1}\bar{a}_{n+1}\bar{Z}x^R q$ is in P_0, there is a derivation $p \overset{*}{\Rightarrow} w_1a_{n+1}\bar{a}_{n+1}\bar{Z}x^R q$ with n+2 steps in G_0.

Conversely, suppose that there exists $w \in \Delta^*$ such that there is a derivation $p \overset{*}{\Rightarrow} wq$ with n+2 steps in G_0. From the construction of G_0, we see that $h_1(w) = a_1 \ldots a_{n+1}$ for some $a_1, \ldots, a_{n+1} \in \Sigma$, and that $\mu(w) \in (V - \Sigma)^*$. Let $v = (\mu(w))^R$. Since G_0 is a left linear grammar, every derivation from p is a left-to-right derivation. Thus, there exists a unique pair $y, z \in \Delta^*$ such that $yz = w$, there is a derivation $p \overset{*}{\Rightarrow} yq$ of length n+1 in G_0, and $q \to zq$ is in P_0. Applying the induction hypothesis to y and considering the three possible forms for z yields the conclusion that there is a left-to-right derivation $S \overset{*}{=>} a_1 \ldots a_n a_{n+1} v$ in G.

This completes our proof of the claim.

To see that $L = h_1(D_t \cap R)$, note that for any $n \geq 1$ and $a_1, \ldots, a_n \in \Sigma$, $a_1 \ldots a_n \in L = L(G)$ if and only if there is a left-to-right derivation $S \overset{*}{=>} a_1 \ldots a_n$ in G. By the Lemma, $S \overset{*}{=>} a_1 \ldots a_n$ in G if and only if there exists $w \in \Delta^*$ such that $\mu(w) = e$, $h_1(w) = a_1 \ldots a_n$, and there is a derivation $p \overset{*}{=>} wq$ with n+1 steps in G_0. Now $p \overset{*}{=>} wq$ in G_0 implies that $p \overset{*}{=>} wq => w$ since $q \to e$ is in P_0, so that $w \in L(G_0) = R$. Since $\mu(w) = e$, $w \in D_t$. Thus, $a_1 \ldots a_n \in L$ if and only if $a_1 \ldots a_n \in h_1(D_t \cap R)$. From the remarks above, this yields Theorem C. □

We now prove Theorem B from Theorem C. Suppose L is a context-free language and $L - \{e\}$ is generated by a grammar $G = (V, \Sigma, P, S)$ in Greibach Normal Form. Let $\Delta = V \cup \{\overline{Z}: Z \in V\}$ and suppose the homomorphisms $h_1: \Delta^* \to \Sigma^*$ and $h_2: \Delta^* \to \Delta_2^*$ and the regular set $R \subseteq \Delta^*$ are as defined in the proof of Theorem C, so that $L - \{e\} = h_1(h_2^{-1}(D_2) \cap R)$. We use a technique of Ginsburg, Greibach, and Hopcroft's [5] to construct a length-preserving homomorphism h_3, a homomorphism h_4, and a regular set R' such that $L - \{e\} = h_3(h_4^{-1}(D_2) \cap R)$.

Let Γ be an alphabet consisting of symbols [yay'] with $a \in \Sigma$, $y, y' \in \Delta^*$, $h_1(y) = h_1(y') = e$, and $0 \leq |y|, |y'| \leq 4$. (Recall that h_1 is 4-limited on $h_2^{-1}(D_2) \cap R$.) Let $R' \subseteq \Gamma^*$ be the regular set $R' = \{[w_1] \ldots [w_n] \mid n \geq 1, w_1, \ldots, w_n \in R\}$. Let $h_3: \Gamma^* \to \Sigma^*$ and $h_4: \Gamma^* \to \Delta_2^*$ be the homomorphisms determined by defining $h_3([yay']) = a$ for $a \in \Sigma$ and $h_4([yay']) = h_2(yay')$. Note that h_3 is a length-preserving homomorphism and

$h_3([w]) = h_1(w)$ for $[w] \in \Gamma$. It is easily verified that
$h_3(h_4^{-1}(D_2) \cap R') = h_1(h_2^{-1}(D_2) \cap R) = L - \{e\}$. Also,
$L \cup \{e\} = h_3(h_4^{-1}(D_2) \cap (R' \cup \{e\}))$. This yields Theorem B.

One should note that Theorem B is the basis for the result stated in Ginsburg and Greibach [4] that the class of context-free languages is a principal abstract family of languages with generator D_2. The use of a Greibach Normal Form grammar in the proof of Theorem C is similar to the use of such grammars in the proof of the main result of Greibach [6].

In the proofs of Theorems B and C, the construction of the homomorphisms depended on the size (number of symbols) of a Greibach Normal Form grammar for $L - \{e\}$. The proof of Theorem C can be altered so that the homomorphisms depend only on the alphabet Σ (where $L \subseteq \Sigma^*$), by using an idea in the proof of the Chomsky-Schützenberger Theorem in Ginsburg [3]. However, the limit on the erasing done by h_1 will then depend on the grammar G, rather than being fixed at 4, and the homomorphisms constructed for Theorem B depend on the amount of erasing.

References

1] N. Chomsky.
Context-free grammars and pushdown storage.
MIT Research Laboratory in Electronics Quarterly Progress Report 65, 1962.

2] N. Chomsky and M. P. Schützenberger.
The algebraic theory of context-free languages.
In P. Braffort and P. Hirschberg, editors, Computer Programming and Formal Systems, 115-161. North-Holland, Amsterdam, 1963.

3] S. Ginsburg.
The Mathematical Theory of Context-Free Languages, 109-114.
McGraw-Hill, 1966.

4] S. Ginsburg and S. Greibach.
Principal AFL.
JCSS 4:308-338, 1970.

5] S. Ginsburg, S. Greibach, and J. Hopcroft.
Studies in Abstract Families of Languages, page 7.
Memoir 87, American Mathematical Society, 1969.

6] S. Greibach.
The hardest context-free language.
SIAM Journal of Computing 2:304-310, 1973.

7] S. Greibach.
A new normal form theorem for context-free phrase structure grammars.
JACM 12:42-52, 1965.

8] A. Salomaa.
Formal Languages, 68-71.
Academic Press, 1973.