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 $\Delta_n$ 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 $P = \{S \to SS, S \to e\} \cup \{S \to a_i \bar{a}_i \mid 1 \leq i \leq n\}$. Alternatively, let be the congruence on $\Delta_n^*$ determined by defining $a_i \bar{a}_i \sim e$ for each $i = 1, \ldots, n$. Then $D_n = \{w \in \Delta_n^* \mid w \sim e\}$. 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 $\Delta^*$ 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 $|y| > |\mu(w)|$. The function $\mu$ 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 $h: \Delta_n^* \to \Delta_2^*$ 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 $h^{-1}(D_n) = \{w \in \Delta_n^* \mid h(w) \in D_2\} = D_n$. Thus, every Dyck set can be obtained from the Dyck set on two letters by applying an inverse homomorphism.

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

2. For any string $x$, the length of $x$ is denoted by $|x|$. 
Let $h: \Sigma^* \rightarrow \Delta^*$ 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 $h$ is $k$-limited on $L$. If there exists $k$ such that $h$ is $k$-limited on $L$, then $h$ is $e$-limited on $L$. If for all $a \in \Sigma$, $|h(a)| = 1$, then $h$ is a length-preserving homomorphism.

A context-free grammar $G = (V, \Sigma, P, S)$ is in Greibach Normal Form (standard 2-form) if each production in $P$ is of the form $Z \rightarrow a$ or $Z \rightarrow aY_1$ or $Z \rightarrow aY_1Y_2$ where $a \in \Sigma$ and $Z, Y_1, Y_2 \in V - \Sigma$. 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.

**Theorem C.** 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.

**Proof.** 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, $\Sigma \subseteq V$ 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 $Z \rightarrow u$ 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 $Z \rightarrow \gamma \in P$. Let $\Rightarrow^*$ be the transitive reflexive closure of $\Rightarrow$. The language generated by $G$ is $L(G) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}$. 

e-limited on $h_2^{-1}(D_2 \cap R)$. 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

$L = h_1^{-1}(h_2^{-1}(D_2 \cap R))$ and $h_1$ is e-limited on $h_2^{-1}(D_2 \cap R)$, then

$L \cup \{e\} = h_1^{-1}(h_2^{-1}(D_2) \cap (R \cup \{e\}))$ 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 $\overline{Z}$ be a new symbol. Let $\Delta = V \cup \{\overline{Z} \mid Z \in V\}$. 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 $P$, $q \to a\overline{a}Zq$ is in $P_0$;

iii) for each $Z, Y \in V - \Sigma$, $a \in \Sigma$ such that $Z \to aY$ is in $P$, $q \to a\overline{a}ZYq$ 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 $P$, $q \to a\overline{a}ZY_2Y_1q$ 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 $\Delta^*$ determined by defining $a\overline{a} \sim ZZ \sim e$ 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 $\Delta$. We claim that $D_1 \cap R$ 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 $h_1(\overline{a}) = h_1(Z) = h_1(\overline{Z}) = e$ for $a \in \Sigma$, $Z \in V - \Sigma$, then we claim that
h₁(Dₜ∩R) = L and h₁ is k-limited on Dₜ∩R for k = 4.

By construction of G₀, it is immediate that h₁ is 4-limited on L(G₀) = R and therefore on Dₜ∩R.

Since G is a Greibach Normal Form grammar, for every n ≥ 1, a₁,...,aₙ ∈ Σ, and v ∈ (V₋Σ)*, S *⇒ a₁...aₙv in G if and only if there is a left-to-right derivation S *⇒ a₁...aₙv with n steps in G.⁴ Thus, to show that h₁(Dₜ∩R) = L, it is sufficient to establish the following technical result.

**Claim.** For each n ≥ 1, a₁,...,aₙ ∈ Σ, v ∈ (V₋Σ)*, there is a left-to-right derivation S *⇒ aₙ...a₁v in G if and only if there exists w ∈ Δ* such that μ(w) = v R, h₁(w) = a₁...aₙ, and there is a derivation p *⇒ wq with n+1 steps in G₀.

The proof of the claim is by induction on n and depends on the construction of G₀. We shall sketch the proof of the induction step and leave the details to the reader. Assume the result for some n ≥ 1.

Suppose that for some a₁,...,aₙ₊₁ ∈ Σ, v ∈ (V₋Σ)*, there is a left-to-right derivation S *⇒ a₁...aₙ₊₁v in G. Thus, for some Z ∈ V₋Σ, u ∈ (V₋Σ)*, there is a left-to-right derivation S *⇒ a₁...aₙZu in G and there is a production Z → aₙ₊₁X in P where X ∈ (V₋Σ)* and Xu = v. By the induction hypothesis, there exists w₁ ∈ Δ* such that μ(w₁) = (Zu) R = u RZ, h₁(w₁) = a₁...aₙ, and there is a derivation p *⇒ w₁q with n+1 steps in G₀.

---

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 \( z \Rightarrow a_{n+1} x \):

- \( x = e \) so that \( z \Rightarrow a_{n+1} \) is in \( P \), \( q \Rightarrow a_{n+1} a_{n+1} \) is in \( P_0 \), and 
  \( \nu = \epsilon u \);
- \( x = Y \) for some \( y \in V - \Sigma \) so that \( z \Rightarrow a_{n+1} y \) is in \( P \), and 
  \( q \Rightarrow a_{n+1} a_{n+1} \) \( \Rightarrow y_{n+1} q \) is in \( P_0 \), and \( \nu = yu \);
- \( x = X \) for some \( y, z \in V - \Sigma \) so that \( z \Rightarrow a_{n+1} y_1 y_2 \) is in \( P \), and 
  \( q \Rightarrow a_{n+1} a_{n+1} \) \( \Rightarrow y_{n+1} y_{n+1} q \) is in \( P_0 \), and \( \nu = y_1 y_2 u \).

In each case, the string \( w = w_1 a_{n+1} a_{n+1} \) is the required string in \( \Delta^* \). To see this, note that \( x^R \in (V - \Sigma)^* \) so that \( \mu(w) = \mu(w_1 a_{n+1} a_{n+1} z) x^R \), and that 
\[
\mu(w_1 a_{n+1} a_{n+1} z) = \mu(w_1 z) = \mu(u^R z) = u^R, \quad \text{so that} \\
\mu(w) = u^R x^R = (xu)^R = \nu^R. \]

Also,
\[
h_1(w) = h_1(w_1) h_1(a_{n+1}) h_1(a_{n+1}) h_1(z) h_1(x^R) = a_n \ldots a_{n+1}. \]

Finally, since there is a derivation \( p \Rightarrow w_1 q \) with \( n+1 \) steps in \( G_0 \) and \( q \Rightarrow a_{n+1} a_{n+1} \) \( \Rightarrow x^R \) \( q \) is in \( P_0 \), there is a derivation \( p \Rightarrow w_1 a_{n+1} a_{n+1} \) \( \Rightarrow x^R \) \( q \) with \( n+2 \) steps in \( G_0 \).

Conversely, suppose that there exists \( w \in \Delta^* \) such that there is a derivation \( p \Rightarrow wq \) with \( n+2 \) steps in \( G_0 \). From the construction of \( G_0 \), we see that \( h_1(w) = a_n \ldots a_{n+1} \) for some \( a_n \ldots, a_{n+1} \in \Sigma \), and that \( \mu(w) \in (V - \Sigma)^* \).

Let \( \nu = (\mu(w))^R \). 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 \( p \Rightarrow yq \) of length \( n+1 \) in \( G_0 \), and \( q \Rightarrow zq \) 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 \( S \Rightarrow a_1 \ldots a_n a_{n+1} v \) 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
\( S \Rightarrow a_1 \ldots a_n \) in \( G \). By the Lemma, \( S \Rightarrow a_1 \ldots a_n \) 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
\( p \Rightarrow wq \) with \( n+1 \) steps in \( G_0 \). Now \( p \Rightarrow wq \) in \( G_0 \) implies that \( p \Rightarrow wq \Rightarrow w \)
since \( q \rightarrow e \) 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. \( \square \)

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 \( h_1 : \Delta^* \rightarrow \Sigma^* \)
and \( h_2 : \Delta^* \rightarrow \Delta_2^* \) 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 \( \Gamma \) 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
\( R' = \{ [w_1] \ldots [w_n] : n \geq 1, w_1, \ldots, w_n \in R \} \). Let \( h_3 : \Gamma^* \rightarrow \Sigma^* \) and \( h_4 : \Gamma^* \rightarrow \Delta_2^* \)
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
\[ h_3([w]) = h_1(w) \text{ for } [w] \in \Gamma. \] It is easily verified that
\[ h_3(h_4^{-1}(D_2) \cap R') = h_1(h_2^{-1}(D_2) \cap R) = L \setminus \{e\}. \] 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 \(L \setminus \{e\}\). The proof of Theorem C can be altered so that the homomorphisms depend only on the alphabet \(\Sigma\) (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.

   The algebraic theory of context-free languages.
   In P. Braffort and P. Hirschberg, editors, Computer Programming and Formal


   Principal AFL.


   The hardest context-free language.

   A new normal form theorem for context-free phrase structure grammars.

8] A. Salomaa.
   Formal Languages, 68-71.