The productions for the grammar in CNF are shown below.

$$S \to C_b A \mid C_a B \qquad D_1 \to AA$$
$$A \to C_a S \mid C_b D_1 \mid a \qquad D_2 \to BB$$
$$B \to C_b S \mid C_a D_2 \mid b \qquad C_a \to a$$
$$C_b \to b$$

4.6 GREIBACH NORMAL FORM

We now develop a normal-form theorem that uses productions whose right-hand sides each start with a terminal symbol perhaps followed by some variables. First we prove two lemmas that say we can modify the productions of a CFG in certain ways without affecting the language generated.

Lemma 4.3 Define an *A-production* to be a production with variable A on the left. Let $G = (V, T, P, S)$ be a CFG. Let $A \to \alpha_1 B \alpha_2$ be a production in P and $B \to \beta_1 \mid \beta_2 \mid \cdots \mid \beta_r$ be the set of all B-productions. Let $G_1 = (V, T, P_1, S)$ be obtained from G by deleting the production $A \to \alpha_1 B \alpha_2$ from P and adding the productions $A \to \alpha_1 \beta_1 \alpha_2 \mid \alpha_1 \beta_2 \alpha_2 \mid \cdots \mid \alpha_1 \beta_r \alpha_2$. Then $L(G) = L(G_1)$.

Proof Obviously $L(G_1) \subseteq L(G)$, since if $A \to \alpha_1 \beta_i \alpha_2$ is used in a derivation of G_1, then $A \underset{G}{\Rightarrow} \alpha_1 B \alpha_2 \underset{G}{\Rightarrow} \alpha_1 \beta_i \alpha_2$ can be used in G. To show that $L(G) \subseteq L(G_1)$, one simply notes that $A \to \alpha_1 B \alpha_2$ is the only production in G not in G_1. Whenever $A \to \alpha_1 B \alpha_2$ is used in a derivation by G, the variable B must be rewritten at some later step using a production of the form $B \to \beta_i$. These two steps can be replaced by the single step $A \underset{G_1}{\Rightarrow} \alpha_1 \beta_i \alpha_2$. □

Lemma 4.4 Let $G = (V, T, P, S)$ be a CFG. Let $A \to A\alpha_1 \mid A\alpha_2 \mid \cdots \mid A\alpha_r$ be the set of A-productions for which A is the leftmost symbol of the right-hand side. Let $A \to \beta_1 \mid \beta_2 \mid \cdots \mid \beta_s$ be the remaining A-productions. Let $G_1 = (V \cup \{B\}, T, P_1, S)$ be the CFG formed by adding the variable B to V and replacing all the A-productions by the productions:

$$1) \quad \left.\begin{matrix} A \to \beta_i \\ A \to \beta_i B \end{matrix}\right\} 1 \le i \le s, \qquad 2) \quad \left.\begin{matrix} B \to \alpha_i \\ B \to \alpha_i B \end{matrix}\right\} 1 \le i \le r.$$

Then $L(G_1) = L(G)$.

Proof In a leftmost derivation, a sequence of productions of the form $A \to A\alpha_i$ must eventually end with a production $A \to \beta_j$. The sequence of replacements

$$A \Rightarrow A\alpha_{i_1} \Rightarrow A\alpha_{i_2}\alpha_{i_1} \Rightarrow \cdots \Rightarrow A\alpha_{i_p}\alpha_{i_{p-1}} \cdots \alpha_{i_1}$$
$$\Rightarrow \beta_j \alpha_{i_p} \alpha_{i_{p-1}} \cdots \alpha_{i_1}$$

in G can be replaced in G_1 by

$$\begin{aligned} A &\Rightarrow \beta_j B \Rightarrow \beta_j \alpha_{i_p} B \Rightarrow \beta_j \alpha_{i_p} \alpha_{i_{p-1}} B \\ &\Rightarrow \cdots \Rightarrow \beta_j \alpha_{i_p} \alpha_{i_{p-1}} \cdots \alpha_{i_2} B \\ &\Rightarrow \beta_j \alpha_{i_p} \alpha_{i_{p-1}} \cdots \alpha_{i_1}. \end{aligned}$$

The reverse transformation can also be made. Thus $L(G) = L(G_1)$. Figure 4.8 shows this transformation on derivation trees, where we see that in G, a chain of A's extending to the left is replaced in G_1 by a chain of B's extending to the right. □

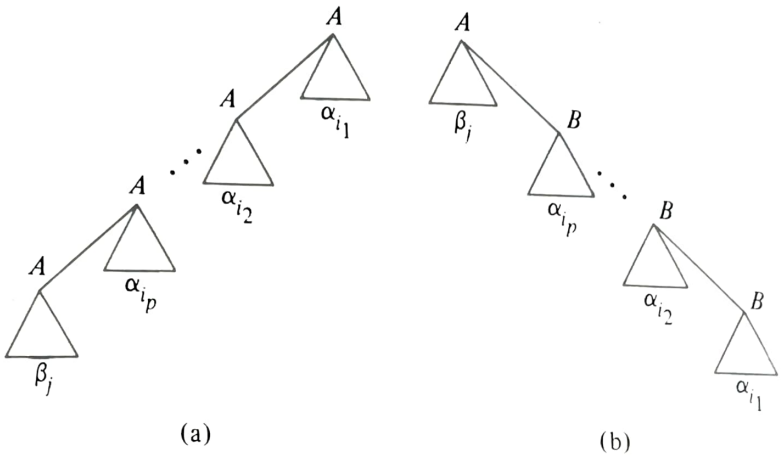

Fig. 4.8 Transformation of Lemma 4.4 on portion of a derivation tree.

Theorem 4.6 (*Greibach normal form or GNF*) Every context-free language L without ϵ can be generated by a grammar for which every production is of the form $A \to a\alpha$, where A is a variable, a is a terminal, and α is a (possibly empty) string of variables.

Proof Let $G = (V, T, P, S)$ be a Chomsky normal form grammar generating the CFL L. Assume that $V = \{A_1, A_2, \ldots, A_m\}$. The first step in the construction is to modify the productions so that if $A_i \to A_j\gamma$ is a production, then $j > i$. Starting with A_1 and proceeding to A_m, we do this as follows. We assume that the productions have been modified so that for $1 \leq i < k$, $A_i \to A_j\gamma$ is a production only if $j > i$. We now modify the A_k-productions.

If $A_k \to A_j\gamma$ is a production with $j < k$, we generate a new set of productions by substituting for A_j the right-hand side of each A_j-production according to Lemma 4.3. By repeating the process $k - 1$ times at most, we obtain productions of the form $A_k \to A_\ell\gamma$, $\ell \geq k$. The productions with $\ell = k$ are then replaced according to Lemma 4.4, introducing a new variable B_k. The precise algorithm is given in Fig. 4.9.

```
    begin
1)    for k:= 1 to m do
        begin
2)        for j:= 1 to k - 1 do
3)            for each production of the form A_k → A_j α do
                begin
4)                for all productions A_j → β do
5)                    add production A_k → βα;
6)                remove production A_k → A_j α
                end;
7)        for each production of the form A_k → A_k α do
            begin
8)                add productions B_k → α and B_k → αB_k;
9)                remove production A_k → A_k α
            end;
10)       for each production A_k → β, where β does not
            begin with A_k do
11)               add production A_k → βB_k
        end
    end
```

Fig. 4.9 Step 1 in the Greibach normal-form algorithm.

By repeating the above process for each original variable, we have only productions of the forms:

1) $A_i \to A_j\gamma$, $\quad j > i$,
2) $A_i \to a\gamma$, $\quad a$ in T,
3) $B_i \to \gamma$, $\quad \gamma$ in $(V \cup \{B_1, B_2, \ldots, B_{i-1}\})^*$.

Note that the leftmost symbol on the right-hand side of any production for A_m must be a terminal, since A_m is the highest-numbered variable. The leftmost symbol on the right-hand side of any production for A_{m-1} must be either A_m or a terminal symbol. When it is A_m, we can generate new productions by replacing A_m by the right-hand side of the productions for A_m according to Lemma 4.3. These productions must have right sides that start with a terminal symbol. We then proceed to the productions for $A_{m-2}, \ldots, A_2, A_1$ until the right side of each production for an A_i starts with a terminal symbol.

As the last step we examine the productions for the new variables, $B_1, B_2, \ldots, B_m$. Since we began with a grammar in Chomsky normal form, it is easy to prove by induction on the number of applications of Lemmas 4.3 and 4.4 that the right-hand side of every A_i-production, $1 \le i \le n$, begins with a terminal or $A_j A_k$ for some j and k. Thus α in line (7) of Fig. 4.9 can never be empty or begin with some

B_j, so no B_i-production can start with another B_j. Therefore all B_i-productions have right-hand sides beginning with terminals or A_i's, and one more application of Lemma 4.3 for each B_i-production completes the construction. □

Example 4.10 Let us convert to Greibach normal form the grammar

$$G = (\{A_1, A_2, A_3\}, \{a, b\}, P, A_1),$$

where P consists of the following:

$$A_1 \to A_2 A_3$$
$$A_2 \to A_3 A_1 | b$$
$$A_3 \to A_1 A_2 | a$$

Step 1 Since the right-hand side of the productions for A_1 and A_2 start with terminals or higher-numbered variables, we begin with the production $A_3 \to A_1 A_2$ and substitute the string $A_2 A_3$ for A_1. Note that $A_1 \to A_2 A_3$ is the only production with A_1 on the left.

The resulting set of productions is:

$$A_1 \to A_2 A_3$$
$$A_2 \to A_3 A_1 | b$$
$$A_3 \to A_2 A_3 A_2 | a$$

Since the right side of the production $A_3 \to A_2 A_3 A_2$ begins with a lower-numbered variable, we substitute for the first occurrence of A_2 both $A_3 A_1$ and b. Thus $A_3 \to A_2 A_3 A_2$ is replaced by $A_3 \to A_3 A_1 A_3 A_2$ and $A_3 \to b A_3 A_2$. The new set is

$$A_1 \to A_2 A_3$$
$$A_2 \to A_3 A_1 | b$$
$$A_3 \to A_3 A_1 A_3 A_2 | b A_3 A_2 | a$$

We now apply Lemma 4.4 to the productions

$$A_3 \to A_3 A_1 A_3 A_2 | b A_3 A_2 | a.$$

Symbol B_3 is introduced, and the production $A_3 \to A_3 A_1 A_3 A_2$ is replaced by $A_3 \to b A_3 A_2 B_3$, $A_3 \to a B_3$, $B_3 \to A_1 A_3 A_2$, and $B_3 \to A_1 A_3 A_2 B_3$. The resulting set is

$$A_1 \to A_2 A_3$$
$$A_2 \to A_3 A_1 | b$$
$$A_3 \to b A_3 A_2 B_3 | a B_3 | b A_3 A_2 | a$$
$$B_3 \to A_1 A_3 A_2 | A_1 A_3 A_2 B_3$$

Step 2 Now all the productions with A_3 on the left have right-hand sides that start with terminals. These are used to replace A_3 in the production $A_2 \to A_3 A_1$ and then the productions with A_2 on the left are used to replace A_2 in the production $A_1 \to A_2 A_3$. The result is the following.

$$A_3 \to bA_3A_2B_3 \qquad A_3 \to bA_3A_2$$
$$A_3 \to aB_3 \qquad A_3 \to a$$
$$A_2 \to bA_3A_2B_3A_1 \qquad A_2 \to bA_3A_2A_1$$
$$A_2 \to aB_3A_1 \qquad A_2 \to aA_1$$
$$A_2 \to b$$
$$A_1 \to bA_3A_2B_3A_1A_3 \qquad A_1 \to bA_3A_2A_1A_3$$
$$A_1 \to aB_3A_1A_3 \qquad A_1 \to aA_1A_3$$
$$A_1 \to bA_3$$
$$B_3 \to A_1A_3A_2 \qquad B_3 \to A_1A_3A_2B_3$$

Step 3 The two B_3-productions are converted to proper form, resulting in 10 more productions. That is, the productions

$$B_3 \to A_1A_3A_2 \quad \text{and} \quad B_3 \to A_1A_3A_2B_3$$

are altered by substituting the right side of each of the five productions with A_1 on the left for the first occurrences of A_1. Thus $B_3 \to A_1A_3A_2$ becomes

$$B_3 \to bA_3A_2B_3A_1A_3A_3A_2, \qquad B_3 \to aB_3A_1A_3A_3A_2.$$
$$B_3 \to bA_3A_3A_2, \qquad B_3 \to bA_3A_2A_1A_3A_3A_2, \qquad B_3 \to aA_1A_3A_3A_2.$$

The other production for B_3 is replaced similarly. The final set of productions is

$$A_3 \to bA_3A_2B_3 \qquad A_3 \to bA_3A_2$$
$$A_3 \to aB_3 \qquad A_3 \to a$$
$$A_2 \to bA_3A_2B_3A_1 \qquad A_2 \to bA_3A_2A_1$$
$$A_2 \to aB_3A_1 \qquad A_2 \to aA_1$$
$$A_2 \to b$$
$$A_1 \to bA_3A_2B_3A_1A_3 \qquad A_1 \to bA_3A_2A_1A_3$$
$$A_1 \to aB_3A_1A_3 \qquad A_1 \to aA_1A_3$$
$$A_1 \to bA_3$$
$$B_3 \to bA_3A_2B_3A_1A_3A_3A_2B_3 \qquad B_3 \to bA_3A_2B_3A_1A_3A_3A_2$$

$B_3 \to aB_3A_1A_3A_3A_2B_3$ $\quad$ $B_3 \to aB_3A_1A_3A_3A_2$

$B_3 \to bA_3A_3A_2B_3$ $\quad$ $B_3 \to bA_3A_3A_2$

$B_3 \to bA_3A_2A_1A_3A_3A_2B_3$ $\quad$ $B_3 \to bA_3A_2A_1A_3A_3A_2$

$B_3 \to aA_1A_3A_3A_2B_3$ $\quad$ $B_3 \to aA_1A_3A_3A_2$

4.7 THE EXISTENCE OF INHERENTLY AMBIGUOUS CONTEXT-FREE LANGUAGES

It is easy to exhibit ambiguous context-free grammars. For example, consider the grammar with productions $S \to A$, $S \to B$, $A \to a$, and $B \to a$. What is not so easy to do is to exhibit a context-free language for which every CFG is ambiguous. In this section we show that there are indeed inherently ambiguous CFL's. The proof is somewhat tedious, and the student may skip this section without loss of continuity. The existence of such a language is made use of only in Theorem 8.16.

We shall show that the language

$$L = \{a^n b^n c^m d^m \mid n \geq 1, m \geq 1\} \cup \{a^n b^m c^m d^n \mid n \geq 1, m \geq 1\}$$

is inherently ambiguous by showing that infinitely many strings of the form $a^n b^n c^n d^n$, $n \geq 1$, must have two distinct leftmost derivations. We proceed by first establishing two technical lemmas.

Lemma 4.5 Let (N_i, M_i), $1 \leq i \leq r$, be pairs of sets of integers. (The sets may be finite or infinite.) Let

$$S_i = \{(n, m) \mid n \text{ in } N_i, m \text{ in } M_i\}$$

and let

$$S = S_1 \cup S_2 \cup \cdots \cup S_r.$$

If each pair of integers (n, m) is in S for all n and m, where $n \neq m$, then (n, n) is in S for all but some finite set of n.

Proof Assume that for all n and m, where $n \neq m$, each (n, m) is in S, and that there are infinitely many n such that (n, n) is not in S. Let J be the set of all n such that (n, n) is not in S. We construct a sequence of sets $J_r, J_{r-1}, \ldots, J_1$ such that

$$J \supseteq J_r \supseteq J_{r-1} \supseteq \cdots \supseteq J_1.$$

Each J_i will be infinite, and for each n and m in J_i, (n, m) is not in

$$S_i \cup S_{i+1} \cup \cdots \cup S_r.$$

For n in J, either n is not in N_r or n is not in M_r; otherwise (n, n) would be in S_r and hence in S. Thus there is an infinite subset of J, call it J_r, such that either for all n in J_r, n is not in N_r, or for all n in J_r, n is not in M_r. Now for n and m in J_r, (n, m) is not in S_r.