5

Context-Free Languages

In this chapter we study context-free grammars and languages. We define derivation trees and give methods of simplifying context-free grammars. The two normal forms—Chomsky normal form and Greibach normal form—are dealt with. We conclude this chapter after proving pumping lemma and giving some decision algorithms.

5.1 CONTEXT-FREE LANGUAGES AND DERIVATION TREES

Context-free languages are applied in parser design. It is also useful for describing block structure in programming languages. It is easy to visualise derivations in context-free languages as we can represent derivations using tree structures.

We recall the definition of a context-free grammar (CFG). G is context-free if every production is of the form $A \to \alpha$, where $A \in V_N$ and $\alpha \in (V_N \cup \Sigma)^*$.

EXAMPLE 5.1 Construct a context-free grammar G generating all integers (with sign).

SOLUTION Let

$$G = (V_N, \Sigma, P, S)$$

where

$$V_N = \{S, \langle \text{sign} \rangle, \langle \text{digit} \rangle, \langle \text{Integer} \rangle\}$$

$$\Sigma = \{0, 1, 2, 3, \ldots, 9, +, -\}$$

P consists of $S \to \langle \text{sign} \rangle \langle \text{integer} \rangle$, $\langle \text{sign} \rangle \to + \mid -$,

$$\langle \text{integer} \rangle \to \langle \text{digit} \rangle \langle \text{integer} \rangle \mid \langle \text{digit} \rangle$$

$$\langle \text{digit} \rangle \to 0 \mid 1 \mid 2 \mid \ldots \mid 9$$

$L(G)$ = the set of all integers. For example, the derivation of -17 can be obtained as follows:

$$S \Rightarrow \langle \text{sign} \rangle \langle \text{integer} \rangle \Rightarrow - \langle \text{integer} \rangle$$

$$\Rightarrow - \langle \text{digit} \rangle \langle \text{integer} \rangle \Rightarrow - 1 \langle \text{integer} \rangle \Rightarrow - 1 \langle \text{digit} \rangle$$

$$\Rightarrow - 17$$

5.1.1 DERIVATION TREES

The derivation in a CFG can be represented using trees. Such trees representing derivations are called derivation trees. We give a rigorous definition of a derivation tree.

Definition 5.1 A derivation tree (also called a parse tree) for a CFG $G = (V_N, \Sigma, P, S)$ is a tree satisfying the following:

(i) Every vertex has a label which is a variable or terminal or Λ.
(ii) The root has label S.
(iii) The label of an internal vertex is a variable.
(iv) If the vertices $n_1, n_2, \ldots, n_k$ written with labels $X_1, X_2, \ldots, X_k$ are the sons of vertex n with label A, then $A \to X_1X_2 \ldots X_k$ is a production in P.
(v) A vertex n is a leaf if its label is $a \in \Sigma$ or Λ; n is the only son of its father if its label is Λ.

For example, let $G = (\{S, A\}, \{a, b\}, P, S)$, where P consists of $S \to aAS \mid a \mid SS$, $A \to SbA \mid ba$. Figure 5.1 is an example of a derivation tree.

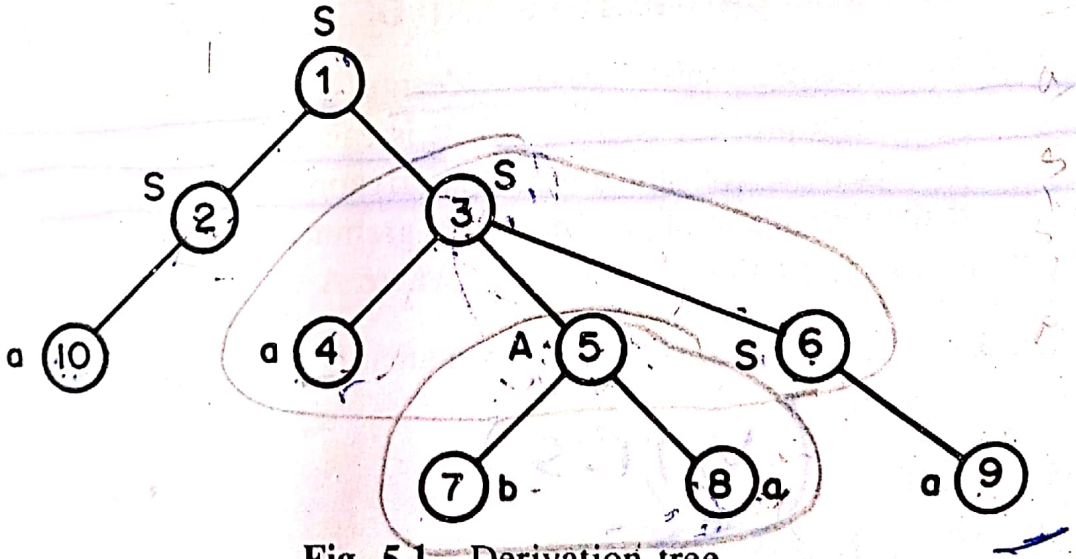

Fig. 5.1 Derivation tree.

NOTE: Vertices 4–6 are sons of 3 written from the left, and $S \to aAS$ is in P. Vertices 7 and 8 are sons of 5 written from the left, and $A \to ba$ is a production in P. Vertex 5 is an internal vertex and its label is A, which is a variable.

Ordering of Leaves from

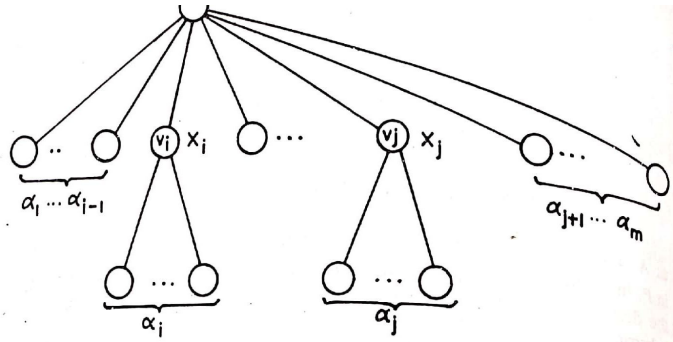

Fig. 5.7 Derivation tree with yield $\alpha_1\alpha_2 \ldots \alpha_n$.

is $A \Rightarrow A_1A_2 \ldots A_n$, then we can write w as $w_1w_2 \ldots w_n$ so that $A_i \overset{*}{\Rightarrow} w_i$. (Actually, in the derivation tree for w, the ith son of the root has label A_i, and w_i is the yield of the subtree whose root is the ith son.)

EXAMPLE 5.2 Consider G whose productions are $S \to aAS|a, A \to SbA|SS|ba$. Show that $S \overset{*}{\Rightarrow} aabbaa$ and construct a derivation tree whose yield is $aabbaa$.

SOLUTION

$$S \Rightarrow aAS \Rightarrow aSbAS \Rightarrow aabAS \Rightarrow a^2bbaS \Rightarrow a^2b^2a^2 \tag{5.2}$$

Hence, $S \overset{*}{\Rightarrow} a^2b^2a^2$. The derivation tree is given in Fig. 5.8.

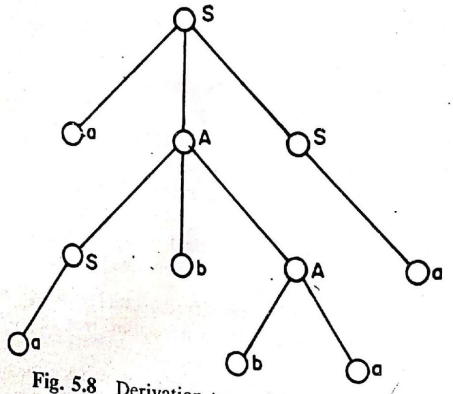

Fig. 5.8 Derivation tree with yield $aabbaa$.

NOTE: Consider G given in Example 5.2. We have seen that $S \overset{*}{\Rightarrow} a^2b^2a^2$, and (5.2) gives a derivation of $a^2b^2a^2$.

Another derivation of $a^2b^2a^2$ is

$$S \Rightarrow aAS \Rightarrow aAa \Rightarrow aSbAa \Rightarrow aSbbaa \Rightarrow aabbaa \tag{5.3}$$

Yet another derivation of $a^2b^2a^2$ is

$$S \Rightarrow aAS \Rightarrow aSbAS \Rightarrow aSbAa \Rightarrow aabAa \Rightarrow aabbaa \tag{5.4}$$

In derivation (5.2), whenever we replace a variable X using a production, there are no variables to the left of X. In derivation (5.3), there are no variables to the right of X. But in (5.4), no such conditions are satisfied. These lead to the following definitions.

Definition 5.3 A derivation $A \overset{*}{\Rightarrow} w$ is called a *leftmost* derivation if we apply a production only to the leftmost variable at every step.

Definition 5.4 A derivation $A \overset{*}{\Rightarrow} w$ is a *rightmost* derivation if we apply production to the rightmost variable at every step.

Relation (5.2), for example, is a leftmost derivation. Relation (5.3) is a rightmost derivation. But (5.4) is neither leftmost nor rightmost. In the second step of (5.4), the rightmost variable S is not replaced. So (5.4) is not a rightmost derivation. In the fourth step, the leftmost variable S is not replaced. So (5.4) is not a leftmost derivation.

Theorem 5.2 If $A \overset{*}{\Rightarrow} w$ in G, then there is a leftmost derivation of w.

PROOF We prove the result for every A in V_N by induction on the number of steps in $A \overset{*}{\Rightarrow} w$. $A \Rightarrow w$ is a leftmost derivation as L.H.S. has only one variable. So there is basis for induction. Let us assume the result for derivations in atmost k steps. Let $A \overset{k+1}{\Rightarrow} w$. The derivation can be split as $A \Rightarrow X_1X_2 \ldots X_m \overset{k}{\Rightarrow} w$.

The string w can be split as $w_1w_2 \ldots w_m$ such that $X_i \overset{*}{\Rightarrow} w_i$ (see remark before Example 5.2). As $X_i \overset{*}{\Rightarrow} w_i$ involves atmost k steps by induction hypothesis, we can find a leftmost derivation of w_i. Using these leftmost derivations, we get a leftmost derivation of w given by

$$A \Rightarrow X_1X_2 \ldots X_m \overset{*}{\Rightarrow} w_1X_2 \ldots X_m \overset{*}{\Rightarrow} w_1w_2X_3 \ldots X_m \ldots \overset{*}{\Rightarrow} w_1w_2 \ldots w_m$$

Hence by induction the result is true for all derivations $A \overset{*}{\Rightarrow} w$. ∎

Corollary Every derivation tree of w induces a leftmost derivation of w.

Once we get some derivation of w, it is easy to get a leftmost derivation of w in the following way: From the derivation tree for w, at every level consider the productions for the variables at that level, taken in the left-to-right ordering. The leftmost derivation is obtained by applying the productions in this order.

EXAMPLE 5.3 Let G be the grammar $S \to OB|1A$, $A \to 0|0S|1AA$, $B \to 1|1S|0BB$. For the string 00110101, find (a) leftmost derivation, (b) rightmost derivation, and (c) derivation tree.

SOLUTION

(a) $S \Rightarrow 0B \Rightarrow 00BB \Rightarrow 001B \Rightarrow 0011S$

$\Rightarrow 0^2 1^2 0B \Rightarrow 0^2 1^2 01S \Rightarrow 0^2 1^2 010B \Rightarrow 0^2 1^2 0101$

(b) $S \Rightarrow 0B \Rightarrow 00BB \Rightarrow 00B1S \Rightarrow 00B10B$

$\Rightarrow 0^2 B101S \Rightarrow 0^2 B1010B \Rightarrow 0^2 B10101 \Rightarrow 0^2 110101.$

The derivation tree is given in Fig. 5.9.

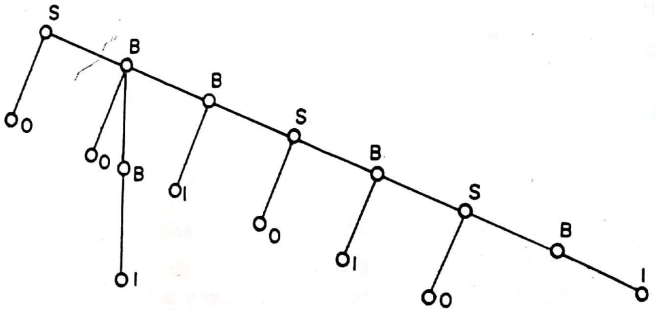

Fig. 5.9 Derivation tree with yield 00110101.

5.2 AMBIGUITY IN CONTEXT-FREE GRAMMARS

Sometimes we come across ambiguous sentences in the language we are using. Consider the following sentence in English: "In books selected information is given." The word 'selected' may refer to books or information. So the sentence may be parsed in two different ways. The same situation may arise in context-free languages. The same terminal string may be the yield of two derivation trees. So there may be two different leftmost derivations of w by Theorem 5.2. This leads to the definition of ambiguous sentences in a context-free language.

Definition 5.5 A terminal string $w \in L(G)$ is ambiguous if there exists two or more derivation trees for w (or there exist two or more leftmost derivation of w).

Consider, for example, $G = (\{S\}, \{a, b, +, *\}, P, S)$, where P consists of $S \to S + S | S * S | a | b$. We have two derivation trees for $a + a * b$ given in Fig. 5.10.

The leftmost derivations of $a + a * b$ induced by the two derivation trees are

$$S \Rightarrow S + S \Rightarrow a + S \Rightarrow a + S * S \Rightarrow a + a * S \Rightarrow a + a * b$$

$$S \Rightarrow S * S \Rightarrow S + S * S \Rightarrow a + S * S \Rightarrow a + a * S \Rightarrow a + a * b$$

Therefore, $a + a * b$ is ambiguous.

Definition 5.6 A context-free grammar G is ambiguous if there exists some $w \in L(G)$, which is ambiguous.

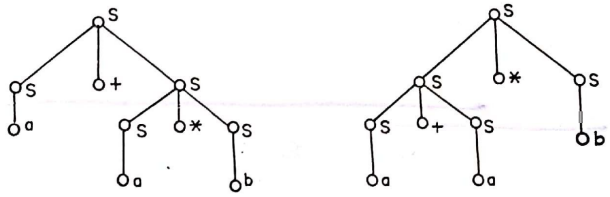

Fig. 5.10 Derivation trees for $a + a * b$.

EXAMPLE 5.4 If G is the grammar $S \to SbS | a$, show that G is ambiguous.

SOLUTION To prove that G is ambiguous, we have to find a $w \in L(G)$, which is ambiguous. Consider $w = abababa \in L(G)$. Then we get two derivation trees for w (see Fig. 5.11). Thus, G is ambiguous.

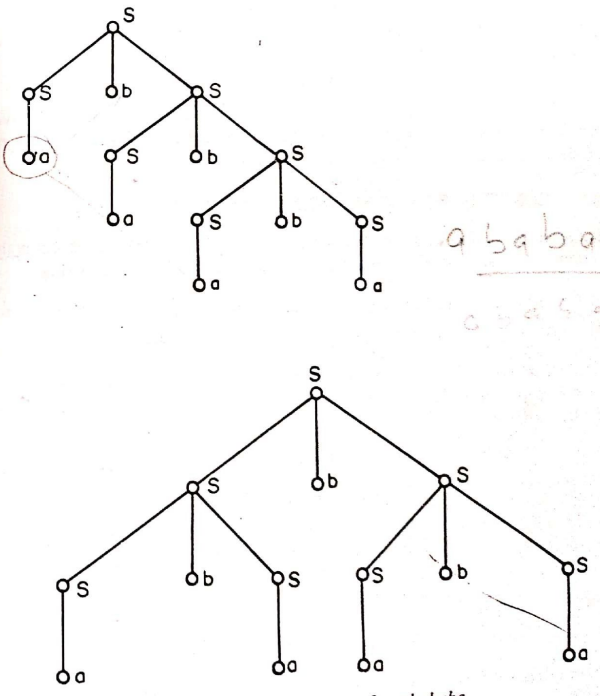

Fig. 5.11 Derivation tree for *abababa*.