Context Free Grammar and Context Free Language

Notation for CFG Derivations

There are a number of conventions in common use that help us remember the role of the symbols we use when discussing CFG's. Here are the conventions we shall use:

1. Lower-case letters near the beginning of the alphabet, a, b, and so on, are *terminal symbols*. We shall also assume that *digits* and *other characters* such as + and parentheses are terminals.
2. Upper-case letters near the beginning of the alphabet, A, B, and so on, are *variables*.
3. Lower-case letters near the end of the alphabet, such as w or z, are *strings of terminals*.
4. Upper-case letters near the end of the alphabet, such as X or Y, are either terminals or variables.
5. Lower-case Greek letters, such as α and β, are *strings* consisting of terminals and/or variables.

Leftmost and Rightmost Derivations

In order to restrict the number of choices we have in deriving a string-

- It is often useful to require that at each step we replace the leftmost variable by one of its production bodies. Such a derivation is called a *leftmost derivation*, and we indicate that a derivation is leftmost by using the relations $\Rightarrow_{lm}$ and $\Rightarrow^{*}_{lm}$, for one or more steps respectively.
- Similarly, it is also possible to require that at each step the rightmost variable is replaced by one of its bodies. If so, we call the derivation *rightmost* and use the symbols $\Rightarrow_{rm}$ and $\Rightarrow^{*}_{rm}$ to indicate one or many rightmost derivation steps, respectively.

<u>Example</u>:

1. $E \rightarrow I$
2. $E \rightarrow E+E$
3. $E \rightarrow E*E$
4. $E \rightarrow (E)$
5. $I \rightarrow a$
6. $I \rightarrow b$
7. $I \rightarrow Ia$
8. $I \rightarrow Ib$
9. $I \rightarrow I0$
10. $I \rightarrow I1$

A context-free grammar for simple expressions.

Leftmost derivation of a*(a+b00):

$E \Rightarrow_{lm} E*E$	$[E \rightarrow E*E]$
$\Rightarrow_{lm} I*E$	$[E \rightarrow I]$
$\Rightarrow_{lm} a*E$	[]
$\Rightarrow_{lm} a*(E)$	[]
$\Rightarrow_{lm} a*(E+E)$	[]
$\Rightarrow_{lm} a*(I+E)$	[]
$\Rightarrow_{lm} a*(a+E)$	[]
. . . .	
$\Rightarrow_{lm} a*(a+b00)$	

The Language of a Grammar

If G = (V, T, P, S) is a CFG, the language of G, denoted by *L(G)*, is the set of terminal strings that have derivations from the start symbol. That is,

$$L(G) = \{w \text{ in } T^* \mid S \Rightarrow^*_G w\}$$

If a language L is a language of some context free grammar, then L is said to be a *context-free language*, or *CFL*.

Sentential Forms

If G = (V, T, P, S) is a CFG, then string α in $(V \cup T)^*$ such that $S \Rightarrow^* \alpha$ is a sentential form.

- If $S \Rightarrow^*_{lm} \alpha$, then α is a *left-sentential form*.
- If $S \Rightarrow^*_{rm} \alpha$, then α is a *right-sentential form*.

Ambiguous Grammar

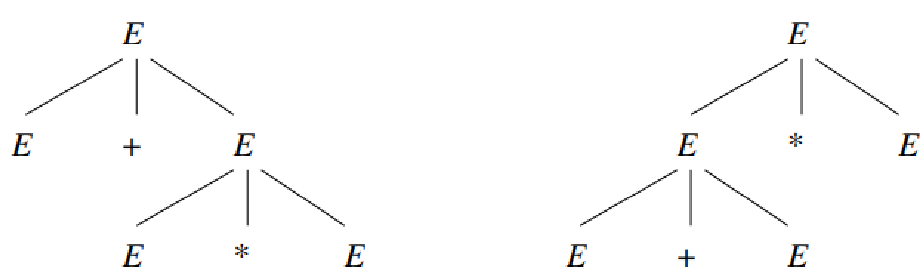
- We assume that a grammar uniquely determines a structure for each string in its language.
- However, we shall see that not every grammar does provide unique structures.
- A grammar is called *ambiguous* when it fails to provide unique structure for each string in its language.

<u>Example</u>:

Let's consider the sentential form **E + E * E**. According to the grammar mentioned above, it has two derivations from E:

1. $E \Rightarrow E + E \Rightarrow E + E * E$
2. $E \Rightarrow E * E \Rightarrow E + E * E$

Notice that in derivation (1), the second E is replaced by E * E, while in derivation (2), the first E is replaced by E + E. The following figure shows the two *parse trees* (tree representation for derivations), which we are two distinct trees.



(a) (b)