1 Normal Forms for Context-free Grammars

The goal of this section is to show that every CFL (without \( \epsilon \)) is generated by a CFG in which all rules/productions are of the form \( A \rightarrow BC \) or \( A \rightarrow a \), where \( A, B \) and \( C \) are variables, and \( a \) is a terminal. This form is called **Chomsky Normal Form**.

To get there, we need to make few simplifications on CFG, which are themselves useful in various ways:

1. Eliminate **useless symbols** - those variables or terminals that do not appear in any derivation of a terminal string from the start symbol.

2. Eliminate \( \epsilon \)-**productions** - those of the form \( A \rightarrow \epsilon \) for some variable \( A \).

3. Eliminate **unit productions** - those of the form \( A \rightarrow B \) for variables \( A, B \).

1.1 Eliminating Useless Symbols

A symbol \( X \) is said to be **useful** for grammar \( G = (V, T, R, S) \) if there is some derivation of the form \( S \xrightarrow{*} \alpha X \beta \xrightarrow{*} w \), where \( w \in T^* \). Note that \( X \) may be in either \( V \) or \( T \), and the sentential form \( \alpha X \beta \) might be the first or last in the derivation. If \( X \) is not useful, we say it is **useless**. Evidently, omitting useless symbols from a grammar will not change the language generated by grammar \( G \), so we may as well detect and eliminate all useless symbols.

**Definition:** We say \( X \) is **generating** if \( X \xrightarrow{*} w \) for some terminal string \( w \). Note that every terminal is generating, since \( w \) can be that terminal itself, which is derived by zero steps.

**Definition:** We say \( X \) is **reachable** if there is a derivation \( S \xrightarrow{*} \alpha X \beta \) for some \( \alpha \) and \( \beta \).

Surely a symbol that is useful will be both **generating** and **reachable**. If we eliminate the symbols that are not generating first, and then eliminate from the remaining grammar those symbols that are not reachable, we shall, as will be proved, have only the useful symbols left.
Example: Consider the grammar

\[
\begin{align*}
S &\rightarrow aAa|aBC \\
A &\rightarrow aS|bD \\
B &\rightarrow aBa|b \\
C &\rightarrow abb|DD \\
D &\rightarrow aDa \\
E &\rightarrow Fa \\
F &\rightarrow c
\end{align*}
\]

- \(B\) is generating because of the rules \(B \rightarrow b\)
- \(C\) is generating because of the rules \(C \rightarrow abb\)
- \(F\) is generating because of the rules \(F \rightarrow c\)
- \(S\) is generating because of the rules \(S \rightarrow aBC\)
- \(A\) is generating because of the rules \(A \rightarrow aS\)
- \(E\) is generating because of the rules \(E \rightarrow Fa\)

Thus the only symbol which is not generating is \(D\). Eliminating \(D\) and those productions involving \(D\) we get the grammar:

\[
\begin{align*}
S &\rightarrow aAa|aBC \\
A &\rightarrow aS \\
B &\rightarrow aBa|b \\
C &\rightarrow abb \\
E &\rightarrow Fa \\
F &\rightarrow c
\end{align*}
\]

Now we find that only \(S, A, B, C, a\) and \(b\) are reachable from \(S\). Eliminating \(E\) and \(F\) and all productions involving them we get

\[
\begin{align*}
S &\rightarrow aAa|aBC \\
A &\rightarrow aS \\
B &\rightarrow aBa|b \\
C &\rightarrow abb
\end{align*}
\]

Theorem: Let \(G = (V, T, R, S)\) be a CFG, and assume that \(L(G) \neq \emptyset\); i.e., \(G\) generates at least one string. Let \(G_1 = (V_1, T_1, R_1, S)\) be the grammar we obtained by the following steps:

1. First eliminate non-generating symbols and all productions involving one or more of those symbols. Let \(G_2 = (V_2, T_2, R_2, S)\) be this new grammar. Note that \(S\) must be generating, since we assume \(L(G)\) has at least one string, so \(S\) has not been eliminated.

2. Second, eliminate all symbols that are not reachable in the grammar \(G_2\).
Then (i) $G_1$ has no useless symbols, and (ii) $L(G_1) = L(G)$.

**Proof:** (i) Suppose $X$ is a symbol that remains; i.e., $X \in V_1 \cup T_1$. We know that $X \Rightarrow_G^* w$ for some $w \in T^*$. Moreover every symbol used in the derivation of $w$ from $X$ is also generating. Thus, $X \Rightarrow_{G_2}^* w$.

Since $X$ was not eliminated in the second step, we also know that there are $\alpha$ and $\beta$ such that $S \Rightarrow_{G_2}^* \alpha X \beta$. Further, every symbol used in this derivation is reachable, so $S \Rightarrow_{G_1}^* \alpha X \beta$.

We know that every symbol in $\alpha X \beta$ is reachable, and we also know that all these symbols are in $V_2 \cup T_2$, so each of them is generating in $G_2$. The derivation of some terminal string, say $\alpha X \beta \Rightarrow_{G_2}^* xwy$, involves only symbols that are reachable from $S$, because they are reached by symbols in $\alpha X \beta$. Thus, this derivation is also a derivation of $G_1$; that is,

$$S \Rightarrow_{G_1}^* \alpha X \beta \Rightarrow_{G_1}^* xwy$$

We conclude that $X$ is useful in $G_1$. Since $X$ is an arbitrary symbol of $G_1$, we conclude that $G_1$ has no useless symbols.

(ii) $L(G_1) \subseteq L(G)$: Since we have only eliminated symbols and productions from $G$ to get $G_1$, it follows that $L(G_1) \subseteq L(G)$.

$L(G) \subseteq L(G_1)$: We must prove that if $w \in L(G)$, then $w \in L(G_1)$. If $w \in L(G)$, then $S \Rightarrow_G^* w$. Each symbol in this derivation is evidently both reachable and generating, so it is also a derivation derivation of $G_1$. That is, $S \Rightarrow_{G_1}^* w$, and thus $w \in L(G_1)$.  