6.2 Two Important Normal Forms

[1] Chomsky Normal Form

Definition 6.4

A context-free grammar is in Chomsky normal form if all productions are of the form

$$A \to BC$$

or

$$A \to a,$$

where A, B, C are in V, and a is in T.

Example 6.7 The grammar

$$S \to AS|a,$$
$$A \to SA|b$$

is in Chomsky normal form. The grammar

$$S \to AS|AAS,$$
$$A \to SA|aa$$

is not; both productions $S \to AAS$ and $A \to aa$ violate the conditions of Definition 6.4.

Theorem 6.6 Any context-free grammar $G = (V, T, S, P)$ with $\lambda \notin L(G)$ has an equivalent grammar $\widehat{G} = \left(\widehat{V}, \widehat{T}, S, \widehat{P}\right)$ in Chomsky normal form.

Proof: Because of Theorem 6.5, we can assume without loss of generality that G has no λ-productions and no unit-productions. The construction of $\widehat{G}$ will be done in two steps.

Step 1: Construct a grammar $G_1 = (V_1, T, S, P_1)$ from G by considering all productions in P in the form

$$A \rightarrow x_1 x_2 \cdots x_n, \tag{6.5}$$

where each x_i is a symbol either in V or in T. If $n = 1$, then x_1 must be a terminal since we have no unit-productions. In this case, put the production into P_1. If $n \geq 2$, introduce new variables B_a for each $a \in T$. For each production of P in the form (6.5) we put into P_1 the production

$$A \to C_1 C_2 \cdots C_n,$$

where

$$C_i = x_i \text{ if } x_i \text{ is in } V,$$

and

$$C_i = B_a \text{ if } x_i = a.$$

For every B_a we also put into P_1 the production

$$B_a \to a.$$

This part of the algorithm removes all terminals from productions whose right side has length greater than one, replacing them with newly introduced variables. At the end of this step we have a grammar G_1 all of whose productions have the form

$$A \to a, \tag{6.6}$$

or

$$A \to C_1 C_2 \cdots C_n, \tag{6.7}$$

where $C_i \in V_1$.

It is an easy consequence of Theorem 6.1 that

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

Step 2: In the second step, we introduce additional variables to reduce the length of the right sides of the productions where necessary. First we put all productions of the form (6.6) as well as all the productions of the form (6.7) with $n = 2$ into $\widehat{P}$. For $n > 2$, we introduce new variables $D_1, D_2, \ldots$ and put into $\widehat{P}$ the productions

$$\begin{aligned}
A &\to C_1 D_1, \\
D_1 &\to C_2 D_2, \\
&\vdots \\
D_{n-2} &\to C_{n-1} C_n.
\end{aligned}$$

Obviously, the resulting grammar $\widehat{G}$ is in Chomsky normal form. Repeated applications of Theorem 6.1 will show that $L(G_1) = L\left(\widehat{G}\right)$, so that

$$L\left(\widehat{G}\right) = L(G).$$

This somewhat informal argument can easily be made more precise. We will leave this to the reader. ■

Example 6.8 Convert the grammar with productions

$$\begin{aligned} S &\rightarrow ABa, \\ A &\rightarrow aab, \\ B &\rightarrow Ac \end{aligned}$$

to Chomsky normal form.

As required by the construction of Theorem 6.6, the grammar does not have any λ-productions or any unit-productions.

In Step 1, we introduce new variables B_a, B_b, B_c and use the algorithm to get

$$
\begin{aligned}
S &\to ABB_a, \\
A &\to B_aB_aB_b, \\
B &\to AB_c, \\
B_a &\to a, \\
B_b &\to b, \\
B_c &\to c.
\end{aligned}
$$

In the second step, we introduce additional variables to get the first two productions into normal form and we get the final result

$$\begin{aligned}
S &\rightarrow AD_1, \\
D_1 &\rightarrow BB_a, \\
A &\rightarrow B_aD_2, \\
D_2 &\rightarrow B_aB_b, \\
B &\rightarrow AB_c, \\
B_a &\rightarrow a, \\
B_b &\rightarrow b, \\
B_c &\rightarrow c.
\end{aligned}$$

$\blacksquare$

[2] ## Greibach Normal Form

Definition 6.5

A context-free grammar is said to be in Greibach normal form if all productions have the form

$$A \to ax,$$

where $a \in T$ and $x \in V^*$.

If we compare this with Definition 5.4, we see that the form $A \to ax$ is common to both Greibach normal form and s-grammars, but Greibach normal form does not carry the restriction that the pair (A, a) occur at most once. This additional freedom gives Greibach normal form a generality not possessed by s-grammars.

If a grammar is not in Greibach normal form, we may be able to rewrite it in this form with some of the techniques encountered above. Here are two simple examples.

Example 6.9 The grammar

$$S \rightarrow AB,$$
$$A \rightarrow aA \,|bB|\, b,$$
$$B \rightarrow b$$

is not in Greibach normal form. However, using the substitution given by Theorem 6.1, we immediately get the equivalent grammar

$$S \rightarrow aAB \,|bBB|\, bB,$$
$$A \rightarrow aA \,|bB|\, b,$$
$$B \rightarrow b,$$

which is in Greibach normal form. ■

Example 6.10 Convert the grammar

$$S \to abSb|aa$$

into Greibach normal form.

Here we can use a device similar to the one introduced in the construction of Chomsky normal form. We introduce new variables A and B that are essentially synonyms for a and b, respectively. Substituting for the terminals with their associated variables leads to the equivalent grammar

$$S \to aBSB|aA,$$
$$A \to a,$$
$$B \to b,$$

which is in Greibach normal form. ■

Theorem 6.7 For every context-free grammar G with $\lambda \notin L(G)$, there exists an equivalent grammar $\widehat{G}$ in Greibach normal form.