Solution to Problem 10

**Task:** Transform the grammar consisting of two rules

\[ S \rightarrow \varepsilon; \quad S \rightarrow (S) \]

into Chomsky normal form.

**Solution.**

**Preliminary step.** First, we introduce a new starting variable \( S_0 \) and a rule \( S_0 \rightarrow S \), where \( S \) is the starting variable of the original grammar. So, the grammar takes the following form:

\[ S \rightarrow \varepsilon; \quad S \rightarrow (S); \quad S_0 \rightarrow S. \]

**Step 0.** We eliminate non-Chomsky rules with right-hand side of length 0, i.e., with right-hand side an empty string and the left-hand side is not a starting variable. In the above grammar, there is one such rule: \( S \rightarrow \varepsilon \). To eliminate this rule, for each rule that has \( S \) in the right-hand side, we add another rule in which this symbol \( S \) is deleted.

In the above grammar, there are two such rules: \( S \rightarrow (S) \) and \( S_0 \rightarrow S \).

- For the rule \( S \rightarrow (S) \), if we delete the letter \( S \) from the right-hand side, we get the rule \( S \rightarrow () \) that we add to our grammar.

- For the rule \( S_0 \rightarrow S \), if we delete the letter \( S \) from the right-hand side, we get the rule \( S_0 \rightarrow \varepsilon \) that we add to our grammar.

After we delete the rule \( S \rightarrow \varepsilon \) and add the new rules \( S \rightarrow () \) and \( S_0 \rightarrow \varepsilon \), we get the following grammar:

\[ S \rightarrow (S); \quad S_0 \rightarrow S; \quad S \rightarrow (); \quad S_0 \rightarrow \varepsilon. \]

**Step 1.** On this step, we eliminate non-Chomsky rules in which the right-hand side has length 1, i.e., in which the right-hand side is a variable. In the above grammar, there is only one such rule: \( S_0 \rightarrow S \). To eliminate this rule, for each rule \( S \rightarrow w \) that has the variable \( S \) is the left-hand side (for any right-hand side \( w \)), we add a rule \( S_0 \rightarrow w \).

1
In the current grammar, we have two rules with $S$ in the left-hand side: $S \rightarrow (S)$ and $S \rightarrow ()$. So, once we eliminate the rule $S_0 \rightarrow S$, we have to add rules $S_0 \rightarrow (S)$ and $S_0 \rightarrow ()$. As a result, we get the following grammar:

$$S \rightarrow (S); \quad S \rightarrow (); \quad S_0 \rightarrow \varepsilon; \quad S_0 \rightarrow (S); \quad S_0 \rightarrow ().$$

**Step 2.** On this step:

- For each terminal symbol $a$, we introduce an auxiliary variable $V_a$ and a rule $V_a \rightarrow a$.
- Then, in each rule in which the right-hand side has 2 or more symbols and at least one of them is a terminal symbol, we replace each terminal symbol with the corresponding variable.

In our grammar, we have two terminal symbols ( and ). So, we introduce two new variables $V_i$ and $V_j$ and two new rules $V_i \rightarrow (\text{ and } V_j \rightarrow )$.

In the rule $S \rightarrow (S)$, we replace ( with $V_i$ and ) with $V_j$, and get the new rule $S \rightarrow V_i SV_j$. We do the same replacement with all other rules in which the right-hand side has 2 or more symbols and at least one of them is a terminal symbol. As a result, we get the following grammar:

$$S \rightarrow V_i SV_j; \quad S \rightarrow V_i V_j; \quad S_0 \rightarrow \varepsilon; \quad S_0 \rightarrow V_i SV_j; \quad S_0 \rightarrow V_i V_j; \quad V_i \rightarrow (); \quad V_j \rightarrow ).$$

**Step 3.** At this step, we replace each rule of the type $V \rightarrow ABC$ with two rules: $V_{AB} \rightarrow AB$ for a new variable $V_{AB}$ and $V \rightarrow V_{AB}C$. According to this algorithm, the rule $S \rightarrow V_i SV_j$ is replaced by two rules: $V_iS \rightarrow V_i S$ and $S \rightarrow V_i SV_j$. After we perform the same replacement for all other rules that have three or more symbols in the right-hand side, we get the following grammar:

$$V_iS \rightarrow V_i S; \quad S \rightarrow V_i SV_j; \quad S \rightarrow V_i V_j; \quad S_0 \rightarrow \varepsilon; \quad S_0 \rightarrow V_i SV_j; \quad S_0 \rightarrow V_i V_j; \quad V_i \rightarrow (); \quad V_j \rightarrow ).$$

This grammar is already in Chomsky normal form, i.e., it only has three types of rules:

- rules of the type $S_0 \rightarrow \varepsilon$, where $S_0$ is the starting variable;
- rules of the type $V \rightarrow a$, where $V$ is a variable and $a$ is a terminal symbol; and
- rules of the type $V \rightarrow AB$, where $V$, $A$, and $B$ are variables.