Conversion to Chomsky Normal Form

Chomsky Normal Form: A context-free grammar $G = (V, \Sigma, R, S)$ is in Chomsky normal form if and only if all rules are of the form:

- $A \rightarrow BC$ for some $B, C \in V$ with $B, C \neq S$,
- $A \rightarrow a$ for some $a \in \Sigma$, or
- $S \rightarrow \epsilon$

Theorem: Every CFL can be generated by some grammar in Chomsky Normal Form.

Proof: Let $G = (V, \Sigma, R, S)$ be a context-free grammar generating L. We give a several step construction for converting G to a grammar G' in Chomsky Normal Form that is a little easier for hand calculation than the one in the text.

Step 1: Create a new start symbol S_0 and add the rule $S_0 \rightarrow S$.

Step 2: For each terminal symbol $a \in \Sigma$ that appears on the right side of a rule of G of size at least 2 create a new variable A, add the rule $A \rightarrow a$ and replace very occurrence of a on the right side of a rule of size at least 2 by A.

Step 3: For each rule $A \rightarrow B_1 \ldots B_k$ with $k > 2$, create new variables $T_2, \ldots T_{k-1}$ and replace the rule by rules $A \rightarrow B_1 T_2$, $T_2 \rightarrow B_2 T_3$,$\ldots T_{k-1} \rightarrow B_{k-1} B_k$. (There are separate symbols T_i for each rule converted in this way.) Now all rules have right-hand sides of length at most 2.

Step 4: Figure out the set of variables $\mathcal{E}$ that can generate the empty string ϵ. (If $A \rightarrow \epsilon$ is a rule then put A in $\mathcal{E}$. Then for every $A \in \mathcal{E}$ if $B \rightarrow w$ is a rule with $w \in \mathcal{E}^*$, also put $B \in \mathcal{E}$. Repeat this until no new variables are added to $\mathcal{E}$.)

If $S_0 \in \mathcal{E}$ then add the rule $S_0 \rightarrow \epsilon$. Remove all rules $A \rightarrow \epsilon$ for $A \neq S_0$. For every rule $A \rightarrow BC$ with $B \in \mathcal{E}$ add the rule $A \rightarrow C$. For every rule $A \rightarrow BC$ with $C \in \mathcal{E}$ add the rule $A \rightarrow B$.

Step 5: A *unit rule* is a rule of the form $A \rightarrow B$ where A and B are variables. We now only need to eliminate all unit rules. To do this we draw a directed graph of all the variables where there is an edge from A to B if $A \rightarrow B$ is a rule. For any variable A, let $\mathcal{D}(A)$ be the set of variables reachable from A in this graph. (This is just like the $\mathcal{D}(A)$ in the text except we ignore terminals.)

Call a right-hand side of a rule *interesting* if the rule is not a unit rule. To make the Chomsky normal form grammar, we define a new grammar with the same variables in which $A \rightarrow w$ if and only if w is an interesting right-hand side of some rule whose left-hand side is in $\mathcal{D}(A)$.

Clearly these rules keep the language generated the same. $\square$