Chomsky Normal Form

Every CFL has a grammar in which the rules are of a very restricted form

$$N \longrightarrow a$$
$$A \longrightarrow BC$$

the only non-terminal allowed to go ε is the start symbol.

We do not allow rules like $A \to B$ or $A \to \varepsilon$ or $A \to$ ⟨very long string⟩

(1) Get rid of rules like $X \to \alpha_1 A_1 \alpha_2 A \cdots \alpha_n A \alpha_{n+1}$ by replacing it by all possible variations with [or] more A's removed

$X \to \alpha_1 A \alpha_2 \alpha_3 \cdots \alpha_n A \alpha_{n+1}$

$X \to \alpha_1 \alpha_2 \alpha_3 A \alpha_4 \cdots \alpha_n A \alpha_{n+1}$

(2) Remove rules like $A \to B$

If we have $A \to B$ & $B \to \beta$ remove $A \to B$ & introduce $A \to \beta$; unless β is a single NT.

Introduce an order on NT's & remove in order.

(3) Get rid of rules like $X \to a A b B \cdots c C$ by introducing new NT's $N_a, N_b, \ldots$ & new rules $N_a \to a$, $N_b \to B$, $X \to N_a A N_b B \cdots N_c C$

(4) Now all rules look like $X \to A_1 A_2 \cdots A_n$ or $X \to a$

Replace $X \to A_1 A_2 \cdots A_n$ by $X \to A_1 B_1$, $B_1 \to A_2 B_2$, $B_2 \to A_3 B_3 \cdots$

If we stop here we get the language of the original G but perhaps without ε. We can add a new start symbol S' at the outset & add rules $S' \to \varepsilon$, $S' \to S$. We omit $S' \to \varepsilon$ in step 1.