

# Chomsky Normal Form

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

$$N \rightarrow a$$
$$A \rightarrow BC$$

The only non-terminal allowed to go  $\epsilon$  is the start symbol.

We do not allow rules like  $A \rightarrow B$  or  $A \rightarrow \epsilon$  or  $A \rightarrow$  (very long string)

(1) Get rid of all rules of the form  $A \rightarrow \epsilon$ . How  
replace  $X \rightarrow \alpha_1 A_1 \alpha_2 A_2 \dots \alpha_n A_n \alpha_{n+1}$  by

$$\left. \begin{array}{l} X \rightarrow \alpha_1 \alpha_2 A \alpha_3 \dots \alpha_n A \alpha_{n+1} \\ X \rightarrow \alpha_1 A \alpha_2 \alpha_3 \dots \alpha_n A \alpha_{n+1} \\ \vdots \\ X \rightarrow \alpha_1 \alpha_2 \alpha_3 A \alpha_4 \dots \alpha_n A \alpha_{n+1} \end{array} \right\} \begin{array}{l} \text{all possible} \\ \text{variations with } \underline{\text{len}} \\ \underline{\text{more } A\text{'s removed}} \end{array}$$

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

If we have  $A \rightarrow B$  &  $B \rightarrow \beta$  remove  $A \rightarrow B$  & introduce  $A \rightarrow \beta$ , unless  $\beta$  is a single NT.

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

(3) Get rid of rules like  $X \rightarrow a A B B \dots c$

by introducing new NT's  $N_a, N_b, \dots$  & new rules  $N_a \rightarrow a, N_b \rightarrow B, X \rightarrow N_a A N_b B \dots N_c C$

(4) Now all rules look like  $X \rightarrow A_1 A_2 \dots A_n$  or  $X \rightarrow a$   
Replace  $X \rightarrow A_1 A_2 \dots A_n$  by  $X \rightarrow A_1 B_1, B_1 \rightarrow A_2 B_2, B_2 \rightarrow A_3 B_3 \dots$

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