

# *Chomsky Normal Form*

*We introduce Chomsky Normal Form, which is used to answer questions about context-free languages*

## Chomsky Normal Form

**Chomsky Normal Form.** *A grammar where every production is either of the form  $A \rightarrow BC$  or  $A \rightarrow c$  (where  $A, B, C$  are arbitrary variables and  $c$  an arbitrary symbol).*

Example:

$$\begin{aligned} S &\rightarrow AS \mid a \\ A &\rightarrow SA \mid b \end{aligned}$$

(If language contains  $\varepsilon$ , then we allow  $S \rightarrow \varepsilon$  where  $S$  is start symbol, and forbid  $S$  on RHS.)

## Why Chomsky Normal Form?

The key advantage is that in Chomsky Normal Form, every derivation of a string of  $n$  letters has exactly  $2n - 1$  steps.

Thus: *one can determine if a string is in the language by exhaustive search of all derivations.*

## Conversion

The conversion to Chomsky Normal Form has four main steps:

1. Get rid of all  $\varepsilon$  productions.
2. Get rid of all productions where RHS is one variable.
3. Replace every production that is too long by shorter productions.
4. Move all terminals to productions where RHS is one terminal.

## 1) Eliminate $\epsilon$ Productions

Determine the nullable variables (those that generate  $\epsilon$ ) (algorithm given earlier).

Go through all productions, and for each, omit every possible subset of nullable variables.

For example, if  $P \rightarrow AxB$  with both  $A$  and  $B$  nullable, add productions  $P \rightarrow xB \mid Ax \mid x$ .

After this, delete all productions with empty RHS.

## 2) *Eliminate Variable Unit Productions*

A unit production is where RHS has only one symbol.

Consider production  $A \rightarrow B$ . Then for every production  $B \rightarrow \alpha$ , add the production  $A \rightarrow \alpha$ . Repeat until done (but don't re-create a unit production already deleted).

### *3) Replace Long Productions by Shorter Ones*

For example, if have production  $A \rightarrow BCD$ , then replace it with  $A \rightarrow BE$  and  $E \rightarrow CD$ .

(In theory this introduces many new variables, but one can re-use variables if careful.)

#### 4) *Move Terminals to Unit Productions*

For every terminal on the right of a non-unit production, add a substitute variable.

For example, replace production  $A \rightarrow bC$  with productions  $A \rightarrow BC$  and  $B \rightarrow b$ .



## Example

Consider the CFG:

$$S \rightarrow aXbX$$

$$X \rightarrow aY \mid bY \mid \varepsilon$$

$$Y \rightarrow X \mid c$$

The variable  $X$  is nullable; and so therefore is  $Y$ .  
After elimination of  $\varepsilon$ , we obtain:

$$S \rightarrow aXbX \mid abX \mid aXb \mid ab$$

$$X \rightarrow aY \mid bY \mid a \mid b$$

$$Y \rightarrow X \mid c$$

## Example: Step 2

After elimination of the unit production  $Y \rightarrow X$ , we obtain:

$$S \rightarrow aXbX \mid abX \mid aXb \mid ab$$

$$X \rightarrow aY \mid bY \mid a \mid b$$

$$Y \rightarrow aY \mid bY \mid a \mid b \mid c$$

## Example: Steps 3 & 4

Now, break up the RHSs of  $S$ ; and replace  $a$  by  $A$ ,  $b$  by  $B$  and  $c$  by  $C$  wherever not units:

$$S \rightarrow EF \mid AF \mid EB \mid AB$$

$$X \rightarrow AY \mid BY \mid a \mid b$$

$$Y \rightarrow AY \mid BY \mid a \mid b \mid c$$

$$E \rightarrow AX$$

$$F \rightarrow BX$$

$$A \rightarrow a$$

$$B \rightarrow b$$

$$C \rightarrow c$$

## Practice

Convert the following CFG into Chomsky Normal Form:

$$S \rightarrow AbA$$

$$A \rightarrow Aa \mid \varepsilon$$

## *Solution to Practice*

After the first step, one has:

$$\begin{aligned} S &\rightarrow AbA \mid bA \mid Ab \mid b \\ A &\rightarrow Aa \mid a \end{aligned}$$

The second step does not apply. After the third step, one has:

$$\begin{aligned} S &\rightarrow TA \mid bA \mid Ab \mid b \\ A &\rightarrow Aa \mid a \\ T &\rightarrow Ab \end{aligned}$$

## *Solution Continued*

And finally, one has:

$$S \rightarrow TA \mid BA \mid AB \mid b$$

$$A \rightarrow AC \mid a$$

$$T \rightarrow AB$$

$$B \rightarrow b$$

$$C \rightarrow a$$

## Summary

There are special forms for CFGs such as Chomsky Normal Form, where every production has the form  $A \rightarrow BC$  or  $A \rightarrow c$ . The algorithm to convert to this form involves (1) determining all nullable variables and getting rid of all  $\varepsilon$ -productions, (2) getting rid of all variable unit productions, (3) breaking up long productions, and (4) moving terminals to unit productions.