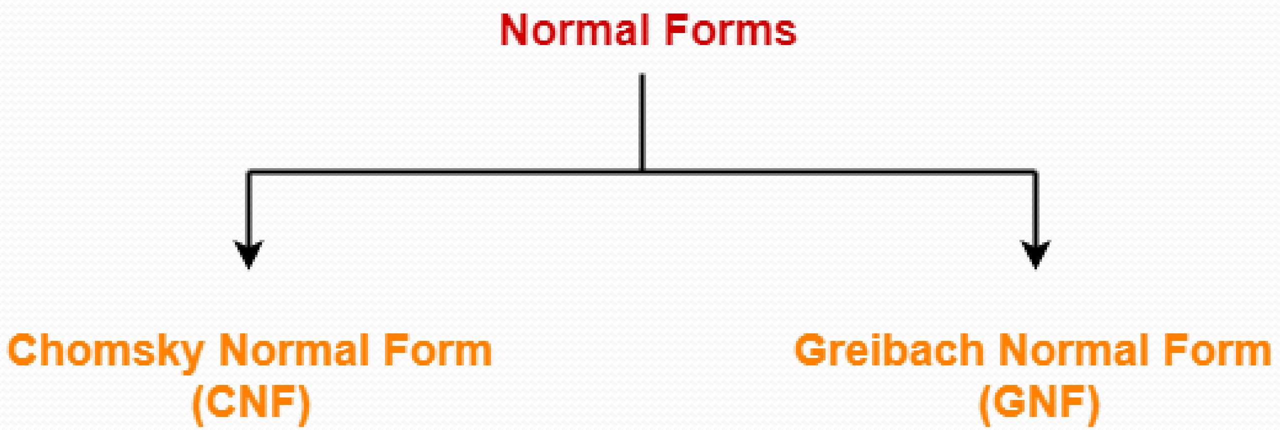
Chomsky Normal Form

---Sakshi Surve

Normal Forms:

- By reducing the grammar, the grammar gets minimized but does not get standardized.
- This is because the RHS of productions have no specific format.
- In order to standardize the grammar, normalization is performed using normal forms.


Normal Forms
Chomsky Normal Form
(CNF)
Greibach Normal Form
(GNF)

Chomsky Normal Form-

- A context free grammar is said to be in chomsky normal form (CNF) if all its productions are of the form-

$$\mathbf{A \rightarrow BC}$$

or

$$\mathbf{A \rightarrow a}$$

where A, B, C are non-terminals

a is a terminal.

Steps :

- Reduce the grammar completely by-
  - Eliminating ∈ productions
  - Eliminating unit productions
  - Eliminating useless productions

- Add to the solution the productions which are already in CNF

- For the remaining non CNF productions,
  - Replace the terminals with some variables
  - Limit the number of variables at RHS to 2

Example 1

- Convert the given grammar to CNF-

$$S \rightarrow aAD$$
$$A \rightarrow aB \; / \; bAB$$
$$B \rightarrow b$$
$$D \rightarrow d$$

- **Step-01:**
  - The given grammar is already completely reduced.
- **Step-02:**

The productions already in chomsky normal form are-

$B \rightarrow b$ ....(1)

$D \rightarrow d$ ....(2)

These productions will remain as they are.

- The productions not in chomsky normal form are-

$S \to aAD$ ....(3)

$A \to aB \,/\, bAB$ ....(4)

We will convert these productions in chomsky normal form.

Step-03:

- Replace the terminal symbols a and b by new variables R_1 and R_2.
- This is done by introducing the following two new productions in the grammar-

$R_1 \to a$ ....(5)

$R_2 \to b$ ....(6)

- Now, the productions (3) and (4) modifies to-

  $S \rightarrow R_1AD$ ....(7)

  $A \rightarrow R_1B \,/\, R_2AB$ ....(8)

- Replace AD and AB by new variables R_3 and R_4 respectively.

  $R_3 \rightarrow AD$ ....(9)

  $R_4 \rightarrow AB$ ....(10)

  $S \rightarrow R_1 R_3$ ....(11)

  $A \rightarrow R_1B \,/\, R_2R_4$ ....(12)

- From (1), (2), (5), (6), (9), (10), (11) and (12), the resultant grammar is-

$S \rightarrow R_1R_3$
$A \rightarrow R_1B \,/\, R_2R_4$
$R_1 \rightarrow a$
$R_2 \rightarrow b$
$R_3 \rightarrow AD$
$R_4 \rightarrow AB$
$B \rightarrow b$
$D \rightarrow d$

- This grammar is in chomsky normal form.

(1) Express CFG to CNF.

$S \to aSbb \mid bSa \mid aa \mid a \mid bb \mid b.$

Productions	Solutions
$S \to a$	$S \to a$
$S \to b$	$S \to b.$
$S \to aa$	$C_1 \to a$
$S \to C_1 C_1$	$S \to C_1 C_1$
$S \to bb$	& $C_2 \to b$
$S \to C_2 C_2$	$S \to C_2 C_2$
$S \to bSa$	
$S \to C_2 S C_1$	$C_3 \to S C_1$
$S \to C_2 C_3$	$S \to C_2 C_3$
$S \to aSbb$	
$S \to C_1 S C_2 C_2$	$S \to C_1 C_4$
$C_4 \to S C_2 C_2$	$C_4 \to S C_5$
$C_5 \to C_2 C_2.$	$C_5 \to C_2 C_2.$

$\therefore$ CFG in CNF is :

$S \rightarrow a \mid b \mid C_1C_1 \mid C_2C_2 \mid C_2C_3 \mid C_1C_4$

$C_1 \rightarrow a$

$C_2 \rightarrow b$

$C_3 \rightarrow SC_1$

$C_4 \rightarrow SC_5$

$C_5 \rightarrow C_2C_2$.

(2) Convert foll. to CNF.

$S \rightarrow aSb \mid bSa \mid ba \mid ab \mid \epsilon$.

Soln:

Step 1: Eliminate Null production.

$S \rightarrow \epsilon$.

$S \rightarrow aSb \mid bSa \mid ab \mid ba$

Now, considering this grammar for CNF conversion -

$S \rightarrow ab$ $C_1 \rightarrow a$ $C_2 \rightarrow b$

$S \rightarrow C_1 C_2$ $S \rightarrow C_1 C_2$.

$S \rightarrow ba$

$S \rightarrow C_2 C_1$ $S \rightarrow C_2 C_1$

$S \rightarrow aSb$

$S \rightarrow C_1 S C_2$ $S \rightarrow C_1 C_3$

$C_3 \rightarrow \cancel{C_1 C} S C_2$ $C_3 \rightarrow S C_2$.

$S \rightarrow bSa$

$S \rightarrow C_2 S C_1$ $S \rightarrow C_2 C_4$

$C_4 \rightarrow S C_1$.

∴ Grammar in CNF becomes -

$S \rightarrow C_1 C_2 \mid C_2 C_1 \mid C_1 C_3 \mid C_2 C_4$

$C_1 \rightarrow a$

$C_2 \rightarrow b$

$C_3 \rightarrow S C_2$

$C_4 \rightarrow S C_1$.

(3) $S \to 11S \mid 00 \mid 00S \mid 11 \mid e$.

Sol^n

Removing e-prod.

$S \to 11S \mid 00 \mid 00S \mid 11$

Now, considering this grammar for CNF conversion.

$S \to 00$

$C_1 \to 0$

~~S~~ $S \to C_1 C_1$

$S \to 11$

$C_2 \to 1$

$S \to C_2 C_2$.

$S \to \cancel{00}\ 11S$

$S \to C_2 C_2 S$

$S \to C_2 C_3$

$C_3 \to C_2 S$.

$S \to 00S$

$S \to C_1 C_1 S$

$S \to C_1 C_4$

$C_4 \to C_1 S$

∴ CFG in CNF becomes -

$S \rightarrow C_1 C_1 \mid C_2 C_2 \mid C_2 C_3 \mid C_1 C_4$.

$C_1 \rightarrow 0$

$C_2 \rightarrow 1$

$C_3 \rightarrow C_2 S$

$C_4 \rightarrow C_1 S$.

(1)

$S \rightarrow \sim S$

$S \rightarrow [\, S \in S \,]$

$S \rightarrow p \mid q$

$S \rightarrow p$

$S \rightarrow q$

$\cancel{S \rightarrow [S \in S]}$

$S \rightarrow \sim S$

$S \rightarrow [S \in S]$

$S \rightarrow p$

$S \rightarrow q$

$C_1 \rightarrow \sim$

$S \rightarrow C_1 S$

$C_2 \rightarrow [\quad C_3 \rightarrow \in \quad C_4 \rightarrow]$

$S \rightarrow C_2 \; S \; C_3 \; S \; C_4$

$S \rightarrow C_2 C_5$

$C_5 \rightarrow S C_6$

$C_6 \rightarrow C_3 C_7$

$C_7 \rightarrow S C_4.$

$S \rightarrow p \mid q \mid C_1 S \mid C_2 C_5$

$C_1 \rightarrow \sim$

$C_2 \rightarrow [$

$C_3 \rightarrow \in$

$C_4 \rightarrow]$

$C_5 \rightarrow S C_6$

$C_6 \rightarrow C_3 C_7$

$C_7 \rightarrow S C_4.$

(5) $S \to ABC \mid aC$

$A \to a$

$B \to b$

$C \to c.$

Step I : This grammar is already simplified.

Step II –

$\cancel{A} \to a$	$A \to a$
$B \to b$	$B \to b$
$C \to c$	$C \to c$
$S \to aC$	$S \to AC$
$S \to ABC$	$S \to AC_1$
	$C_1 \to BC$

$S \to AC_1 \mid AC$

$A \to a$

$B \to b$

$C \to c$

$C_1 \to BC.$

(6) $S \rightarrow AACD$

$A \rightarrow aAb \mid \epsilon$

$C \rightarrow aC \mid a$

$D \rightarrow aDa \mid bDb \mid \epsilon.$

Step I Simplify G.

Removing ϵ-productions.

$S \rightarrow AACD \mid ACD \mid AAC \mid CD \mid AC \mid C$

$A \rightarrow aAb \mid ab$

$C \rightarrow aC \mid a$

$D \rightarrow aDa \mid bDb \mid aa \mid bb$

(b) Removing Unit Productions.

$S \rightarrow AACD \mid ACD \mid AAC \mid CD \mid AC \mid aC \mid a$

$A \rightarrow aAb \mid ab$

$C \cancel{B} \rightarrow aC \mid a$

$D \rightarrow aDa \mid bDb \mid aa \mid bb.$

(c) Removing useless symbols.

There are no useless symbols.

Step 2. $S \rightarrow AACD$

$S \rightarrow AR_1$

$R_1 \rightarrow AR_2$

$R_2 \rightarrow CD$

$S \rightarrow ACD$

$S \rightarrow AR_2$

$S \to AAC$

$S \to AC$

$S \to CD$

$S \to aC$

$S \to a.$

$S \to AR_3$

$R_3 \to AC.$

$S \to AC$

$S \to CD$

$S \to R_4 C$

$R_4 \to a$

$A \rightarrow aAb$

$A \rightarrow R_4 R_5$

$R_5 \rightarrow A R_6$

$R_6 \rightarrow b.$

$R_4 \rightarrow a.$

$A \rightarrow ab$

$A \rightarrow R_4 R_6.$

$C \rightarrow aC$

$\cancel{C \rightarrow R_7} C$ $C \rightarrow R_4 C$

$R_4 \rightarrow a$ (Already done)

$C \rightarrow a$

$D \rightarrow aDa$

$D \rightarrow \cancel{R_7 R_8}\ R_4 R_7$

$R_7 \rightarrow D R_4$

$D \rightarrow bDb$

$D \rightarrow R_6 R_8$

$R_8 \rightarrow D R_6.$

$D \rightarrow aa$

$D \rightarrow R_4 R_4$

$D \rightarrow bb$

$D \rightarrow R_6 R_6.$

$D \rightarrow bb$

$D \rightarrow R_6 R_6$.

$S \rightarrow CD \mid AC \mid AR_1 \mid AR_2 \mid AR_3 \mid R_4 C \mid a$

$R_1 \rightarrow AR_2$

$R_2 \rightarrow CD$

$R_3 \rightarrow AC$

$R_4 \rightarrow a$

$R_5 \rightarrow AR_6$

$R_6 \rightarrow b$

$A \rightarrow R_4 R_5 \mid R_4 R_6$

$R_5 \rightarrow AR_6$

$R_6 \rightarrow b$.

$C \rightarrow R_4 C$

$D \rightarrow R_4 R_7 \mid R_6 R_8 \mid R_4 R_4 \mid R_6 R_6$.

$R_7 \rightarrow D R_4$

$R_8 \rightarrow D R_6$.

(7)

$S \rightarrow bA \mid aB$

$A \rightarrow bAA \mid aS \mid \epsilon$.

$B \rightarrow aBB \mid bS \mid \epsilon$.

(a) Remove nullable symbols :-

$S \rightarrow bA \mid aB \mid b \mid a$

$A \rightarrow bAA \mid aS \mid bA \mid b$

$B \rightarrow aBB \mid aB \mid bS \mid a$

(b) Remove Useless symbols :-

No Useless symbols.

(c) No Unit productions.

Converting to CNF :

$S \rightarrow a$	$S \rightarrow a$
$S \rightarrow b$	$S \rightarrow b$
$S \rightarrow bA$	$R_1 \rightarrow b$
	$S \rightarrow R_1 A$
$S \rightarrow aB$	$R_2 \rightarrow a$
	$S \rightarrow R_2 B$
$A \rightarrow b$	
$A \rightarrow bA$	$A \rightarrow R_1 A$
$A \rightarrow aS$	$A \rightarrow R_2 S$
$A \rightarrow bAA$	$A \rightarrow R_3 A$
	$R_3 \rightarrow AA$
$B \rightarrow a$	
$B \rightarrow bS$	$B \rightarrow R_1 S$
$B \rightarrow aB$	$B \rightarrow R_2 B$
$B \rightarrow aBB$	$B \rightarrow \cancel{R_4 B}\ R_2 R_5$
	$\cancel{R_4 \rightarrow R_2 B}\ R_5 \rightarrow BB$

Hence, the CNF is :-

$$S \rightarrow R_1 A \mid R_2 B \mid a \mid b.$$
$$R_1 \rightarrow b$$
$$R_2 \rightarrow a$$
$$A \rightarrow R_1 A \mid R_2 S \mid R_3 A \mid b$$
$$R_3 \rightarrow AA$$
$$B \rightarrow R_1 S \mid R_2 R_5 \mid R_2 B \mid a.$$
$$R_5 \rightarrow BB.$$

1. The original CFG G_6 is shown on the left. The result of applying the first step to make a new start variable appears on the right.

$$
\begin{aligned}
S &\rightarrow ASA \mid \mathrm{a}B \\
A &\rightarrow B \mid S \\
B &\rightarrow \mathrm{b} \mid \varepsilon
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{S_0} &\rightarrow \boldsymbol{S} \\
S &\rightarrow ASA \mid \mathrm{a}B \\
A &\rightarrow B \mid S \\
B &\rightarrow \mathrm{b} \mid \varepsilon
\end{aligned}
$$

2. Remove ε-rules $B \rightarrow \varepsilon$, shown on the left, and $A \rightarrow \varepsilon$, shown on the right.

$$
\begin{aligned}
S_0 &\rightarrow S \\
S &\rightarrow ASA \mid \mathrm{a}B \mid \mathbf{a} \\
A &\rightarrow B \mid S \mid \boldsymbol{\varepsilon} \\
B &\rightarrow \mathrm{b} \mid \varepsilon
\end{aligned}
$$

$$
\begin{aligned}
S_0 &\rightarrow S \\
S &\rightarrow ASA \mid \mathrm{a}B \mid \mathrm{a} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \mid \boldsymbol{S} \\
A &\rightarrow B \mid S \mid \varepsilon \\
B &\rightarrow \mathrm{b}
\end{aligned}
$$

3a. Remove unit rules $S \to S$, shown on the left, and $S_0 \to S$, shown on the right.

$$\begin{array}{rcl} S_0 & \to & S \\ S & \to & ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \mid S \\ A & \to & B \mid S \\ B & \to & \mathrm{b} \end{array}$$

$$\begin{array}{rcl} S_0 & \to & S \mid ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ S & \to & ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ A & \to & B \mid S \\ B & \to & \mathrm{b} \end{array}$$

3b. Remove unit rules $A \to B$ and $A \to S$.

$$\begin{array}{rcl} S_0 & \to & ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ S & \to & ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ A & \to & B \mid S \mid \mathrm{b} \\ B & \to & \mathrm{b} \end{array}$$

$$\begin{array}{rcl} S_0 & \to & ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ S & \to & ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ A & \to & S \mid \mathrm{b} \mid ASA \mid \mathrm{a}B \mid \mathrm{a} \mid SA \mid AS \\ B & \to & \mathrm{b} \end{array}$$

Pumping Lemma for Context-Free Languages

- **Pumping Lemma:**
  Let G = (V, T, P, S) be a CFG in CNF, and let $n = 2^{|V|}$. If *z* is a string in L(G) and $|z| \geq n$, then there exist substrings u, v, w, x and y in T* such that z=uvwxy and:

  - $|vx| \geq 1$ (i.e., $|v| + |x| \geq 1$, or, non-null)
  - $|vwx| \leq n$ (the loop in generating this substring)
  - uv^iwx^iy is in L(G), for all $i \geq 0$

  - *Note: u could be of any length, so, vwx is not a prefix*
    - *unlike that (uv of uvw) in RL pumping lemma*

Context-Free Pumping Game for L:

1. C chooses an integer $p \geq 0$.
2. N chooses a string $s \in L$ such that $|s| \geq p$.
3. C chooses strings u, v, x, y, z such that $s = uvxyz$, $|vxy| \leq p$, and $|vy| >$
4. N chooses an integer $i \geq 0$ such that $uv^ixy^iz \notin L$.

Pumping Claim 2: If L is context-free, then C has a winning strategy.

Simple Example:

- Given Language is L = $\{0^n 1^n 0^n 1^n \mid n >= 0\}$
  - Let p be pumping length of L by pumping lemma
  - It is enough to show that string s = $0^p 1^p 0^p 1^p$ cannot be pump
  - Remember uv^ixy^iyz
  - s in form s = $uvxyz$
    - If both v and y contain at most one type of alphabet symbol, string wi uv^2xy^2z.
    - If either v or y contain more than one type of alphabet symbol, string which is not in correct order