

Chomsky normal form

Back (for review), by popular demand

What is Chomsky Normal Form?

It's a way to express the rules of a CFG

Every CFG may be written in normal form

What is the structure of Chomsky Normal Form?

CFG is in Chomsky normal form if every rule takes form:

$$A \rightarrow BC$$

$$A \rightarrow a$$

- B and C may not be the start variable
- Only the start variable can transition to ε
- Each variable goes to two other variables or two one terminal
- The start variable may not point back to the start variable


Consider a typical CFG

- Variables and terminals mix $A \rightarrow xBy$
- Some variables point to other single variables $A \rightarrow C$
- Start variable can point to itself $S \rightarrow SS \mid y$
- Any variable can transition to ε $B \rightarrow \varepsilon$

Converting to Chomsky Normal Form

- $S_0 \rightarrow S$ where S was original start variable
- Remove $A \rightarrow \varepsilon$

We won't use this rule in this class, will use all others



- For each multiple-occurrence of A , add new rules with A deleted

$$R \rightarrow uAvAw \quad \text{change to} \quad R \rightarrow uvAw \mid uAvw \mid uvw$$

- Shortcut all unit rules

$$\text{Given } A \rightarrow B \text{ and } B \rightarrow u, \text{ add } A \rightarrow u$$

- Replace rules $A \rightarrow u_1u_2u_3 \dots u_k$ with:

$$A \rightarrow u_1A_1, A_1 \rightarrow u_2A_2, A_2 \rightarrow u_3A_3, \dots, A_{k-2} \rightarrow u_{k-1}u_k$$

Let's Chomsky-ize a non-Chomsky form grammar

G1:

$S \rightarrow AB$

$A \rightarrow cAn \mid c$

$B \rightarrow BB \mid n$

Replace terminals-
variable mixes with
variables only

$S \rightarrow AB$

$A \rightarrow U_c A U_n \mid c$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$S \rightarrow AB$

$A \rightarrow U_c D \mid c$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow A U_n$

Convert 3-variable rules
to 2-variable rules

More on G1: Typical to add

Replace S in
 $S_0 \rightarrow S$ rule

G1:

$S \rightarrow AB$

$A \rightarrow U_c D \mid c$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow AU_n$

new start state

$S_0 \rightarrow S$

$S \rightarrow AB$

$A \rightarrow U_c D \mid c$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow AU_n$

Final answer!

$S_0 \rightarrow AB$

$S \rightarrow AB$

$A \rightarrow U_c D \mid c$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow AU_n$

This already fits
normal form!

Reminder:

CFG is in Chomsky normal form if every rule takes form:

$$A \rightarrow BC$$

$$A \rightarrow a$$

- B and C may not be the start variable
- Only the start variable can transition to ε
- Each variable goes to two other variables or two one terminal
- The start variable may not point back to the start variable

Let's get more complicated with grammar G2

G2:

$S \rightarrow AB$

$A \rightarrow cAn \mid c \mid \varepsilon$

$B \rightarrow BB \mid n$

Let's add new start state first this time:

$S_0 \rightarrow S$

$S \rightarrow AB$

$A \rightarrow cAn \mid c \mid \varepsilon$

$B \rightarrow BB \mid n$

$S_0 \rightarrow S$

$S \rightarrow AB \mid \varepsilon B$

$A \rightarrow cAn \mid c \mid \varepsilon \mid c\varepsilon n$

$B \rightarrow BB \mid n$

To remove ε , first plug it in wherever it applies

More on G2

Finish removing ϵ

Replace terminals-variable mixes with variables only

$$S_0 \rightarrow S$$

$$S \rightarrow AB \mid \epsilon B$$

$$A \rightarrow cAn \mid c \mid \epsilon \mid c\epsilon n$$

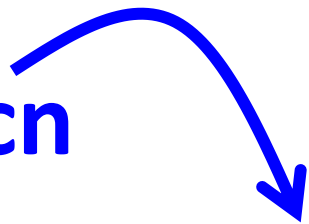
$$B \rightarrow BB \mid n$$


$$S_0 \rightarrow S$$

$$S \rightarrow AB \mid B$$

$$A \rightarrow cAn \mid c \mid cn$$

$$B \rightarrow BB \mid n$$


$$S_0 \rightarrow S$$

$$S \rightarrow AB \mid B$$

$$A \rightarrow U_c A U_n \mid c \mid U_c U_n$$

$$B \rightarrow BB \mid n$$

$$U_c \rightarrow c$$

$$U_n \rightarrow n$$

More on G2

Convert 3-variable rules
to 2-variable rules

Replace single variables
on right side (S, B)

$S_0 \rightarrow S$

$S \rightarrow AB \mid B$

$A \rightarrow U_c A U_n \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$S_0 \rightarrow S$

$S \rightarrow AB \mid B$

$A \rightarrow U_c D \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow AU_n$

$S_0 \rightarrow AB \mid BB \mid n$

$S \rightarrow AB \mid BB \mid n$

$A \rightarrow U_c D \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow AU_n$

G2 final answer

$S_0 \rightarrow AB \mid BB \mid n$

$S \rightarrow AB \mid BB \mid n$

$A \rightarrow U_c D \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow AU_n$

Produces same
CFL as:

$S \rightarrow AB$

$A \rightarrow cAn \mid c \mid \varepsilon$

$B \rightarrow BB \mid n$

Reminder:

CFG is in Chomsky normal form if every rule takes form:

$$A \rightarrow BC$$

$$A \rightarrow a$$

- B and C may not be the start variable
- Only the start variable can transition to ε
- Each variable goes to two other variables or two one terminal
- The start variable may not point back to the start variable

Let's get even more complicated

G3:

$S \rightarrow AB \mid BS$

$A \rightarrow cAn \mid c \mid \varepsilon$

$B \rightarrow BB \mid n$

Add new start state:

$S_0 \rightarrow S$

$S \rightarrow AB \mid BS$

$A \rightarrow cAn \mid c \mid \varepsilon$

$B \rightarrow BB \mid n$

Remove ε

$S_0 \rightarrow S$

$S \rightarrow AB \mid BS \mid \mathbf{B}$

$A \rightarrow cAn \mid c \mid \mathbf{cn}$

$B \rightarrow BB \mid n$

More on G3

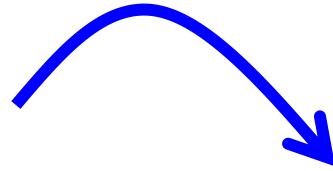
Replace terminals-
variable mixes with
variables only

$S_0 \rightarrow S$

$S \rightarrow AB \mid BS \mid B$

$A \rightarrow cAn \mid c \mid cn$

$B \rightarrow BB \mid n$



$S_0 \rightarrow S$

$S \rightarrow AB \mid BS \mid B$

$A \rightarrow U_c A U_n \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

More on G3

Replace single variables
on right side (S, B)

$$S_0 \rightarrow S$$

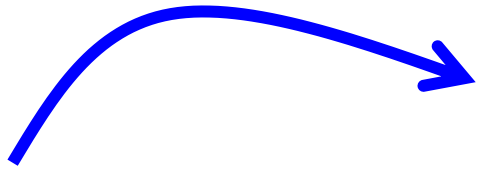
$$S \rightarrow AB \mid BS \mid \mathbf{B}$$

$$A \rightarrow \mathbf{U}_c A \mathbf{U}_n \mid c \mid \mathbf{U}_c \mathbf{U}_n$$

$$B \rightarrow BB \mid n$$

$$\mathbf{U}_c \rightarrow c$$

$$\mathbf{U}_n \rightarrow n$$



$$S_0 \rightarrow \mathbf{AB} \mid \mathbf{BS} \mid \mathbf{BB} \mid n$$

$$S \rightarrow AB \mid BS \mid \mathbf{BB} \mid n$$

$$A \rightarrow \mathbf{U}_c A \mathbf{U}_n \mid c \mid \mathbf{U}_c \mathbf{U}_n$$

$$B \rightarrow BB \mid n$$

$$\mathbf{U}_c \rightarrow c$$

$$\mathbf{U}_n \rightarrow n$$

More on G3

Convert 3-variable rules
to 2-variable rules



$S_0 \rightarrow AB \mid BS \mid BB \mid n$

$S \rightarrow AB \mid BS \mid BB \mid n$

$A \rightarrow U_c A U_n \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$S_0 \rightarrow AB \mid BS \mid BB \mid n$

$S \rightarrow AB \mid BS \mid BB \mid n$

$A \rightarrow U_c D \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$

$D \rightarrow A U_n$

Reminder:

CFG is in Chomsky normal form if every rule takes form:

$$A \rightarrow BC$$

$$A \rightarrow a$$

- B and C may not be the start variable
- Only the start variable can transition to ε
- Each variable goes to two other variables or two one terminal
- The start variable may not point back to the start variable