Context free grammar $G = (N, \Sigma, P, S)$

Parse tree is a tree satisfying the following Conditions.

1. Each interior node is labelled with an element of N
2. Each leaf node is labelled with Σ or ϵ.
3. if an interior node is labelled A and its children are labelled $B_1, \ldots B_k$.

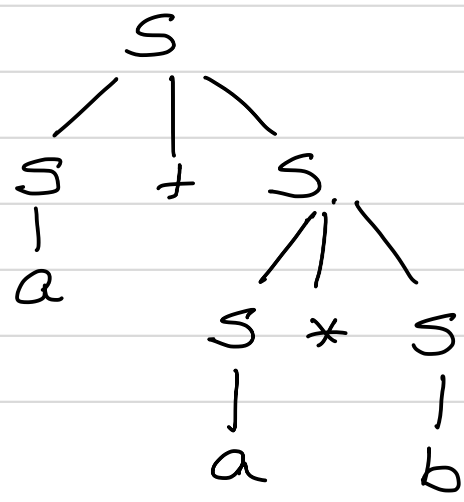
Then $A \rightarrow B_1 B_2, \ldots B_k \in P$.

$$S \rightarrow S+S \mid S*S \mid (S) \mid I$$

$$I \rightarrow a \mid b$$

String $a + a*b$

$$S \rightarrow S+S \rightarrow S + S*S \xrightarrow{*} a+a*b.$$



$S \underset{G}{\rightarrow} SS \underset{G}{\rightarrow} 0S \underset{G}{\rightarrow} 01$

$G = (N, \Sigma, P, S)$

$S \rightarrow SS \mid 0 \mid 1 \mid \epsilon.$

CNF

$S \rightarrow SS \mid 0 \mid 1$

$S \xrightarrow[G]{1} SS \xrightarrow[G]{1} SSS \xrightarrow[G]{1} SSSS \xrightarrow[G]{1} SSS \xrightarrow[G]{1} SS \rightarrow 0S \rightarrow 01$

Chomsky Normal Form. (CNF)

$G = (N, \Sigma, P, S)$ is in CNF if all productions

$A \rightarrow BC \quad A \rightarrow a \quad A, B, C \in N, \; a \in \Sigma.$

Example. $S \rightarrow [S] \mid SS \mid \epsilon \Rightarrow G_1$

G_2

$S \rightarrow AB \mid AC \mid SS \quad C \rightarrow SB \quad A \rightarrow [\quad B \rightarrow]$

$L(G_1) = L(G_2)$.

One step progress. $\begin{cases} \text{- \# of Nonterminals increase by 1} \\ \text{- \# of terminal increase by 1.} \end{cases}$

Theorem. For any CFG G, there is a CFG G' in Chomsky normal form s.t $L(G') = L(G) - \{\epsilon\}$.

Lemma 1. For any CFG $G = (N, \Sigma, P, S)$ there is a CFG G' with no ϵ-productions or unit-productions such that $L(G') = L(G) - \{\epsilon\}$

Proof. Let $\hat{P}$ be the smallest set of productions containing P and closed under the rules:

(a) if $A \to \alpha B \beta$ and $B \to \epsilon$ are in $\hat{P}$ then $A \to \alpha\beta \in \hat{P}$

(b) if $A \to B$ and $B \to \gamma$ are in $\hat{P}$ then $A \to \gamma \in \hat{P}$.

Note: $\hat{P}$ is finite

Finitely many new production rules are added
Each new RHS is a substring of an old RHS.

$\hat{G} = (N, \Sigma, \hat{P}, S)$

We have $L(G) \subseteq L(\hat{G})$ since $P \subseteq \hat{P}$

$L(G) = L(\hat{G})$. each new production was included because of rule (a) or (b) - can be simulated in 2 steps by two productions that caused it to be included.

Claim 2. For any non-null $x \in \Sigma^*$, any derivation $S \xrightarrow[\hat{G}]{*} x$ of minimum length does not use ϵ- or unit productions.

Proof. Let $x \neq \epsilon$. Let $S \xrightarrow[\hat{G}]{*} x$ be the minimum length derivation.

Suppose an ϵ-production $B \to \epsilon$ is used at somepoint

$$S \xrightarrow[\hat{G}]{*} \gamma B \delta \xrightarrow[\hat{G}]{1} \gamma \delta \xrightarrow[\hat{G}]{*} x.$$

At least one of γ or δ is non-null $\Rightarrow$ B was introduced from a production of the form $A \to \alpha B \beta$.

$$S \xrightarrow[\hat{G}]{m} \eta A \theta \xrightarrow[\hat{G}]{1} \eta \alpha B \beta \theta \xrightarrow[\hat{G}]{n} \gamma B \delta \xrightarrow[\hat{G}]{1} \gamma \delta \xrightarrow[\hat{G}]{k} x$$

for $m, n, k \geq 0$

By rule (a) $A \to \alpha\beta \in \hat{P}$.

But then we have a strictly shorter derivation of x

$$S \xrightarrow[\hat{G}]{m} \eta A \theta \xrightarrow[\hat{G}]{1} \eta \alpha \beta \theta \xrightarrow[\hat{G}]{n} \gamma \delta \xrightarrow[\hat{G}]{k} x.$$

This gives a contradiction.

Unit Productions

Let $x \neq \epsilon$. Consider a derivation $S \xrightarrow[\hat{G}]{*} x$ of minimum length.

Suppose a unit production $A \to B$ is used at some point

$$S \xrightarrow[\hat{G}]{*} \alpha A \beta \xrightarrow[\hat{G}]{1} \alpha B \beta \xrightarrow[\hat{G}]{*} x.$$

B must be removed later by applying a production $B \to \gamma$.

$$S \xrightarrow[\hat{G}]{m} \alpha A \beta \xrightarrow[\hat{G}]{1} \alpha B \beta \xrightarrow[\hat{G}]{n} \eta B \theta \xrightarrow[\hat{G}]{1} \eta \gamma \theta \xrightarrow[\hat{G}]{k} x.$$

By rule (b), $A \to \gamma \in \hat{P}$.

But then, there is a shorter derivation of x

$$S \xrightarrow[\hat{G}]{m} \alpha A \beta \xrightarrow[\hat{G}]{1} \alpha \gamma \beta \xrightarrow[\hat{G}]{n} \eta \gamma \theta \xrightarrow[\hat{G}]{k} x.$$

This is a contradiction.

Claim 2 implies we can remove the ϵ-productions and unit productions from $\hat{P}$ without changing the language.

Chomsky Normal Form.

By Lemma 1, $L(G) = L(\hat{G})$ and $\hat{P}$ does not have ϵ-productions or unit productions.

For each terminal $a \in \Sigma$ introduce a new nonterminal A_a and add the production rule $A_a \to a$.

Replace all occurences of a on the RHS of old productions (except productions of the form $B \to a$) with A_a. Then all productions are of the form:

$$A \to a \quad \text{or} \quad A \to \underbrace{B_1 B_2 \cdots B_k}_{\text{nonterminals.}} \quad k \geq 2$$

For any production of the form $A \to B_1 B_2 \cdots B_k$ with $k \geq 3$, introduce a new nonterminal C and replace with

$$A \to B_1 C \quad \text{and} \quad C \to B_2 \cdots B_k.$$

Repeat until all RHS of all productions are of length atmost 2.

(a) if $A \to \alpha B \beta$ and $B \to \epsilon$ are in $\hat{P}$ then $A \to \alpha\beta \in \hat{P}$

(b) if $A \to B$ and $B \to \gamma$ are in $\hat{P}$ then $A \to \gamma \in \hat{P}$.

Example 1: $\{a^n b^n \mid n \geq 0\} - \{\epsilon\} = \{a^n b^n \mid n \geq 1\}$.

$S \to aSb \mid \epsilon$

$S \to aSb \mid ab$

Add nonterminals A, B

$S \to ASB \mid AB \quad A \to a \quad B \to b$

Add nonterminal C, replace $S \to ASB$ with $S \to AC$, $C \to SB$

G': $S \to AB \mid AC$, $C \to SB$, $A \to a$, $B \to b$

Balanced Paranthesis $S \to [S] \mid SS \mid \epsilon$

$S \to [S] \mid SS \mid [\,]$

Add new nonterminals A, B

$S \to ASB \mid SS \mid AB$, $A \to [$, $B \to]$

Add a new nonterminal C. Replace $S \to ASB$ with $S \to AC$ and $C \to SB$.

G': $S \to AB \mid AC \mid SS$, $C \to SB$, $A \to [$, $B \to]$