1 Context-free grammar

A **context-free grammar** (CFG) can be defined as the tuple $G = (\Gamma, \Sigma, P, S)$, where:

- Γ : variables (A,B,...)
- Σ : terminals (a,b,...)
- P : productions (e.g A $\rightarrow$ aB)
- S : start symbol ($S \in \Gamma$)

Example (Palindromes):

$P \rightarrow \varepsilon$ (the empty string is a palindrome)
$P \rightarrow 0$
$P \rightarrow 1$
$P \rightarrow 0P0$ (any palindrome surrounded by two 0s is also a palindrome)
$P \rightarrow 1P1$ (any palindrome surrounded by two 1s is also a palindrome)

We could write this in a simpler way as:

$P \rightarrow \varepsilon \mid 0 \mid 1 \mid 0P0 \mid 1P1$
(the vertical line represents OR, allowing you to merge multiple productions with the same head)

Other example:

$\Sigma = \{a, b, 0, 1, (,), *\}$

$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$E \rightarrow I \mid E * E \mid E + E \mid (E)$

Leftmost derivation $E \Rightarrow_{lm} E * E \Rightarrow_{lm} I * E \Rightarrow_{lm} a * E \Rightarrow_{lm} a * (E) \Rightarrow_{lm} a * (E + E) \Rightarrow_{lm} ...$

Rightmost derivation $E \Rightarrow_{rm} E * E \Rightarrow_{rm} E * (E) \Rightarrow_{rm} E * (E + E) \Rightarrow_{rm} E * (E + I) \Rightarrow_{rm} ...$

Continue the above derivations, and show that $a * (a + b00)$ is in the language of E.

2 Chomsky Normal Form

Every context-free language without ε has a grammar G in which all productions have one of the following forms:

1. $A \rightarrow BC$, where A, B and C are all variables, or
2. $A \rightarrow a$, where A is variable and a is a terminal

Further, G also has no useless symbols. Such a grammar is in *Chomsky Normal Form* (CNF).

We say that X is a useful symbol if: $S \Rightarrow^* \alpha X \beta \Rightarrow^* w$
X is generating if: $X \Rightarrow^* w$
X is reachable if: $S \Rightarrow^* \alpha X \beta$
Useful symbols are both generating and reachable.

What do you think of this grammar?
$S \to AB \mid a$
$A \to b$

Answer:
B is a useless symbol. It's reachable ($S \to AB$), but not generating. Because of this, we should delete $S \to AB$. This leaves us with
$S \to a$
$A \to b$
Again, A is useless. It's generating ($A \to b$), but it's not reachable from S. Deleting $A \to b$ leaves us with
$S \to a$
which is the final form of our grammar.

Variable A is nullable if $A \Rightarrow^* \varepsilon$

To convert a grammar to CNF, we should:

1. remove useless symbols
2. remove ε productions (e.g. $A \to \varepsilon$)
3. remove unit productions (e.g. $A \to B$)
4. modify bodies of length 2 or more to contain only variables
5. break bodies of length 3 or more

Example 1:
$S \to AB$
$A \to aAA \mid \varepsilon$
$B \to bBB \mid \varepsilon$

Answer:
All symbols are useful. A and B are reachable ($S \Rightarrow AB$), and they are both generating ($A \Rightarrow^* a$ and $B \Rightarrow^* b$).
We have to get rid of ε-productions $A \to \varepsilon$ and $B \to \varepsilon$
$S \to AB \mid A \mid B$
$A \to aAA \mid aA \mid a$
$B \to bBB \mid bB \mid b$
We should remove unit productions $S \to A$ and $S \to B$
$S \to AB \mid aAA \mid aA \mid a \mid bBB \mid bB \mid b$
$A \to aAA \mid aA \mid a$
$B \to bBB \mid bB \mid b$
To modify bodies of length 2 or more to contain only variables, we introduce $A_1 \to a$ and $B_1 \to b$
$S \to AB \mid A_1AA \mid A_1A \mid a \mid B_1BB \mid B_1B \mid b$
$A \to A_1AA \mid A_1A \mid a$
$B \to B_1BB \mid B_1B \mid b$
$A_1 \to a$
$B_1 \to b$
To break up bodies of length 3 or more, we introduce $A_2 \to A_1A$ and $B_2 \to B_1B$
$S \to AB \mid A_2A \mid A_1A \mid a \mid B_2B \mid B_1B \mid b$
$A \to A_2A \mid A_1A \mid a$
$B \to B_2B \mid B_1B \mid b$
$A_1 \to a$

$B_1 \to b$
$A_2 \to A_1 A$
$B_2 \to B_1 B$
The above grammar is in CNF, because it satisfies the above requirements. If we still had more productions with body length 3 or more, we would have to repeat this last step again in a similar fashion.

Example 2:
$S \to aXbX$
$X \to aY \mid bY \mid \varepsilon$
$Y \to X \mid c$

Example 3:
$S \to AbA \mid a$
$A \to Aa \mid \varepsilon$