Context-Free Grammars (CFG)

Formally, a CFG $G$ is a 4-tuple $G = (\Sigma, N, P, S)$

- $\Sigma$ - non-empty, finite alphabet of *terminal* symbols
- $N$ - non-empty, finite set of *non-terminals* (often also called *variables*)
- $P$ - a finite set of production rules of the form:
  \[ A \to \beta \text{ where } A \in N, \beta \in (N \cup \Sigma)^* \]
- $S$ - a start non-terminal, $S \in N$

We frequently use $(N \cup \Sigma)$, and denote it as $V$ for "vocabulary";
e.g. $V^* = (N \cup \Sigma)^*$. 
Notation

Shortform: Multiple rules with the same *lethand side* (LHS), can be written on a single line with the *righthand sides* (RHS) separated by | (meaning OR).

\[
S \rightarrow \epsilon \\
S \rightarrow (S) \\
S \rightarrow SS
\]

Shortform: \( S' \rightarrow \epsilon \ | \ (S) \ | \ SS \)
Conventions

We often follow the following conventions:

- Early lowercase letters: $a, b, c, \ldots$ are symbols from $\Sigma$
- Late lowercase letters: $\ldots, w, x, y, z$ are words from $\Sigma^*$
- Uppercase letters: $A, B, \ldots, S, \ldots$ are non-terminals from $N$
- $S$ is the starting non-terminal
- Lowercase Greek letters: $\alpha, \beta, \gamma, \ldots$ are elements of $V^*$ used in the RHS of production rules which may contain both terminals and non-terminals
Derivations

The application of production rules is called a derivation; i.e. from some initial form, we apply a production rule to obtain the next form. The symbol $\Rightarrow$ means “derives”.

- $\alpha \Rightarrow \beta$ means $\beta$ can be derived from $\alpha$ by the application of one production rule
- $\alpha \Rightarrow^k \beta$ means $\beta$ can be derived from $\alpha$ by the application of $k$ production rules
- $\alpha \Rightarrow^* \beta$ means $\beta$ can be derived from $\alpha$ by the application of 0 or more production rules

Example: $\alpha \Rightarrow^k \beta$ means $\alpha = \delta_0 \Rightarrow \delta_1 \Rightarrow \ldots \Rightarrow \delta_k = \beta$
Intermediate Forms

We typically start from $S$ and apply production rules until a word $w$ of only terminals is derived. The intermediate strings, however, may contain both terminals and non-terminals.

i.e. $S \Rightarrow^* w$ where $S \in N$ and $w \in \Sigma^*$

Intermediate steps may have the form:

$$\alpha A \beta \Rightarrow \alpha \gamma \beta$$

if there is a production $A \rightarrow \gamma$ in $P$

- $\alpha, \beta, \gamma \in V^* = (N \cup \Sigma)^*$ and $A \in N$
- RHS is derivable from LHS in one step

Know your notation: do not confuse $\Rightarrow$ with $\rightarrow$. 
The language of CFG $G$ is the set of strings (terminals only) that we can be derived from the starting non-terminal $S$, i.e.

$$L(G) = \{ w \in \Sigma^* \mid S \Rightarrow^* w \}.$$ 

A language $L$ is context-free if $L = L(G)$ for some CFG $G$.

The language $L = \{ \text{strings of balanced parentheses} \}$ is context-free:

- **CFG $G$:** $S \rightarrow \epsilon \mid (S) \mid SS$
- $L = L(G)$