In general, to describe a language, there are two possible approaches:

1) **Recognition**: describe rules (or a mechanism) to determine whether or not a certain string belongs to a language.

   e.g. an automaton is such a mechanism.

2) **Generation**: define rules to generate all strings of a language.

A grammar is a formalism for defining a language in terms of rules that generate all strings of the language.

Since 1950, various formal methods based on the notion of rewriting or derivation have been proposed by Noam Chomsky, Emil Post, A.A. Markov.

In the mid 1950s, the linguist Noam Chomsky introduced the notion of formal grammar with the aim of formalizing natural language. Formal grammars are in fact too simplistic to capture natural language, but they were adapted as the main formal tool to define syntactic properties of artificial languages (e.g. programming languages).
Definition: Given an alphabet \( \Sigma \), a (formal) grammar \( G \)
 is a quadruple \( G = (V_N, V_T, P, S) \) where
- \( V_T \subseteq \Sigma \) is a finite nonempty set of symbols called terminals
- \( V_N \) is a finite nonempty set of symbols s.t. \( V_N \cap \Sigma = \emptyset \),
  called variables or nonterminals, or syntactic categories.
  Each variable represents a language
- \( S \in V_N \) is called start symbol or axiom, and represents
  the language being defined by \( G \)
- \( P \) is a binary relation over
  \[
  (V_N \cup V_T)^* \times V_N \times (V_N \cup V_T)^* \times (V_N \cup V_T)^*
  \]
  Each element \((\alpha, \beta) \in P\) is called a production or rule,
  and is generally written as: \( \alpha \rightarrow \beta \).

Note: \( \alpha \) ... sequence of terminals and nonterminals with
  at least one nonterminal

\( \beta \) ... sequence of terminals and nonterminals

Definition: The language \( L(G) \) generated by a grammar \( G \)
 is the set of strings of terminals only that can be generated
 starting from the axiom by a finite sequence of rule applications
 Each application of a rule \( \alpha \rightarrow \beta \) consists in replacing an
 occurrence of \( \alpha \) with \( \beta \).
Example: Palindromes:

A palindrome is a word that reads the same both forwards and backwards. (AILALI, AMORONGA)

$L_{pal} = \{w \in \{0, 1\}^* \mid w^R = w\}$

Grammar $G_{pal} = (V_N, V_T, P, S)$, where $P$ consists of:

1. $S \rightarrow \epsilon$
2. $S \rightarrow 0$
3. $S \rightarrow 1$
4. $S \rightarrow 0SO$
5. $S \rightarrow 1SI$

Example of derivation:

0110 : $S \rightarrow 0SO \rightarrow 01SI0 \rightarrow 0110$

1011 : $S \rightarrow 1SI \rightarrow 11SI1 \rightarrow 1011$

Exercise E5.1: Prove that the above grammar generates all and only palindromes over $\{0, 1\}$.

Hint: use induction on the length of the derivation.

Example: Natural language generation

Sentence $\rightarrow$ NounPhrase VerbPhrase

NounPhrase $\rightarrow$ Adjective NounPhrase

NounPhrase $\rightarrow$ Noun

Noun $\rightarrow$ car

Noun $\rightarrow$ train

Adjective $\rightarrow$ big

Adjective $\rightarrow$ broken
Notation:

1) To denote the set of productions
\[ \alpha \rightarrow \beta_1 \mid \alpha \rightarrow \beta_2 \mid \ldots \mid \alpha \rightarrow \beta_m \]
we use
\[ \alpha \rightarrow \beta_1 \mid \beta_2 \mid \ldots \mid \beta_m \]

2) We use \[ V = V_N \cup V_T \]

A production of the form \[ \alpha \rightarrow \epsilon \], with \( \alpha \in V^* \cup \{ \epsilon \} \), is called \underline{\epsilon-production}.

Example: \[ L_{eq} = \{ w \in \{0,1\}^* \mid w \text{ has equal number of 0's and 1's} \} \]

We have already seen that this language is not regular.

Idea to define \( G_{eq} \) s.t.: \( L(G_{eq}) = L_{eq} \): use induction

\[ L_{eq} \]

base: \( \epsilon \in \epsilon \cdot L_{eq} \)

induction: \( \text{OW}_a \in L_{eq} \) if \( w_a \) has one more 1 than 0
\[ \text{OW}_b \in L_{eq} \] if \( w_b \) has one more 0 than 1

Characterize else languages for \( w_a \) and \( w_b \) inductively

Grammar \( G_{eq} = (\{S,A,B\}, \{0,1\}, \{0,1\}, S) \) with \( P \)

\[ S \rightarrow \epsilon \mid OA \mid 1B \] \((A \text{ generates strings with one more 1 than 0.})\)
\[ A \rightarrow AS \mid 0AA \] \((B \text{ generates strings with one more 0 than 1.})\)
\[ B \rightarrow OS \mid 1BB \]

Exercise E5.2: Prove that \( L(G_{eq}) = L_{eq} \) (by induction)
Definition: Given $G$, the direct derivation for $G$ is the binary relation on $(V^* \circ V_\kappa \circ V^*) \times V^*$ defined as follows:

$(\varphi, \psi)$ is in the relation if there are

$\alpha \in V^* \circ V_\kappa \circ V^*$, \hspace{1em} \beta, \gamma, \delta \in V^*$

such that $\varphi = \rho \alpha \delta$, $\psi = \rho \beta \delta$ and $\alpha \Rightarrow \beta \in P$.

We write $\varphi \Rightarrow \psi$ and say that $\psi$ directly derives from $\varphi$ by $G$.

Definition: We call derivation the reflexive, transitive closure of direct derivation. In other words, $\psi$ derives from $\varphi$ by $G$, written $\varphi \Rightarrow^* \psi$, if

a) $\varphi = \psi$, or

b) there are $\varphi_1, \ldots, \varphi_n \in V^*$ such that $\varphi_1 = \varphi$, $\varphi_n = \psi$, and $\varphi_i \Rightarrow \varphi_{i+1}$, $\forall i \in 1 \leq i < n$.

Definition: Given a grammar $G$, the language generated by $G$ in $L(G) = \{w \in V_T^* \mid S \Rightarrow^* w\}$.

Notice: words in $L(G)$ are constituted by terminals only.

Terminology:

- sentence: any word $w \in V_T^*$ such that $S \Rightarrow^* w$, i.e. $w \in L(G)$

- sentential form: any $\alpha \in V^* = (V_T \cup V_N)^*$ such that $S \Rightarrow \alpha$

Notation:

- terminals: $a, b, c, \ldots$

- nonterminals: $A, B, C, \ldots$

- strings of terminals: $a, a, a, w$, $x, y, z$

- symbols of $V = V_T \cup V_N$: $X, Y, Z$

- sentential forms: $A, B, C$
Example: Productions for $G_{eq}$:

$$S \rightarrow \varepsilon \mid 0A \mid 1B$$
$$A \rightarrow 1S \mid 0AA$$
$$B \rightarrow 0S \mid 1BB$$

derivation:

1) $001SA \Rightarrow 00151S$ (using $A \rightarrow 1S$)
2) $00151S \Rightarrow 001MS$ (using $S \rightarrow \varepsilon$)
3) $001SA \Rightarrow 001MS$ (using (1) and (2))
4) $S \Rightarrow 001110$

Example: Grammar for $L_{3n} = \{a^n b^n c^n \mid n \geq 1\}$

$G_{3n} = (\{A, B, C, S\}, \{a, b, c\}, \{P, S\}, P, S)$

with:

1) $S \rightarrow aSBBC$  
2) $S \rightarrow aBC$  
3) $CB \rightarrow BC$  
4) $aB \rightarrow ab$  
5) $bB \rightarrow bb$  
6) $bC \rightarrow bc$  
7) $CC \rightarrow cc$

generate: $aaa\ldots BBBCBCBC$  

Example of derivation of $aaa bbb ccc$:

$$S \overset{1}{\Rightarrow} aSBBC \overset{2}{\Rightarrow} aaaaBBBCBCBC \overset{3}{\Rightarrow} aaaaBBBCBCBC$$
$$\overset{3}{\Rightarrow} aaaaBBBCBCBC \overset{4}{\Rightarrow} aaaaBBBCBCBC$$
$$\overset{5}{\Rightarrow} aaaaBBBCBCBC \overset{5}{\Rightarrow} aaaaBBBCBCBC$$
$$\overset{5}{\Rightarrow} aaaaBBBCBCBC \overset{6}{\Rightarrow} aaaaBBBCBCBC$$
$$\overset{6}{\Rightarrow} aaaaBBBCBCBC \overset{7}{\Rightarrow} aaaaBBBCBCBC$$
$$\overset{7}{\Rightarrow} aaaaBBBCBCBC$$
Note: not each sequence of direct derivations leads to a sentence in $L(G_{3,2})$

e.g. with the previous grammar we could generate

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aeeBCBCBC$
$\Rightarrow eeeBCBBCC \Rightarrow eeeBCCBBCC$
$\Rightarrow eeeBCCBBCC$

we cannot apply any other production

Also, the application of productions could go on forever (e.g. rule 1 in the previous example)

Classification of Chomsky grammars into 4 groups, depending on the form of the productions:

- grammars of type 0: no limitations
- 1: context-sensitive
- 2: context-free
- 3: regular (or right linear)

Definition: grammar of type 0.

Productions have the most general form $A \rightarrow \beta$,
with $A \in V \cup V_M \cup V^*$ and $\beta \in V^*$

Grammars of type 0 allow for derivations that shorten the sentential form.

A language generated by a grammar of type 0 is called a language of type 0.
Definition: grammar of type 1, or context sensitive

Productions have the form $A \rightarrow \beta$, with

$$\alpha \in V^* \cdot V_M \cdot V^*, \quad \beta \in V^+,$$

$|\alpha| \leq |\beta|$

These productions cannot shorten the length of the sentential form to which they are applied.

A language generated by a grammar of type 1 is called of type 1, or context sensitive.

Example: $G_{3n}$ is context sensitive. Obviously, it is also of type 0.

Definition: grammar of type 2, or context-free

Productions have the form $A \rightarrow \beta$, with $A \in V_M$, $\beta \in V^+$.

These productions are productions of type 1, with the additional requirement that on the left there is a single nonterminal.

A language generated by a grammar of type 2 is called of type 2, or context-free.

Example: $L_{2n} = \{a^n b^n | n \geq 1\}$ is of type 1, since the following grammar $G_{2n}^\prime$ generates $L_{2n}$

$$S \rightarrow aB \mid SAB \quad aA \rightarrow aA \quad aB \rightarrow aB$$

$L_{2n}$ is also of type 2, since it is generated by

$$S \rightarrow aSB \mid aB$$
We said that grammars of type 1 are also called context-sensitive (in contrast to context-free grammars). This is justified by the original definition by Chomsky for context-sensitive grammars.

**Definition:** Chomsky CS-grammar

Productions have the form \( \gamma_1 \alpha \gamma_2 \rightarrow \gamma_3 \beta \gamma_4 \)

with \( \gamma_1, \gamma_2, \gamma_3, \gamma_4 \in V^* \), \( \alpha \in V_N \), \( \beta \in V^+ \)

Intuitively, \( \alpha \) is replaced by \( \beta \) only if it appears "in the context" of \( \gamma_1 \) and \( \gamma_2 \).

**Theorem:** Grammars of type 1 and Chomsky CS grammars generate the same class of languages.

**Proof:** We show that for every language \( L \):

There is a type-1 grammar \( G_1 \) s.t. \( L = L(G_1) \) iff there is a Chomsky CS grammar \( G_C \) s.t. \( L = L(G_C) \)

"If": immediate, since each Chomsky CS grammar is of type 1.

"Only if": let \( G_1 \) be a type-1 grammar for \( L \).

We construct from \( G_1 \) a Chomsky CS grammar \( G_C \) as follows:

1) for each \( \alpha \in V_T \), add a new nonterminal \( N_\alpha \)
2) replace in each production of \( G_1 \), each \( \alpha \in V_T \) by \( N_\alpha \)

Now all productions have the form

\[ A_1 A_2 \ldots A_m \rightarrow B_1 B_2 \ldots B_n \]

with \( m \leq n \)

and all \( A_1, B_j \in V_N \)
3) For each such production \( A_1 \cdots A_m \rightarrow B_1 \cdots B_n \), introduce a new nonterminal \( N \), and replace the production by the following ones:

\[
\begin{align*}
A_1 A_2 \cdots A_m & \rightarrow NA_2 \cdots A_m \\
NA_2 \cdots A_m & \rightarrow N B_2 A_3 \cdots A_m \\
N B_2 A_3 \cdots A_m & \rightarrow N B_2 B_3 A_4 \cdots A_m \\
& \vdots \\
N B_2 \cdots B_{m-1} A_m & \rightarrow N B_2 \cdots B_{m-1} B_m \cdots B_n \\
N B_2 \cdots B_n & \rightarrow B_1 B_2 \cdots B_n
\end{align*}
\]

(note that, due to the presence of \( N \), these productions will not "interfere" with other ones)

Observe that all such productions are of the form

\[ \gamma_1 A \gamma_2 \rightarrow \gamma_1 B \gamma_2 \text{ with } \gamma_1, \gamma_2 \in \mathcal{V}^*, \ \alpha \in \mathcal{V}_N, \ \beta \in \mathcal{V}^+ \]

4) For each \( \alpha \in \mathcal{V}_N \), add the production

\[ N_\epsilon \rightarrow \alpha \]  

(where \( N_\epsilon \) is the new non-terminal associated to \( \epsilon \))

It is not difficult to see that \( \mathcal{L}(G_1) = \mathcal{L}(G_\epsilon) \)

(the proof is by induction on the length of the derivation of \( \epsilon \) string \( w \in \mathcal{L}(G_1) \) : (resp., \( \mathcal{L}(G_\epsilon) \))

\[ END \text{ OPTIONAL] \]
Definition: grammar of type 3, or regular, or right-linear productions have the form \( A \rightarrow \gamma \) with \( A \in V_N \), 
\( \gamma \in V_T \cup (V_T \circ V_N) \)
(i.e., \( A \rightarrow \alpha B \) or \( A \rightarrow \alpha \), with \( A, B \in V_N \), \( \alpha \in V_T \))

A language generated by a grammar of type 3 is called of type 3 or regular.

Example: \( \{ a^n b^n | n \geq 0 \} \) is of type 3, since it is generated by the grammar
\[
S \rightarrow aS \\
S \rightarrow b
\]

Note: a grammar of type 3 is called linear, because on the right-hand side of a production there is at most one non-terminal. It is called right-linear because the non-terminal is on the right of the terminal.

Exercise: E5.3: Show that grammars of type 3 generate the class of regular languages that do not contain \( \epsilon \).

Hint: given \( G = (V_N, V_T, P, S) \), construct an NFA
\[
A_\epsilon = (V_N \cup \{ F \}, V_T, \delta, S, \{ F \}) \text{ with }
\]
\( B \in \delta(A, \alpha) \) iff \( A \rightarrow \alpha B \) and \( F \in \delta(A, \alpha) \) iff \( A \rightarrow \alpha \)

show by induction on \( |w| \) that \( w \in L(A_\epsilon) \) iff \( w \in L(G) \).

Conversely, given an NFA \( A \), construct a grammar \( G_A \) by having again non-terminals correspond to states of \( A \).
Note on $E$-productions (for grammars of type 1, 2, 3)

As we have defined them, grammars of type 1 (req. 2, 3) cannot generate the empty string $E$.

We could extend the definition by allowing also the generation of $E$:

- if the start symbol $S$ does not appear on the right-hand side of productions, we allow also for a production $S \to E$ \textit{(E-production)}

- if the start symbol $S$ appears on the right-hand side of productions, we introduce a new non-terminal $S_{\text{new}}$, make it the new start symbol, add a production $S_{\text{new}} \to S$, and allow for $S_{\text{new}} \to E$.

Hence, an $E$-production used just to generate $E$ is harmless.

Note that, allowing for $E$-productions for every non-terminal is not that harmless.

\textbf{Exercise:} E5.4. Show that, for every language $L$ of type 0 there is a grammar of type 1 extended with $E$-productions on arbitrary non-terminals that generates $L$.

Hint: introduce a new non-terminal $N_{\text{e}}$ that is eliminated through an $E$-production $N_{\text{e}} \to E$, and use $N_{\text{e}}$ to make the right-hand side of productions as long as the left-hand side.

\textbf{END OPTIONAL}
A context-free grammar (CFG) is a CFG, the productions have the form $A \rightarrow \beta$ with $A \in V_N$, $\beta \in V^*$. (note: we allow for $\epsilon$-productions)

Example: CFG for arithmetic expressions over variable $i$

$G = (\{E, T, F\}, \{+, \cdot, (, )\}, \{P, E\}, P)$, where $P$ is:

- $E \rightarrow T \mid E + T$  \hspace{1cm} E ... expression
- $T \rightarrow F \mid T \cdot F$  \hspace{1cm} T ... term
- $F \rightarrow i \mid (E)$  \hspace{1cm} F ... factor

This grammar generates, e.g., $i + i \cdot i$

$E \rightarrow E + T \rightarrow T + T \rightarrow E + T \rightarrow i + T \rightarrow$

$\rightarrow i + T \cdot F \rightarrow i + i \cdot F \rightarrow i + i \cdot i$

We can also represent a derivation of a string by a CFG by means of a tree, called parse-tree:

In a tree whose nodes are labeled by elements of $V \cup \mathcal{E}$ satisfying:

1) each interior node is labeled by a non-terminal
2) each leaf is labeled by a non-terminal, a terminal, or $E$. If it is labeled by $E$, then it is the only child of its parent
3) if an interior node is labeled $A$, and its children from left to right are labeled $X_1, X_2, \ldots, X_k$, then there is a production $A \rightarrow X_1 X_2 \ldots X_k$ in $P$.

Example: parse tree for $i + i \cdot i$
We call A-tree a subtree of the parse tree rooted at non-terminal A.

Yield (or frontier) of a tree:

is the sequence of labels of the leaves from left to right.

Example:

```
E
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
```

Theorem: $\alpha \in V^+$ is the yield of an A-tree $\Rightarrow A \Rightarrow^* \alpha$

Proof: by induction on the height of the tree

(see textbook)

8/1/2008

Note: a parse tree does not specify a unique way to derive $\alpha$ from A. (the order in which non-terminals are expanded is not specified).

The parse tree specifies, however, which rule is applied for each non-terminal.

Specific derivation orders:

- Leftmost derivation: obtained by traversing the tree depth-first, by first going to the left subtree, and then to the right one.
  
  e.g.: $E \Rightarrow^* E + T \Rightarrow^* I + T \Rightarrow^* E + T \Rightarrow^* I + T \Rightarrow \cdots$

- Rightmost derivation: defined similarly. $E \Rightarrow^* E + T \Rightarrow E + T \Rightarrow \cdots$
Theorem: the following are all equivalent statements for a CFG $G = (V, T, P, S)$ and a string $w \in T^*$

1) $w \in L(G)$ (or $S \Rightarrow^* w$)
2) $S \xrightarrow{km} w$
3) $S \xrightarrow{rm} w$
4) There exists an S-tree with yield $w$.

Proof: the equivalence of (1) and (4) follows from the previous theorem. The other equivalences are obvious.

Thus, we could always use $km$-derivation as a canonical way to derive any $w \in L(G)$; i.e. as a canonical way to interpret a parse tree for $w$.

Ambiguous grammars:

2) $w \in L(G)$ could have two distinct parse trees, and hence two distinct $km$-derivations.

Example: another grammar for arithmetic expressions

$$E \rightarrow i \mid (E) \mid E + E \mid E \cdot E$$

$$w = i + i \cdot i$$

These parse trees correspond to two different $km$-derivations, and also to two ways of interpreting $w$. 
Definition: A CFG G is ambiguous if for some w ∈ L(G) there exist two distinct parse trees.

Ambiguity has to be avoided in compilers, since it corresponds to different ways of interpreting a string.

Sometimes grammars can be redesigned to remove ambiguity. (e.g., for arithmetic expressions)

This is not always possible:

Definition: A CFG language is (inherently) ambiguous if all its grammars are ambiguous

Example: \[ L = \{a^m b^n c^n d^n \mid m, n \geq 1\} \cup \{a^m b^n c^n d^n \mid m, n \geq 1\} \]

\[ L \text{ is } CF \quad \text{(show for exercise)} \]

Consider strings of the form abbc bcd. We cannot tell whether they come from first or second types of strings in L, and any CFG must allow for both possibilities.
Properties of context-free languages

We will study
1) Normal forms for CFGs (useful for proving properties of CFLs)
2) Expressive power => pumping lemma for CFLs
3) Closure and decision properties

Normal forms for CFGs

We look at how to simplify CFGs, while preserving the generated language.
- gain efficiency in parsing
- simplify proving properties

9) Eliminate useless symbols:

We say that \( X \in V \) is useful if
\[
S \Rightarrow^* \alpha K \beta \Rightarrow^* w \quad \text{with} \quad w \in V_T^*, \alpha, \beta \in V^*
\]

Thus, a symbol is useless (not useful) if it does not participate in any derivation of strings of the language.
\Rightarrow \text{can be eliminated}

Definition: \( X \in V \) is generating if \( X \Rightarrow^* w \), for \( w \in V_T^* \)
\( X \in V \) is reachable if \( S \Rightarrow^* \alpha K \beta \), for \( \alpha, \beta \in V^* \)

Hence, \( X \) is useful if it is both generating and reachable.
We identify useless symbols by

1) eliminating non-generating symbols and all their production

2) unreachable

Note: it is important to do these two steps in the above order

Example: \[
\begin{align*}
S & \rightarrow AB | b \\
A & \rightarrow \varepsilon
\end{align*}
\]

Let us consider what happens if we do first (2) and then (1)

- eliminate unreachable symbols: all are reachable
- non-generating

we eliminate B and \( S \rightarrow AB \)

\[ \Rightarrow \text{we obtain: } S \rightarrow b \\
A \rightarrow \varepsilon \]

But, if we do it in right order:

1) eliminate non-generating symbols: B and \( S \rightarrow AB \)

2) unreachable: A and \( A \rightarrow \varepsilon \)

\[ \Rightarrow \text{we obtain: } S \rightarrow b \]

1) Eliminating non-generating symbols:

Recursive algorithm to construct the set of generating symbols:

basis: mark all terminals as generating

recursive step: for each production \( A \rightarrow X_1 \ldots X_n \)
if all of \( X_1, \ldots, X_n \) are marked as generating
then mark \( A \) as generating

terminate: when no new generating symbol is found
Example: \( G_1: \) 
\[
\begin{align*}
S &\rightarrow AB \mid AC \mid CD \\
A &\rightarrow BB \\
B &\rightarrow AC \mid \varepsilon b \\
C &\rightarrow Ca \mid CC \\
D &\rightarrow Bc \mid \varepsilon d \\
\{a, b, d\} \\
\{\varepsilon, b, C, D\} \\
\{\varepsilon, a\} \\
\{\varepsilon, S\} &\Rightarrow \text{C is non-generating} \\
\Rightarrow \text{Remove C and all productions involving C}
\end{align*}
\]

2) Eliminating unreachable symbols

Recursive algorithm to construct the set of reachable symbols:

**Basis:** mark \( S \) as reachable

**Recursive step:** for each production \( A \rightarrow X_1 \ldots X_n \)

- if \( A \) is marked as reachable
  - then mark \( X_1, \ldots, X_n \) as reachable

Terminate when no new reachable symbol is found

Example: \( G_2: \) 
\[
\begin{align*}
S &\rightarrow AB \\
A &\rightarrow BB \\
B &\rightarrow \varepsilon b \\
D &\rightarrow b \mid d \\
\{S\} \\
\{S, A, B\} \\
\{S, A, B, \varepsilon, b\} &\Rightarrow \text{D, d are unreachable} \\
\Rightarrow \text{Remove D, d and all productions involving them}
\end{align*}
\]
2) Eliminate $E$-productions

$E$-production: $A \rightarrow E$ slows down parsing

**Definition:** $X \in V_N$ is **nullable** if $X \Rightarrow^*_E$

We first need to find all nullable symbols:

Recursive algorithm to construct the set of nullable symbols:

**basis:** if $P$ contains $A \rightarrow E$, then mark $A$ as nullable

**inductive step:** for each production $A \rightarrow X_1 \ldots X_n$

if all of $X_1, \ldots, X_n$ are marked as nullable

then mark $A$ as nullable

terminate when no new nullable symbol is found

**Example:** $G_1:

\[
S \rightarrow ABC \mid BCB
\]

\[
A \rightarrow aB \mid e
\]

\[
B \rightarrow CC \mid b
\]

\[
C \rightarrow S \mid \epsilon
\]

\[
\{C\}
\]

\[
\{-C, B\}
\]

\[
\{C, B, S\}
\]

Knowing the nullable symbols allows us to compensate for the elimination of $E$-transitions.

**Example:** in $G_1$, since $B$ and $C$ are nullable, we can derive

\[
S \Rightarrow^* BCB, \quad S \Rightarrow^* CB, \quad S \Rightarrow^* BC, \quad S \Rightarrow^* BB,
\]

\[
S \Rightarrow^* C, \quad S \Rightarrow^* B, \quad S \Rightarrow^* \epsilon
\]

Hence, if we eliminate $C \rightarrow \epsilon$, we have to add direct productions for the above derivations.
Algorithm to eliminate E-productions

1) Identify all nullable symbols

2) Replace each production $A \to X_1 \ldots X_k$
   by the set of all productions of the form $A \to \alpha_1 \ldots \alpha_k$
   where $\alpha_i = X_i$, if $X_i$ is not nullable
   $\alpha_i = X_i$ or E, if $X_i$ is nullable

3) If the resulting grammar contains $S \to E$, introduce a new
   start symbol $S'$ and add the productions $S' \to S(E$

4) Remove all E-productions, except possibly the one for $S'$.

Example: by applying steps (1) and (2) to $G_1$, we get

$S \to ABC | AB | AC | A | BC | BC | BB | CB | B | C | E$

$A \to aB | e$

$B \to CC | C | E | b$

$C \to S | E$

Since we have $S \to E$, we add

$S' \to S(E$

and remove the remaining E-productions.

3) Eliminate unit productions

Unit production: $A \to B$ slows down parsing

Algorithm to eliminate unit-productions

1) Remove E-productions

2) For all $A, B \in V_M$
   if $A \Rightarrow^* B$ and $B \Rightarrow \alpha$ is not unit
   then add $A \to \alpha$

3) Eliminate all unit-productions
How do we determine whether $A \Rightarrow^* B$ holds?

Since we have no $E$-productions, we have that

$A \Rightarrow^* B$ if and only if

$A \Rightarrow B_1 \Rightarrow B_2 \Rightarrow \cdots \Rightarrow B_{k-1} \Rightarrow B_k \Rightarrow B$

where all $B_i$'s are pairwise distinct, hence $k \leq |V_N|$.

(If we had a sequence where two $B_i$'s are the same, we could eliminate all steps in between, and get a new sequence where all $B_i$'s are pairwise distinct.)

Each single derivation step $B_i \Rightarrow B_{i+1}$ must correspond to a unit production $B_i \Rightarrow B_{i+1}$ of $G$.

Hence, we can detect whether $A \Rightarrow^* B$ by checking whether $B$ is reachable from $A$ in the graph of the unit productions:

- nodes: non-terminals
- edges: one edge $\text{A} \rightarrow \text{B}$ for each unit prod. $A \rightarrow B$

Example: $G_1$:

\[
\begin{align*}
S & \rightarrow A | B \\
A & \rightarrow Sa | a \\
B & \rightarrow S | b
\end{align*}
\]

Graph of unit productions:

Reachability: $S \Rightarrow^* A$, $S \Rightarrow^* B$, $B \Rightarrow^* A$

We get:

\[
\begin{align*}
S & \rightarrow A | B | Sa | a | S | b \\
A & \rightarrow Sa | a \\
B & \rightarrow S | b | A | B | Sa | a
\end{align*}
\]

Removing unit productions, we get:

\[
\begin{align*}
S & \rightarrow Sa | a | b \\
A & \rightarrow Sa | a \\
B & \rightarrow Sa | a | b
\end{align*}
\]

Note: $A$ and $B$ have become unreachable.
We have seen: removal of - useless symbols, E-prod, unit-prod.

Does the order of the steps matter?

Observation:

- removing useless - does not add productions at all
  (and therefore not E-prod. or unit-prod)
- removing E-prod: could add unit-prod
- removing unit-prod: needs removing E-prod first
- could create useless symbols
- cannot create E-prod.

⇒ The right order for removal is

1) E-productions
2) unit-productions
3) useless symbols: first non-generating
    then unreachable

Chomsky Normal Form

Definition: A CFG $G_i$ is in Chomsky Normal (CNF) if all its productions are of the form

\[ A \rightarrow a \]
\[ A \rightarrow BC \quad \text{with} \quad a \in V_T \]
\[ A, B, C \in V_N \]

Given a CFG $G_i$, we can always construct a CFG $G_c$ that is in CNF and such that $L(G_c) = L(G_i) \setminus \{ \varepsilon \}$.

Note: Since a CFG in CNF cannot generate $\varepsilon$, if $G_i$ generates $\varepsilon$,
then we cannot have that $L(G_c) = L(G_i)$. However, apart from $\varepsilon$, the two languages are equal.
Starting from $G_1$, we construct $G_2$ in several steps:

1) Eliminate $E-$productions (without introducing the new start symbol $S'$ with $S' \rightarrow SE$)

2) Eliminate unit-productions

$\Rightarrow$ all productions are of the form

$A \rightarrow a$

$A \rightarrow X_1 \ldots X_k$ with $k \geq 2$

with $A \in V_N$, $e \in V_T$, $X_1, \ldots, X_k \in V$

3) Remove non-generating symbols

4) Remove unreachable symbols

5) Remove "mixed bodies"

for each $e \in V_T$, add a new non-terminal $V_e$ and production $V_e \rightarrow e$

in each production $A \rightarrow X_1 \ldots X_k$, replace $e$ with $V_e$

$\Rightarrow$ all productions are of the form

$A \rightarrow e$

$A \rightarrow A_1 \ldots A_k$ with $k \geq 2$

with $e \in V_T$, $A_1, A_2, \ldots, A_k \in V_N$

6) "Factor" long productions

for each $A \rightarrow A_1 \ldots A_k$ with $k \geq 3$

- add new non-terminals $B_1, \ldots, B_{k-2}$

- replace $A \rightarrow A_1 \ldots A_k$

with $A \rightarrow A_1 B_1$

$B_1 \rightarrow A_2 B_2$

$\vdots$

$B_{k-2} \rightarrow A_{k-1} A_k$

The grammar we get is in CNF by construction.

It is easy to show that the language is preserved, apart possibly for the empty string $\epsilon$, which cannot be generated by a grammar in CNF.
Example: \( G, \) \[
\begin{align*}
S & \rightarrow ABB \mid ve \\
A & \rightarrow Be \mid Ve \\
B & \rightarrow v \alpha V_e B
\end{align*}
\]

Steps 1-4 nothing to do

Step 5:
\[
\begin{align*}
V_e & \rightarrow e \\
V_b & \rightarrow b \\
S & \rightarrow ABB \mid Ve V_b \\
A & \rightarrow BV_e \mid V_e V_e \\
B & \rightarrow Ve AV_e B
\end{align*}
\]

Step 6:
\[
\begin{align*}
V_e & \rightarrow e \\
V_b & \rightarrow b \\
S & \rightarrow AB, \mid Ve V_b \\
B & \rightarrow BB \\
A & \rightarrow BV_e \mid V_e V_e \\
B & \rightarrow Ve C_1 \\
C_1 & \rightarrow AC_2 \\
C_2 & \rightarrow V_e B
\end{align*}
\]

Note: If the original grammar generated \( E, \) and we want that the grammar in CNF also generates \( E, \) we can execute Step 1 by introducing the new start symbol \( S', \) and the productions: \( S' \rightarrow S \mid E \)

Step 2 will then replace the unit production \( S' \rightarrow S, \) by other productions, but none of the transformations 2-5 will introduce an additional production where \( S' \) is on the right side. Therefore, in the end, we will have a grammar in CNF, except for the production \( S' \rightarrow E. \)