Chomsky Normal Form

It's more what you'd call 'guidelines' than actual rules.
Consider the CFG $G_1$:

$$S \rightarrow aSb \mid SS \mid \varepsilon$$

And the CFG $G_2$:

$$S \rightarrow aSb \mid SX \mid \varepsilon$$
$$X \rightarrow X \mid S \mid \varepsilon$$

1. Are $G_1$ and $G_2$ equivalent? (That is, do they generate the same language?)
2. Is there an advantage of one CFG over the other?
3. Can the disadvantage of $G_2$ be eliminated?
4. How could that procedure be formalized? Chomsky Normal Form.
Chomsky Normal Form
Simplifying and bounding grammars.

A context-free grammar is in **Chomsky Normal Form** if every rule is of the form:

\[ S \rightarrow \varepsilon \]
\[ A \rightarrow BC \]
\[ A \rightarrow a \]

Where \( a \) is any terminal and \( A, B, \) and \( C \) are any variables, except that \( B \) and \( C \) may not be the start variable.

In this example, \( S \) is the start variable.

**Implied by the definition:**

- In \( A \rightarrow BC \), only two nonterminals are allowed in the RHS.
- In \( A \rightarrow a \), only one terminal is allowed in the RHS.
- In \( S \rightarrow \varepsilon \), only the start state may go to \( \varepsilon \).
- \( S \) may not appear on the RHS.
Theorem: Any CFG to CNF

Idea: Make conversion in stages, systematically changing or removing rules that do not match the requirements.

Stages:
1. Add new start $S_0$ and rule $S_0 \rightarrow S$, where $S$ was the original start variable. (Guarantees $S_0$ isn’t on RHS.)

2. Eliminate $\epsilon$-rules. For all $A \rightarrow \epsilon$, where $A$ is not $S_0$, then remove the rule and for each occurrence of $A$ in any RHS, add a new rule with the occurrence deleted. (e.g., for every rule $R \rightarrow uAv$, we would add rule $R \rightarrow uv$.)

3. Remove unit rules. For all rules of the form $A \rightarrow B$, remove the rule and for all $B \rightarrow u$, add the rule $A \rightarrow u$ unless this was a unit rule previously removed ($u$ is a string of variables and terminals).

(cont.) Replacement happens for each occurrence of $A$. so $R \rightarrow uAvAw$ would add $R \rightarrow uvAw | uAvw | uvw$.

If $R \rightarrow A$ exists, then add $R \rightarrow \epsilon$ (unless this rule has already been removed). Repeat until all $\epsilon$-rules are removed.

Repeat until all unit rules removed.
Theorem: Any CFG to CNF (continued)

4. Convert all remaining rules into the proper form.
   Reminder: Proper form is
   \[ A \rightarrow BC \]
   \[ A \rightarrow a \]

Replace each rule \( A \rightarrow u_1 u_2 \ldots u_k \), where \( k \geq 3 \) and each \( u_i \) is a variable or terminal, with rules:

\[ A \rightarrow u_1 A_1 \]
\[ A_1 \rightarrow u_2 A_2 \]
\[ A_2 \rightarrow u_3 A_3 \]
\[ \vdots \]
\[ A_{k-2} \rightarrow u_{k-1} u_k \]

4. (cont.) \( A_i \)'s are new variables.

Replace any terminal \( u_i \) in the preceding rules with the new variable \( U_i \) and add the rule \( U_i \rightarrow u_i \).

Repeat until all rules are in the proper form.
Example: $G_6$ to CNF: Step 1

Before:

\[
\begin{align*}
S & \rightarrow ASA \mid aB \\
A & \rightarrow B \mid S \\
B & \rightarrow b \mid \varepsilon
\end{align*}
\]

After (new starting state):

\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow ASA \mid aB \\
A & \rightarrow B \mid S \\
B & \rightarrow b \mid \varepsilon
\end{align*}
\]
Example: $G_6$ to CNF: Step 2

Previous:
\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow ASA | aB \\
A & \rightarrow B | S \\
B & \rightarrow b | \varepsilon
\end{align*}
\]

After (remove $\varepsilon$-rules, $B \rightarrow \varepsilon$):
\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow ASA | aB | a \\
A & \rightarrow B | S | \varepsilon \\
B & \rightarrow b | \varepsilon
\end{align*}
\]

After (remove $\varepsilon$-rules, $A \rightarrow \varepsilon$):
\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow ASA | aB | a | SA | AS | S \\
A & \rightarrow B | S | \varepsilon \\
B & \rightarrow b
\end{align*}
\]
Example: $G_6$ to CNF: Step 3

Previous:

\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow ASA \mid aB \mid a \mid SA \mid AS \mid S \\
A & \rightarrow B \mid S \\
B & \rightarrow b
\end{align*}
\]

After (remove unit rules, $S \rightarrow S$):

\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
A & \rightarrow B \mid S \\
B & \rightarrow b
\end{align*}
\]

After (remove unit rules, $S_0 \rightarrow S$):

\[
\begin{align*}
S_0 & \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
S & \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
A & \rightarrow B \mid S \\
B & \rightarrow b
\end{align*}
\]
Example: $G_6$ to CNF: Step 3 (cont.)

Previous:
\[
S_0 \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
S \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
A \rightarrow B \mid S \\
B \rightarrow b
\]

After (remove unit rules, $A \rightarrow B$):
\[
S_0 \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
S \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
A \rightarrow B \mid S \mid b \\
B \rightarrow b
\]

After (remove unit rules, $A \rightarrow S$):
\[
S_0 \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
S \rightarrow ASA \mid aB \mid a \mid SA \mid AS \\
A \rightarrow S \mid b \mid ASA \mid aB \mid a \mid SA \mid AS \\
B \rightarrow b
\]
Example: $G_6$ to CNF: Step 4

Previous:

$S_0 \rightarrow ASA | aB | a | SA | AS$
$S \rightarrow ASA | aB | a | SA | AS$
$A \rightarrow b | ASA | aB | a | SA | AS$
$B \rightarrow b$

Previous:

$S_0 \rightarrow AA_1 | aB | a | SA | AS$
$S \rightarrow AA_1 | aB | a | SA | AS$
$A \rightarrow b | AA_1 | UB | a | SA | AS$
$A_1 \rightarrow SA$
$B \rightarrow b$

After (Eliminate $ASA$ on RHS):

$S_0 \rightarrow ASA | AA_1 | aB | a | SA | AS$
$S \rightarrow ASA | AA_1 | aB | a | SA | AS$
$A \rightarrow b | ASA | AA_1 | aB | a | SA | AS$
$A_1 \rightarrow SA$
$B \rightarrow b$

After (Eliminate $aB$ on RHS):

$S_0 \rightarrow AA_1 | aB | UB | a | SA | AS$
$S \rightarrow AA_1 | aB | UB | a | SA | AS$
$A \rightarrow b | AA_1 | UB | a | SA | AS$
$A_1 \rightarrow SA$
$U \rightarrow a$
$B \rightarrow b$
Example: $G_6$ to CNF

Before:

$S \rightarrow ASA \mid aB$
$A \rightarrow B \mid S$
$B \rightarrow b \mid \varepsilon$

After (Final):

$S_0 \rightarrow AA_1 \mid UB \mid a \mid SA \mid AS$
$S \rightarrow AA_1 \mid UB \mid a \mid SA \mid AS$
$A \rightarrow b \mid AA_1 \mid UB \mid a \mid SA \mid AS$
$A_1 \rightarrow SA$
$U \rightarrow a$
$B \rightarrow b$

Important: Upper size bound is the square of the original Grammar size.
Convert this to CNF:

\[ S \rightarrow \emptyset S_1 | \varepsilon \]

1. Add new start \( S_0 \)
2. Eliminate \( \varepsilon \)-rules.
3. Remove unit rules.
4. Convert all remaining rules into the proper form.
Convert $S \rightarrow \emptyset S1 \mid \varepsilon$ to CNF

Step 1:
\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow \emptyset S1 \mid \varepsilon
\end{align*}
\]

Step 2:
\[
\begin{align*}
S_0 & \rightarrow S \mid \varepsilon \\
S & \rightarrow \emptyset S1 \mid 01
\end{align*}
\]

Step 3:
\[
\begin{align*}
S_0 & \rightarrow \emptyset S1 \mid 01 \mid \varepsilon \\
S & \rightarrow \emptyset S1 \mid 01
\end{align*}
\]

Step 4, part 1:
\[
\begin{align*}
S_0 & \rightarrow \emptyset A \mid 01 \mid \varepsilon \\
S & \rightarrow \emptyset A \mid 01 \\
A & \rightarrow S1
\end{align*}
\]

Step 4, part 2:
\[
\begin{align*}
S_0 & \rightarrow BA \mid BC \mid \varepsilon \\
S & \rightarrow BA \mid BC \\
A & \rightarrow SC \\
B & \rightarrow 0 \\
C & \rightarrow 1
\end{align*}
\]
CNF in Retrospect

1. The book does **NOT** give a good proof. Why is it a bad example of a proof?

2. Wikipedia gives a better overview of the proof. 

3. What benefit does CNF have?
   Guaranteed maximum final grammar size.
   Guaranteed maximum derivation length.

4. Why do we need CNF?
   Proofs! Algorithms! Big-O Analysis!

5. BUT!!!! Does CNF eliminate ambiguity?
   **No.**
Chomsky Language Hierarchies

Formal Grammar Classifications
Each class is a subset of the class above it.

https://commons.wikimedia.org/wiki/File:Chomsky-hierarchy.svg