Simplification of CFGs

Eliminating Unit Productions

- A *unit production* is a production of the form $A \rightarrow B$, where both A and B are variables. So, E → a is not a unit production.
- The technique to eliminate unit productions involves finding all those pairs of variables A and B such that $A \Rightarrow^* B$ using a sequence of unit productions only.
- It is possible for $A \Rightarrow^* B$ to be true even though no unit productions are involved. For instance, $A \rightarrow BC$ and $C \rightarrow \varepsilon$.
- Following algorithm is for the inductive construction of the pairs (A, B) such that $A \Rightarrow^* B$ using only unit productions. Such a pair is called *unit pair*.

  Basis: (A, A) is a unit pair for any variable A. That is $A \Rightarrow^* A$ by zero steps.

  Induction: Suppose it is determined that (A, B) is a unit pair, and $B \rightarrow C$ is a production, where C is a variable. Then (A, C) is a unit pair.

Example:
Consider the grammar:

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

Unit pairs according to Basis: (E, E), (T, T), (F, F) and (I, I).
For the inductive step, following inferences can be made:

1. (E, E) and the production $E \rightarrow T$ gives us unit pair (E, T).
2. (E, T) and the production $T \rightarrow F$ gives us unit pair (E, F).
3. (E, F) and the production $F \rightarrow I$ gives us unit pair (E, I).
4. (T, T) and the production $T \rightarrow F$ gives us unit pair (T, F).
5. (T, F) and the production $F \rightarrow I$ gives us unit pair (T, I).
6. (F, F) and the production $F \rightarrow I$ gives us unit pair (F, I).

The following table summarizes the final part of the procedure, where new set of productions are created by using first member of a pair as the head and all the non-unit bodies for the second member of the pair as the production bodies.

Pair	Productions
(E, E)	$E \rightarrow E + T$
(E, T)	$E \rightarrow T * F$
(E, F)	$E \rightarrow (E)$
(E, I)	$E \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(T, T)	$T \rightarrow T * F$
(T, F)	$T \rightarrow (E)$
(T, I)	$T \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(F, F)	$F \rightarrow (E)$
(F, I)	$F \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(I, I)	$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

The final grammar:

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

The goal is to convert any CFG into an equivalent CFG that has no useless symbols, ε-productions, or unit productions. Some care must be taken in the order of application of the constructions. A safe order is:

1. Eliminate ε-productions.
2. Eliminate unit productions.
3. Eliminate useless symbols.

We must order the three steps above as shown, or the result might still have some of the features we thought we were eliminating.

Chomsky Normal Form

A grammar G is said to be in *Chomsky Normal Form (CNF)* if it has no ε-productions, unit productions or useless symbols, and all the productions of G are in one of two simple forms, either:

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

Every production of such a grammar is either of the form $A \rightarrow a$, which is already in a form allowed by CNF, or it has a body of length 2 or more. The tasks are to:

a) Arrange that all bodies of length 2 or more consist only of variables.

b) Break bodies of length 3 or more into a cascade of productions, each with a body consisting of two variables.

Example:

Let's convert the following grammar to CNF. Already in Chomsky Normal Form

$$\begin{array}{l}
E \rightarrow E + T \mid T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1 \\
T \rightarrow T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1 \\
F \rightarrow (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1 \\
I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1
\end{array}$$

There are eight terminals, *a, b, 0, 1, +, *, (,* and *),* each of which appears in a body that is not a single terminal. Thus, we must introduce eight new variables, corresponding to these terminals, and eight productions in which the new variable is replaced by its terminal. Using the obvious terminals as the new variables, we introduce:

$$\begin{array}{llll}
A \rightarrow a & B \rightarrow b & Z \rightarrow 0 & O \rightarrow 1 \\
P \rightarrow + & M \rightarrow * & L \rightarrow (& R \rightarrow)
\end{array}$$

And we get the grammar:

$$\begin{array}{lll}
E \rightarrow EPT \mid TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO & A \rightarrow a & B \rightarrow b \\
T \rightarrow TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO & Z \rightarrow 0 & O \rightarrow 1 \\
F \rightarrow LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO & P \rightarrow + & M \rightarrow * \\
I \rightarrow a \mid b \mid IA \mid IB \mid IZ \mid IO & L \rightarrow (& R \rightarrow)
\end{array}$$

Here, all bodies either a single terminal or several variables. So, all productions are in Chomsky Normal Form except for those with the bodies of length 3: *EPT*, *TMF* and *LEF*. These bodies are needed to be broken into a cascade of productions, each with a body consisting of two variables. For EPT:

$\boldsymbol{E \rightarrow EPT}$ **will be replaced with** $\boldsymbol{E \rightarrow EC_1}$ ***and*** $\boldsymbol{C_1 \rightarrow PT}$.

Breaking other two bodies, we get the final grammar what will be in Chomsky Normal Form.

$$\begin{array}{llll}
E \rightarrow EC_1 \mid TC_2 \mid LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO & A \rightarrow a & B \rightarrow b & C_1 \rightarrow PT \\
T \rightarrow TC_2 \mid LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO & Z \rightarrow 0 & O \rightarrow 1 & C_2 \rightarrow MF \\
F \rightarrow LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO & P \rightarrow + & M \rightarrow * & C_3 \rightarrow ER \\
I \rightarrow a \mid b \mid IA \mid IB \mid IZ \mid IO & L \rightarrow (& R \rightarrow) &
\end{array}$$

So, the productions have either a terminal or two variables only in the bodies.

E > EPTZ will be replaced with E > EC1 and C1 > PTZ

C1 > PTZ then C2 > PC3 and C3 > TZ

E > EPTZ then E > EC1, C1 > PC2, C2 > TZ