Chomsky Normal Form (CNF)

A CFG G is said to be in CNF if it has no useless symbols and every production is of the form A -> B C or A -> a.

Given any CFL L s.t. L – {ε} is nonempty we can construct a CFG G in CNF such that L(G) =L – {ε}.

Method of constructing such a G :

1) Start with any CFG G_1 such that $L(G_1) = L$.
2) Simplify G_1 to get G_2 such that $L(G_2) = L – \{\varepsilon\}$ and every production of G_2 is of the form $A \to X_1 X_2 \ldots\ldots X_k$ $k > 1$ or A -> a. If any $X_i = b$ then introduce a variable B_i and a production $B_i \to b$. Then we get an equivalent CFG G_3 where every production is of the form $A \to B_1 B_2 \ldots\ldots B_k$ $k > 1$ or A -> a.
3) If $k > 2$ for any production $A \to B_1 B_2 \ldots\ldots B_k$ replace this by $A \to B_1 B_1'$, $B_1' \to B_2 B_2'$, …….. , $B_{k-2}' \to B_{k-1} B_k$. This way we get a Grammar G_4 which is in CNF and $L(G_4) = L(G_1) – \{\varepsilon\} = L – \{\varepsilon\}$.

Example 1 : S -> a A a | b B b | ε

A -> C | a

B -> C | b

C -> C D | ε

D -> A | B | ε

First we have to simplify this which was given as a Homework in the last lesson.
The result is : S -> a A a | b B b | a a | b b

A -> a | C D | a b | b

B -> b | C D | a b | a

C -> C D | a b | a | b

D -> C D | a b | a | b

Converting this to CNF is now easy. For the terminals a and b in the body of productions with more than one symbol, we introduce variables U and V and productions U -> a and V -> b getting the Grammar

S -> UAU | VBV | UU |VV

A -> a | CD | UV | b

B -> b | CD | UV| a

C -> CD | UV | a | b

D -> CD | UV | a | b

U -> a

V -> b

After that we merely have to take care of the productions S -> U A U | V B V. These yield S -> A_1 U | B_1 V , A_1 -> U A and B_1 ->V B. The final Grammar in CNF is

S -> A_1 U | B_1 V | U U | V V

A_1 -> U A

B_1 -> V B

A -> a | C D | U V | b

B -> b | C D | U V | a

C -> C D | U V | a | b

D -> C D | U V | a | b

U -> a , V -> b

Example 2 : S -> A B C | B a B

A -> a A |B a C | a a a

B -> b B b | a | D

C -> C A | A C

D -> ε

This has to be first simplified. This was also given as a Homework. The result is

S -> B a B | B a | a B | a

B –> b B b | b b | a

For the terminals a and b in the bodies of productions of length > 1 we introduce the variables U, V and the productions U -> a and V -> b getting the Grammar

S -> B U B | B U | U B | a

B -> V B V | V V | a, U -> a, V -> b

Proceeding in the standard way now for bodies of certain productions having > 2 variables we get the final equivalent Grammar in CNF

S -> B B_1 | B U | U B | a

B_1-> U B

B -> V B_2 | V V | a

B_2-> B V

U -> a

V -> b