7.1 Chomsky normal form. A CFG $G = (V, \Sigma, R, S)$ is in Chomsky normal form if all rules have one of the following three forms:

$$S \rightarrow \varepsilon$$
$$A \rightarrow \sigma \ \text{ for } \sigma \in \Sigma$$
$$A \rightarrow BC \ \text{ for } B, C \in V$$

For every CFG, we can construct an equivalent CFG in Chomsky normal form. The construction[1] has the following stages:

1. We add a new start variable S_0 and the rule $S_0 \rightarrow S$, where S was the original start variable.

2. We eliminate all ε-*rules* of the form $A \rightarrow \varepsilon$, where A is not the start variable. That is, if $R \rightarrow uAv$ is a rule in which u and v are strings of variables and terminals, we add rule $R \rightarrow uv$. We do so for *each occurence* of an A on the right-hand side of a rule.

3. We eliminate all *unit rules* of the form $A \rightarrow B$, where B is a variable. In other words, whenever a rule $B \rightarrow u$ appears, we add the rule $A \rightarrow u$ unless this was a unit rule previously removed.

4. We finally convert all remaining rules into the proper form. We replace each rule $A \rightarrow u_1 u_2 \ldots u_k$, where $k \geq 3$ and each u_i is a variable or a terminal symbol, with the rules $A \rightarrow u_1 A_1$, $A_1 \rightarrow u_2 A_2$, ..., and $A_{k-2} \rightarrow u_{k-1} u_k$. The A_i's are new variables. If $k = 2$, we replace any terminal u_i in the preceding rule(s) with the new variable U_i and add the rule $U_i \rightarrow u_i$.

   The construction given above can be performed in quadratic time.

7.2 Pumping lemma. If the CFG G is in Chomsky normal form, then all parse trees are binary trees. Hence, given any word $w \in L(G)$ with at least $2^{|V|+1}$ letters, the parse tree of w has to have at least height $|V|+1$. In other words, some branch of the parse tree of w must be a sequence of at least $|V|+1$ nonterminals, which implies that some nonterminal must occur on the branch more than once. This leads to the following pumping lemma.

> **Lemma.**
> For every context-free language L,
> there exists a number $p \geq 1$ such that
> for every word $w \in L$ with at least p letters
> there exist x, y, z, u, v with $w = xyzuv$ and $|yu| > 0$ and $|yzu| \leq p$ such that
> for every number $i \geq 0$
> $xy^i z u^i v \in L$.

We call p the pumping number of L, and $xyzuv$ the pumping decomposition of w. Suppose that we want to prove that a language L is not context-free. We can do this by showing that the pumping lemma does not hold for L; that is, we prove the negation of the pumping lemma:

> for all numbers $p \geq 1$
> there exists a word $w \in L$ with at least p letters such that
> for all x, y, z, u, v with $w = xyzuv$ and $|yu| > 0$ and $|yzu| \leq p$
> there exists a number $i \geq 0$ such that
> $xy^i z u^i v \notin L$.

---

[1]illustrated by an example on pages 110–111 of M. Sipser's book

Note that the alternations of "for all" and "there exists" are the same as in the pumping lemma for the regular languages. Hence, context-free pumping proofs have the same shape as regular pumping proofs. We have to consider all possibilities for the pumping number p, and all possibilities for the pumping decomposition x, y, z, u, v (by case analysis), but we are free to choose a single word w and a single iteration number i. Choosing a suitable w is usually the crux of the proof; for i, we can typically choose $i = 0$ or $i = 2$.

7.3 Example. We prove that the language

$$L_5 = \{0^n 1^n 2^n \mid n \geq 0\}$$

is not context-free. We show that the pumping lemma does not hold for L_5. Consider any pumping number $p \geq 1$. Choose $w = 0^p 1^p 2^p$. Consider any pumping decomposition $w = xyzuv$; all we know about $xyzuv$ is that $|yu| > 0$ and $|yzu| \leq p$. Since $|yzu| \leq p$, there are three possibilities:

(a) yzu contains no 0s. (b) yzu contains no 1s. (c) yzu contains no 2s.

Choose $i = 2$. We need to show that $xy^2 zu^2 v$ is not in L_5. Since $|yu| > 0$,

in case (a), $xy^2 zu^2 v$ contains either more 1s than 0s, or more 2s than 0s;
in case (b), $xy^2 zu^2 v$ contains either more 0s than 1s, or more 2s than 1s;
in case (c), $xy^2 zu^2 v$ contains either more 0s than 2s, or more 1s than 2s. ∎

7.4 Example. The language

$$L_6 = \{ww \mid w \in \{0, 1\}^*\}$$

is not context-free. To show that the pumping lemma does not hold for L_6, choose $w = 0^p 1^p 0^p 1^p$.