CS 172: Computability and Complexity, Spring 2010 **S. A. Seshia & O. Etesami**

HW 4: Context Free Languages

Assigned: February 17, 2010 *Due: February 24, 2010*

Note: Take time to write clear and concise solutions. Confused and long-winded answers may be penalized. Consult the course webpage for course policies on collaboration.

1. (3 points)

   Define the *size* of a context-free grammar to be the total number of characters used in writing the rules of the grammar down (including variables, terminals, | and $\rightarrow$). For example, the one-rule grammar $A \rightarrow A1 \mid \varepsilon$ has size six because it uses six characters.

   Consider a grammar that generates *only* the string "manamana banana" and no other strings. Here the set of terminals is the set of small letters in the English alphabet and the whitespace character (denoted explicitly here by $\sqcup$), i.e., it is the set {a, b, c, ...,z, $\sqcup$} The smallest context-free grammar that generates only this string has size *sixteen*. Write the rules for this grammar.

2. (9 points)

   Let $\Sigma = \{0, 1, \#\}$. Design a context-free grammar that generates the language $L = \{x\#y : x, y \in \{0, 1\}^*, x \neq y\}$. Explain why your grammar is correct.

   Also, give the state-diagram of a PDA that generates L. (No need to prove that your PDA is correct.)

3. (12 points) Let G be a CFG in Chomsky Normal Form.

   (a) (4 points) Prove that, if w is a string in $L(G)$ of length n where $n \geq 1$, then any derivation of w in G takes exactly $2n - 1$ steps.

   (b) (8 points) Suppose G contains less than m variables. Prove that, if G generates a string with a derivation having at least 2^m steps, $L(G)$ is infinite.

   [Hint: Use part (a) and the statement and proof idea of the CFL pumping lemma.]

4. (6 points) Recall the discussion we had in class about families of computer viruses that can be characterized as languages. A computer program is viewed as a string of instructions drawn from a finite alphabet Σ. A language characterizes a subset (family) of virus programs.

   Suppose you work at anti-virus software company ImmuneSystems. You have already characterized the members of two virus families: one family is a *context-free language* A and the second is a *regular language* B. However, to foil your anti-virus software, the virus creators

have released a new "mutation" tool that generates a new family of programs defined as the following language:

$$L = \{w \mid \exists x \, s.t. \, wx \in A \text{ and } x \in B\}$$

You are tasked with updating your software to recognize the new virus language L. However, your co-worker Ignoramus complains that the detection problem is "too hard" because there is no context-free grammar that can recognize L.

Prove Ignoramus wrong; i.e., show that the set of new viruses *is* a context-free language.

[Hint: Show how to construct a PDA or CFG generating the set of new viruses from A and B. You can use the equivalence of PDAs and CFGs if needed.]