

Assignment 4: CFG grammars and languages

(200 points)

You are required to solve the following problems:

1. (20 points) Recall the CFG $G_4 = (\{E, T, F\}, \{+, x, (,), a\}, R, E)$ where R is the set of productions:

$$\begin{aligned} E &\longrightarrow E + T|T \\ T &\longrightarrow T * F|F \\ F &\longrightarrow (E)|a \end{aligned}$$

Give the leftmost and the rightmost derivations for each of the following strings:

- (a) a

Solution: $E \Rightarrow T \Rightarrow a$

Note: the leftmost and rightmost derivations are the same

- (b) a + a

Solution:

Leftmost derivation: $E \Rightarrow E + T \Rightarrow T + T \Rightarrow F + T \Rightarrow a + T \Rightarrow a + F \Rightarrow a + a$

Rightmost derivation: $E \Rightarrow E + T \Rightarrow E + F \Rightarrow E + a \Rightarrow T + a \Rightarrow F + a \Rightarrow a + a$

- (c) a+a+a

Solution:

Leftmost derivation: $E \Rightarrow E + T \Rightarrow E + T + T \Rightarrow T + T + T \Rightarrow F + T + T \Rightarrow a + T + T \Rightarrow a + F + T \Rightarrow a + a + T \Rightarrow a + a + F \Rightarrow a + a + a$

Rightmost derivation: $E \Rightarrow E + T \Rightarrow E + F \Rightarrow E + a \Rightarrow E + T + a \Rightarrow E + F + a \Rightarrow E + a + a \Rightarrow T + a + a \Rightarrow F + a + a \Rightarrow a + a + a$

- (d) ((a))

Solution: $E \Rightarrow T \Rightarrow F \Rightarrow (E) \Rightarrow (T) \Rightarrow (F) \Rightarrow ((E)) \Rightarrow ((T)) \Rightarrow ((F)) \Rightarrow ((a))$

Note: the leftmost and rightmost derivations are the same

You get 5 points for each correct leftmost and rightmost derivations.

2. (a) (10 points) Use the languages $A = \{a^m b^n c^n | m, n \geq 0\}$, $B = \{a^n b^n c^m | m, n \geq 0\}$ and $C = \{a^n b^n c^n | n \geq 0\}$ to show that class of context-free languages is not closed under intersection.

Solution sketch: A is recognized by the CFG $G_A = (\{S, R, T\}, \{a, b, c\}, \{S \rightarrow RT, R \rightarrow aR|\epsilon, T \rightarrow bTc|\epsilon\}, S)$, and B is recognized by the CFG $G_B = (\{S, T, R\}, \{a, b, c\}, \{S \rightarrow TR, T \rightarrow aTb|\epsilon, R \rightarrow cR|\epsilon\}, S)$.

Both A and B are context-free languages and $A \cap B = C$. We know from Example 2.36 that the language C is not context free. We have found two CFGs whose intersection is not context free. Therefore the class of context-free languages is not closed under intersection.

- (b) (10 points) Use part (1) above and DeMorgan's law to show that the class of context-free languages is not closed under complementation.

Solution sketch: First, the context-free languages are closed under the union operation. Let $G_1 = (V_1, \Sigma, R_1, S_1)$ and $G_2 = (V_2, \Sigma, R_2, S_2)$ be two arbitrary context-free grammars. We construct a CFG G that recognizes their union. Formally, $G = (V, \Sigma, R, S)$ where:

- i. $V = V_1 \cup V_2$
- ii. $R = R_1 \cup R_2 \cup \{S \rightarrow S_1, S \rightarrow S_2\}$

Next, we show that CFGs are not closed under complementation. Assume, for a contradiction, that CFGs are closed under complementation. Then, if G_1 and G_2 are context-free grammars, it would follow that $\overline{L(G_1)}$ and $\overline{L(G_2)}$ are context free, so $\overline{L(G_1) \cup L(G_2)}$ is also context free. That, by our assumption, implies that $\overline{\overline{L(G_1) \cup L(G_2)}}$ is context free. But by DeMorgan's laws, $\overline{\overline{L(G_1) \cup L(G_2)}} = L(G_1) \cap L(G_2)$. However, if G_1 and G_2 are chosen as in part (a), $\overline{L(G_1) \cup L(G_2)}$ is not context free. This contradiction shows that the context-free languages are not closed under complementation.

3. (20 points) Let $E = \{a^i b^j \mid i \neq j \text{ and } 2i \neq j\}$. Show that E is a context-free language.

Solution sketch: According to the fact that $i \neq j$ and $2i \neq j$, there could be three cases for the relationship of i and j :

- (a) $i < j/2$
- (b) $j/2 < i < j$
- (c) $i > j$

We can construct a CFG $G = (\{S, X, Y, Z\}, \{a, b\}, R, S)$ that recognizes E . The rule $S \rightarrow X|Y|Z$ gives the three possible cases, in which X , Y and Z generate cases (a), (b) and (c) above, respectively. The rules are as follows:

$$\begin{aligned} X &\rightarrow aXbb|Xb|b \\ Y &\rightarrow aYbb|aYb|aabb \\ Z &\rightarrow aZ|aZb|a \end{aligned}$$

We can easily show that X , Y and Z generate the three cases, thus the CFG G generates E .

4. (35 points) Give context-free grammars that generate the following languages. In all parts the terminal set is $\{0, 1\}$. You get 5 points for each correct context-free grammar you have constructed.

- (a) $L_1 = \{w \mid w \text{ contains at least three 1-s}\}$

Solution: $G_{L_1} = (\{S, R\}, \{0, 1\}, \{S \rightarrow R1R1R1R, R \rightarrow 0R|1R|\epsilon\}, S)$

(b) $L_2 = \{w \mid w \text{ starts and ends with the same symbol}\}$

Solution: $G_{L_2} = (\{S, R\}, \{0, 1\}, \{S \rightarrow 0R0 \mid 1R1 \mid 0 \mid 1 \mid \epsilon, R \rightarrow 0R \mid 1R \mid \epsilon\}, S)$

(c) $L_3 = \{w \mid \text{the length of } w \text{ is odd}\}$

Solution: $G_{L_3} = (\{S\}, \{0, 1\}, \{S \rightarrow 0 \mid 1 \mid 00S \mid 01S \mid 10S \mid 11S\}, S)$

(d) $L_4 = \{w \mid \text{the length of } w \text{ is odd and its middle symbol is } 0\}$

Solution: $G_{L_4} = (\{S\}, \{0, 1\}, \{S \rightarrow 0 \mid 0S0 \mid 0S1 \mid 1S0 \mid 1S1\}, S)$

(e) $L_5 = \{w \mid w \text{ contains more 1-s than 0-s}\}$

Solution: $G_{L_5} = (\{S, B\}, \{0, 1\}, \{S \rightarrow B1B, B \rightarrow BB \mid 0B1 \mid 1B0 \mid 1 \mid \epsilon\}, S)$

(f) $L_6 = \{w \mid w = w^R, \text{ that is } w \text{ is a palindrome}\}$

Solution: $G_{L_6} = (\{S\}, \{0, 1\}, \{S \rightarrow 0S0 \mid 1S1 \mid 0 \mid 1 \mid \epsilon\}, S)$

(g) $L_7 = \emptyset$

Solution: $G_{L_7} = (\{S\}, \{0, 1\}, \{S \rightarrow S\}, S)$

5. (15 points) Give a context-free grammar that generates the language:

$$A = \{a^i b^j c^k \mid i, j, k \geq 0 \wedge (i = j \vee j = k)\}$$

Is your grammar ambiguous? Why or why not?

Solution sketch: $G = (V, \Sigma, R, S)$, where:

(a) $V = \{S, E_{ab}, E_{bc}, C, A\}$

(b) $\Sigma = \{a, b, c\}$

(c) $R = \{S \rightarrow E_{ab}C \mid AE_{bc}; E_{ab} \rightarrow aE_{ab}b \mid \epsilon; E_{bc} \rightarrow bE_{bc}c \mid \epsilon; C \rightarrow Cc \mid \epsilon; A \rightarrow Aa \mid \epsilon\}$

(d) S is the start symbol (the axiom) of the grammar

Note: $S \rightarrow E_{ab}C$ can generate any string that has an equal number of a -s followed by an equal number of b -s appended by a number of c -s, i.e., $a^i b^i c^k \in L(G)$. $S \rightarrow AE_{bc}$ generates any string with an equal number of b -s and c -s prepended by a -s, i.e., $a^k b^i c^i \in L(G)$.

The grammar is ambiguous because $S \xRightarrow{*} E_{ab}C \xRightarrow{*} \epsilon$ and $S \xRightarrow{*} AE_{bc} \xRightarrow{*} \epsilon$ are two distinguished leftmost derivations of ϵ . In general, any string $a^i b^j c^k$ with $i = j = k$ can be derived ambiguously in this grammar.

6. (30 points) Consider the CFG $G = (V, \Sigma, R, S)$, where $V = \{A, B\}$, $\Sigma = \{0\}$, $R = \{A \rightarrow BAB \mid B \mid \epsilon; B \rightarrow 00 \mid \epsilon\}$, and $S = A$. You are required to convert G into an equivalent CFG G' , where G' is in Chomsky normal form. Use the procedure given in Theorem 2.9.

Solution sketch: $G' = (\{S, A, B, C, D\}, \{0\}, R', S)$ where R' is

$$\begin{aligned} S &\rightarrow AB|CC|BA|BD|BB|\epsilon \\ A &\rightarrow AB|CC|BA|BD|BB \\ B &\rightarrow CC \\ C &\rightarrow 0 \\ D &\rightarrow AB \end{aligned}$$

7. (40 points) Use pumping lemma to prove that the following languages are not context-free. You get 10 points for each correct proof.

(a) $L_1 = \{0^n 1^n 0^n 1^n | n \geq 0\}$

Solution: Assume that L_1 is context free and let p be the pumping length given by the pumping lemma. We show that $s = 0^p 1^p 0^p 1^p \in L_1$ cannot be pumped. Let $s = uvxyz$. If either v or y contains more than one type of terminal symbols then $uv^2xy^2z \notin L_1$ because it does not contain the symbols in the correct order. If both v and y contain at most one type of terminal symbol then $uv^2xy^2z \notin L_1$ because it contains runs of 0-s and 1-s of unequal length. Hence, this is a contradiction and L_1 cannot be context free.

(b) $L_2 = \{0^n \# 0^{2n} \# 0^{3n} | n \geq 0\}$

Solution: Assume that L_2 is context free and let p be the pumping length given by the pumping lemma. Let $s = 0^p \# 0^{2p} \# 0^{3p} \in L_2$. We show that $s = uvxyz$ cannot be pumped. Neither v nor y can contain more than two $\#$ -s. Therefore if we divide s into three segments by $\#$ -s: 0^p , 0^{2p} and 0^{3p} , at least one of the segments is not contained within either v or y . Hence, $uvxyz \notin L_2$ because the 1:2:3 length ratio of the segments is not maintained. This is a contradiction so L_2 cannot be context free.

(c) $L_3 = \{w \# x | w \text{ is a substring of } x, w, x \in \{a, b\}^*\}$

Solution: Assume that L_3 is context free and let p be the pumping length given by the pumping lemma. Let $s = a^p b^p \# a^p b^p \in L_3$. We show that $s = uvxyz$ cannot be pumped.

Neither v nor y can contain $\#$, otherwise uv^0xy^0z does not contain $\#$ and therefore it is not in L_3 . If both v and y are nonempty and occur in the left-hand side of $\#$, then $uv^2xy^2z \notin L_3$ because it is longer in the left-hand side and thus cannot be a substring of the string in the right-hand side. Similarly, if both v and y occur in the right-hand side of $\#$, then $uv^0xy^0z \notin L_3$ because it is shorter in the right-hand side and thus the left-hand side cannot be the substring of the right-hand side. If only one of v or y is empty (both cannot be empty), treat them as if both occurred on the same side of $\#$.

The only remaining case is where both v and y are non-empty and straddle $\#$. Then v consists of \mathbf{b} -s and y consists of \mathbf{a} -s because of the pumping lemma condition $|vxy| \leq p$. Hence, $uv^2xy^2z \notin L_3$ because it contains more \mathbf{b} -s on the left-hand side of $\#$. This is a contradiction and so L_3 cannot be context free.

(d) $L_4 = \{x_1\#x_2\#\dots\#x_k \mid x_i \in \{a, b\}^* \text{ and for some } i \neq j, x_i = x_j\}$.

Solution: Assume that L_4 is context free and let p be the pumping length given by the pumping lemma. Let $s = a^p b^p \# a^p b^p \in L_4$. We show that $s = uvxyz$ cannot be pumped. Use the same reasoning as in part (c).

8. (20 points) Show that, if G is a CFG in Chomsky normal form, then for any string $w \in L(G)$ of length $n \geq 1$, exactly $2n - 1$ steps are required for any derivation of w .

Solution sketch: A CFG in Chomsky normal form has all rules of the form $A \rightarrow BC$ or $A \rightarrow a$ where A, B, C are nonterminals and a is terminal. Hence any derivation of a work w of length n needs to use $n - 1$ applications of the rules of the form $A \rightarrow BC$ and n applications of the rules of the form $A \rightarrow a$, and $n + n - 1 = 2n - 1$.

9. (40 points) Let $D = \{xy \mid x, y \in \{0, 1\}^* \wedge |x| = |y| \text{ but } x \neq y\}$. Show that D is a context-free language by constructing a PDA that recognizes it.

Solution sketch: From the fact $|x| = |y|$ and $x \neq y$, we know that there must exist at least one position in x with the symbol different from the symbol at the same position in y , i.e., $x = \Sigma^i 0 \Sigma^j, y = \Sigma^i 1 \Sigma^j$, or $x = \Sigma^i 1 \Sigma^j, y = \Sigma^i 0 \Sigma^j$. Observe that the two corresponding positions are $n/2$ symbols apart, where n is the length of the entire input. Hence, by ensuring that the number of symbols between the two corresponding positions is equal to the number of other symbols, we can construct a CFG $G = (V, \{0, 1\}, R, S)$ that generates D .

- (a) Set two nonterminals, say A and B , in V and the appropriate rules in R as follows:

$$\begin{aligned} A &\rightarrow XAX|0 \\ B &\rightarrow XBX|1 \\ X &\rightarrow 0|1 \end{aligned}$$

Note: $A \xRightarrow{*} X^i 0 X^i$ and $B \xRightarrow{*} X^j 1 X^j$, thus $AB = X^i 0 X^i X^j 1 X^j = X^i 0 X^{i+j} 1 X^j$. Let the first half of the string AB be x , and the second half be y , then $x = X^i 0 X^j$ and $X^i 1 X^j$. It is obvious that $|x| = |y|$. The number of symbols between the specific 0 and 1 in AB is $i + j$, which is equal to the number of the rest symbols. Hence, we can be sure that the 0 in A and the 1 in B are on the same position in x and y , which achieves the condition $x \neq y$. The string BA can be shown with the similar reason.

- (b) Now set the grammar axiom to be S and the S -rules to be $S \rightarrow AB|BA$. The two shortest strings in D are 01 and 10, all the longer strings can also be generated.

- (c) The grammar thus obtained is:

$G_D = (\{S, A, B, X\}, \{0, 1\}, \{S \rightarrow AB|BA, A \rightarrow XAX|0, B \rightarrow XBX|1, X \rightarrow 0|1\}, S)$.

Note: one can easily proof that $D \subseteq L(G_D)$ and $L(G_D) \subseteq D$ therefore G_D generates the required language. Now, one can use the construction from Lemma 2.21 to construct a PDA that recognizes the language D .

10. (20 points) Show that $F = \{a^i b^j \mid i = kj \text{ for some } k > 0\}$ is not context free.

Solution sketch: Assume to the contrary that F is a CFL. Let p be the pumping length given by the pumping lemma. We show that $s = a^{2p^2} b^{2p} \in F$ cannot be pumped. Let $s = uvxyz$. The following lists all the cases which satisfy the condition $|vxy| \leq p$.

- (a) If either v or y contains two types of symbols, uv^2xy^2z contains the symbols out of order and is not in F .
- (b) If both v and y contain only **a**-s, uv^2xy^2z has the form $a^{2p^2+l}b^{2p}$ where $l \leq p < 2p$. But $2p^2 + l$ is not a multiple of $2p$ if $l < 2p$, so $uv^2xy^2z \notin F$.
- (c) If both v and y contain only **b**-s, uv^mxy^mz for $m = 2p^2$ has the form $a^i b^j$ where $i < j$ and so it cannot be in F .
- (d) If v contains only **a**-s and y contains only **b**-s, let $t = uv^mxy^mz$. We show that string t cannot be member of F for any sufficiently large value of m . Write s as $a^{g+|v|}b^{h+|y|}$, then t is $a^{g+m|v|}b^{h+m|y|}$. If $t \in F$ then $g + m|v| = k(h + m|y|)$ for some integer $k > 0$. In other words,

$$k = \frac{g + m|v|}{h + m|y|}.$$

If m is sufficiently large, g is a tiny fraction of $m|v|$ and h is a tiny fraction of $m|y|$ so we have

$$k = \frac{m|v|}{m|y|} = \frac{|v|}{|y|}.$$

We know that $k < p$ because $|v| + |y| \leq p$. From $g + m|v| = k(h + m|y|)$, we rewrite it as $g - hk = m(k|y| - |v|)$, which must be 0, due to the fact that $k = \frac{|v|}{|y|}$. That is to say, $g - hk = 0$.

Moreover, from $g + |v| = 2p^2$ and $h + |y| = 2p$, we get $g + |v| = ph + p|y|$. Then replace g with hk and $|v|$ with $|y|k$, we have $hk + |y|k = ph + p|y|$, which shows that $k = p$. This contradicts with $k < p$ shown above, so t cannot be member of F .

Thus none of the cases can occur, the string s cannot be pumped and F is not context free.

Due date: Monday 2 November 2009.