Final #2

Mark all correct answers in each of the following questions.

Unless stated otherwise, $G = (N, T, R, S)$ is a context-free grammar without useless letters.

4. (a) If $L(G)$ is infinite, and every word in $L(G)$ has at least two parse trees, then there exists at least one word in $L(G)$ that has infinitely many parse trees.

   (b) The grammar defined by the rules

   $E \to E + T + T \mid T,$
   $T \to T * F * F \mid F * F,$
   $F \to a \mid b \mid c,$

   is unambiguous. If the rule $T \to F * F$ is replaced by the rule $T \to \varepsilon$, then we again obtain an unambiguous grammar.

   (c) The grammar defined by the rules

   $S \to AD \mid EC,$
   $A \to aA \mid \varepsilon,$
   $C \to cC \mid \varepsilon,$
   $D \to bDc \mid \varepsilon,$
   $E \to aEb \mid \varepsilon,$

   is ambiguous. However, the language consisting of all words in $L(G)$, having more than one parse tree, is context-free.

   (d) Let $G_i = (N_i, T, R_i, S_i)$ for $i = 1, 2$, where $N_1 \cap N_2 = \emptyset$. It is given that $L(G_1)$ consists of all words in $\{a, b\}^*$ for which $|w|_b = |w|_a$, such that no proper prefix u of w (i.e., $u \neq \varepsilon, w$) satisfies $|u|_b = |u|_a$. The language $L(G_2)$ consists of all words in $\{a, b\}^*$ for which

$|w|_b = |w|_a + 1$. (Here $|v|_\sigma$ denotes the number of occurrences of the letter σ in the word v.) Then the grammar

$$G = (N_1 \cup N_2 \cup \{S\}, \{a, b\}, R_1 \cup R_2 \cup \{S \to S_1S_2, S \to S_2S_1S_1\}, S),$$

(where $S \notin N_1 \cup N_2$), is ambiguous.

5. (a) The grammar defined by the rules
   $S \to abcdeS \mid abcedS \mid \ldots \mid edcbaS \mid f$
   (namely, there are 121 rules, the right-hand sides of the first 120 of which are the 5! permutations of the word *abcde*, all followed by S, and that of the last one is f), is $LL(5)$ but not $LL(4)$.

   (b) Consider the grammar defined by the rules
   $S \to AS \mid A,$
   $A \to w_1A \mid w_2A \mid \ldots \mid w_mA \mid \varepsilon,$
   where $w_1, w_2, \ldots w_m$ are distinct non-empty words in T^*. In general, the grammar may be ambiguous. However, if none of the words w_i may be written as a concatenation of several other w_j's (perhaps with repetitions), then the grammar is unambiguous, and even $LL(k)$ for sufficiently large k.

   (c) Let $G' = (N, T, R', S)$, where

   $$R' = \{A \to w \in R : w \in T^*\} \cup \{A \to \alpha' : \exists A \to \alpha \in R, \alpha \underset{l}{\Longrightarrow} \alpha'\}.$$

   (Here $\alpha \underset{l}{\Longrightarrow} \alpha'$ means that α yields α' by employing a single leftmost derivation.) Suppose G is an $LL(1)$ grammar. Then G' is $LL(2)$, but not necessarily $LL(1)$.

   (d) Let $G' = (N, T, R', S)$, where R' is obtained from R as follows: Each rule $A \to \alpha$ is replaced by a rule $A \to \alpha'$, such the first letter of α' coincides with that of α. (In particular, if $|\alpha| \le 1$ then $\alpha' = \alpha$.) Then G' is $LL(1)$ if and only if G is such.

6. (a) Denote (for the purposes of this question):

   $$LC(aA) = \{\beta \in (N \cup T)^* : S' \overset{*}{\underset{r}{\Longrightarrow}} \beta aAw,\ (w \in T^*)\}, \qquad a \in T, A \in N.$$

   Then the language $LC(aA)$ is regular for every $a \in T, A \in N$.

(b) If $S \to SaS \in R$ (in addition to other rules), then $LC(S) \supseteq L(G)\{a\}$.

(c) The grammar defined by the rules
$S \to abSc \mid cbSa \mid bSac \mid bca$
is $LR(0)$.

(d) The grammar defined by the rules
$S \to SSSa \mid b$
is $LR(0)$.

Solutions

4. (a) Given any unambiguous grammar, we can turn it into a grammar accepting the same language, with exactly two parse trees for every word. In fact, let $G = (N, T, R, S)$ be the initial grammar. Take two "copies" of G, say $G_i = (N_i, T, R_i, S_i), i = 1, 2$, where $N_1 \cap N_2 = \emptyset$ and the sets of non-terminals and of rules of each G_i are "equivalent" to those of G. That is, if $A \to \alpha \in R$, then $A_i \to \alpha_i \in R_i$, where α_i is obtained from α by replacing each non-terminal B by B_i. Now let $G' = (N_1 \cup N_2 \cup \{S'\}, T, R_1 \cup R_2 \cup \{S' \to S_1, S' \to S_2)$. It is easy to verify that G' satisfies the claim.

For example, the grammar defined by the rules

$S \to SA \mid A,$
$A \to aAb \mid ab,$

is easily seen to be unambiguous. The grammar defined by the rules

$S \to S_1 \mid S_2,$
$S_1 \to S_1A_1 \mid A_1,$
$A_1 \to aA_1b \mid ab,$
$S_2 \to S_2A_2 \mid A_2,$
$A_2 \to aA_2b \mid ab,$

accepts the same language, each word in exactly two ways.

(b) To show that the grammar is unambiguous, suppose we are given a word $w \in L(G)$. Suppose in the process of deriving this word, the rule $E \to E + T + T$ is applied n times. Since no other rule has a '+' on the right-hand side, this means that w must contain exactly $2n$ occurrences of this symbol. In other words, the number of occurrences of '+' in w determines uniquely the number of times the rule $E \to E + T + T$ must be applied. Now we need to show that from a word of the form $T + T + \ldots + T$ we can produce w in a unique way. In fact, similarly to the preceding stage, we see that the number of occurrences of the symbol '$*$' between any two consecutive occurrences of '+' in w determines uniquely the number of times the rule $T \to T * F * F$ has been applied to the initial T between them. Finally, each occurrence of a, b, c is due to an application of the rules producing these letters from F.

The situation if the rule $T \to T * F * F$ is replaced by $T \to \varepsilon$ is very similar. The difference is that the string between any two consecutive occurrences of '+' may be empty, and it starts ("unnaturally") with a '$*$' if it is non-empty.

(c) From the non-terminal A, one can produce the language $\{a\}^*$, from C – the language $\{c\}^*$, from D – the language $\{b^n c^n : n \geq 0\}$, and from E – the language $\{a^n b^n : n \geq 0\}$. Thus, from AD one can produce the language $\{a\}^*\{b^n c^n : n \geq 0\}$, and from BC – the language$\{a^n b^n : n \geq 0\}\{c\}^*$. It is readily seen that each word in the latter two languages can be produced in a unique way from the mentioned string. Since the intersection of the two languages is $\{a^n b^n c^n : n \geq 0\}$, this language is exactly the set of all words with two parse trees. Summing up, G is ambiguous, and the language consisting of all words with more than one parse tree (in our case – exactly two such trees) is non-context-free.

(d) Clearly, $ab \in L(G_1)$ and $abb, babab \in L(G_2)$. Hence the word $abbabab$ can be produced in G in two different ways,

$$S \Longrightarrow S_1 S_2 \overset{*}{\Longrightarrow} (ab)(babab) = abbabab,$$

and

$$S \Longrightarrow S_2 S_1 S_1 \overset{*}{\Longrightarrow} (abb)(ab)(ab) = abbabab.$$

Thus, (b) and (d) are true.

5. (a) We claim that the grammar is $LL(4)$. Indeed, assume we have to decide which rule to use. If the next input letter is f, then clearly we must use the rule $S \rightarrow f$. If not, then the next 4 letters are 4 distinct letters out of the 5 letters a, b, c, d, e. The letter following these must be the one not represented among the 4. Thus, based on the next 4 letters we know that the next 5 are going to be, say, $\sigma_1\sigma_2\sigma_3\sigma_4\sigma_5$. We now must use the rule $S \rightarrow \sigma_1\sigma_2\sigma_3\sigma_4\sigma_5 S$.

   We mention that the grammar is clearly not $LL(3)$.

   (b) The grammar may be ambiguous also due to the concatenation of several w_i's being equal to that of several others. For example, suppose $w_1 = a, w_2 = ab, w_3 = bc, w_4 = c$. Then the word abc has two distinct leftmost derivations,

$$S \Longrightarrow AS \Longrightarrow aAS \Longrightarrow aS \Longrightarrow aA \Longrightarrow abcA \Longrightarrow abc,$$

   and

$$S \Longrightarrow AS \Longrightarrow abAS \Longrightarrow abS \Longrightarrow abA \Longrightarrow abcA \Longrightarrow abc.$$

   (c) Consider the grammar defined by the rules

   $S \rightarrow abS \mid b.$

   Clearly, the grammar is $LL(1)$. The grammar obtained from it according to the process in the question is defined by the rules

   $S \rightarrow ababS \mid abb \mid b.$

   This grammar is not $LL(2)$ as, for words starting with ab, the first two letters do not determine the rule to use already at the first step.

   (d) Consider the grammar defined by the rules

   $S \rightarrow Ab \mid c,$

   $A \rightarrow a \mid \varepsilon.$

   It is readily verified that the grammar is $LL(1)$. However, the grammar defined by the rules

$S \to Ac \mid c,$
$A \to a \mid \varepsilon,$

obtained from it when replacing the rule $S \to Ab$ by $S \to Ac$, is not even unambiguous (as the word c has two parsing trees).

Thus, none of the claims is true.

6. (a) By the definition of $LC(aA)$, this language consists of all words in $LC(A)$ ending with a, with this a omitted. Since $LC(A)$ is regular, so is the intersection $LC(A) \cap (N \cup T)^*\{a\}$. Now, by omitting the last letter from all words in some regular language, we obtain again a regular language. Hence $LC(aA)$ is regular.

   (b) Suppose G is defined by the rules

   $S \to SaS \mid b.$

   Then $LC(S) = \{\varepsilon\} \cup LC(S)\{Sa\}$, which yields $LC(S) = \{Sa\}^*$. On the other hand, $L(G)$ includes the word b, and hence $LC(S)$ does not contain $L(G)\{a\}$.

   (c) We have

$$LC(S) = \{\varepsilon\} \cup LC(S)\{ab\} \cup LC(S)\{cb\} \cup LC(S)\{b\},$$

   which yields

$$LC(S) = \{ab, b, cb\}^*.$$

   It follows that:

$$\begin{array}{lcl} LR(0)\text{-}C(S \to abSc) & = & \{ab, b, cb\}^*\{abSc\}, \\ LR(0)\text{-}C(S \to cbSa) & = & \{ab, b, cb\}^*\{cbSa\}, \\ LR(0)\text{-}C(S \to bSac) & = & \{ab, b, cb\}^*\{bSac\}, \\ LR(0)\text{-}C(S \to bca) & = & \{ab, b, cb\}^*\{bca\}. \end{array}$$

   Denote these four languages by L_1, L_2, L_3, L_4. A word α in one of the first three of these languages is clearly not a prefix of another word in the same language due to the location of S in the word. A word in L_4 ends with ca, but does not contain this subword anywhere else, so that words in L_4 are not prefixes of each other. For the same reason, a word in L_4 is not a prefix of a word in

L_1, L_2, L_3. In the other direction, a word $\alpha \in L_1, L_2, L_3$ is clearly not a prefix of a word $\beta \in L_4$, as α includes an S, whereas β does not. To see that a word in L_3 is not a prefix of a word in L_1, L_2, we just need to note the location of the single occurrence of S in the two words. Similar reasoning shows that a word in either L_1 or L_2 is not a prefix of a word in the other, and a word in L_1 is not a subword of a word in L_3.

Now we show that a word $\alpha \in \ L_2$ is not a prefix of some $\beta \in L_3$. In fact, assume α is a prefix of β. Then there exists a word $w \in \{ab, b, cb\}^*$ such wc also belongs to $\{ab, b, cb\}^*$ and $\alpha = wcbSa, \beta = wcbSac$. However, no word in $\{ab, b, cb\}^*$ ends with c, and consequently this situation is also impossible.

Finally, $LR(0)\text{-}C(S' \to S) = S$. Since S is not a prefix of any word in L_1, L_2, L_3, L_4, neither is any such word a prefix of S, the condition for a grammar to be $LR(0)$ is satisfied, so that G is such.

(d) The grammar is clearly not $LR(0)$. Suppose the input word is, say, b. We reduce the b to S, but then we do not know whether we should reduce this S to S' or shift.

Using the criterion for a grammar to be $LR(0)$, we first see that

$$LC(S) = \{\varepsilon\} \cup LC(S)\{S\} \cup LC(S)\{SS\},$$

which yields

$$LC(S) = \{S\}^*.$$

It follows that:

$$\begin{array}{lcl} LR(0)\text{-}C(S \to SSSa) & = & \{S\}^*\{SSSa\}, \\ LR(0)\text{-}C(S \to b) & = & \{S\}^*\{b\}. \end{array}$$

Obviously, no word in any of these two languages is a prefix of some other word in the same language or the other. However, the word $S \in LR(0)\text{-}C(S' \to S)$ is a prefix of all these words, which implies that the grammar is not $LR(0)$.

Thus, (a) and (c) are true.