CSE340 Fall 2020 - Homework 1

Due: Wednesday September 16 2020 by 11:59 PM on Gradescope

All submissions should be typed. Exception can only be made for drawing parse trees, which can be hand drawn and scanned in the submitted document.

When you submit your solution on Gradescope, you should indicate for each problem the page on which the solution is. Remember that no late submissions are accepted for homework assignments.

Problem 1. Consider the list of tokens

T1 = { "abc", abcd1e" }

T2 = { "abd" }

ID = Set of strings that consist of a letter or underscore that is followed zero or more letters, underscore or digits.

NUM = Set of strings that consist of a non-zero digit that is followed by 1 or more digits or the string "0".

Consider the input

abcd22abc 123 00abc abd

and the following sequence of calls:

t1 = lexer.GetToken();
t2 = lexer.GetToken();
t3 = lexer.peek(1);
t4 = lexer.peek(3);
t5 = lexer.peek(5);
t6 = lexer.peek(7);
lexer.UngetToken(2);
t7 = lexer.GetToken();
t8 = lexer.GetToken();

Assume that space is a separator, but is otherwise ignored.

1. What are the values of t1, t2, t3, t4, t5 and t6?

   Answer

   If GetToken() is called repeatedly, the sequence of tokens that is returned is the following

abcd22abc	123	0	0	abc	abd
ID	NUM	NUM	NUM	T1	T2

   This sequence is the basis for answering the two parts of the question

   t1 = { ID, "abcd22abc" }
   t2 = { NUM, "123" }
   t3 = { NUM, "0" }
   t4 = { T1, "abc" }
   t5 = EOF
   t6 = EOF

2. What are the values of t7 and t8?

   Answer After the lexer.UngetToken(2) is executed, the two tokens that were previously consumed are returned to the input and the input will look as if nothing was consumed. The values of t7 and t8 will be the same as t1 and t2:

   t7 = { ID, "abcd22abc" }
   t8 = { NUM, "123" }

Problem 2. Consider the grammar

S → A B C
A → a A b | ε
B → b B | b
C → b C a | b

1. What is the start symbol? Explain briefly!

   Answer. The start symbol is S. By convention, and unless otherwise specified, the left hand side of the first rule is the start symbol

2. What are the non-terminals? Explain briefly!

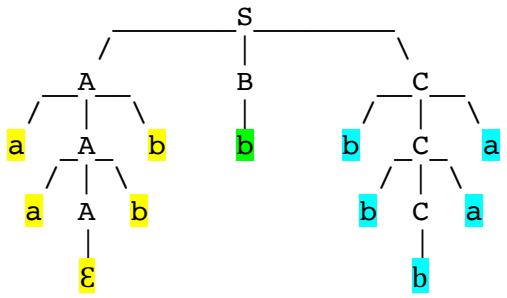
   Answer. The non-terminals are S, A, B and C. By convention, and unless otherwise specified, the left hand side of the rules are the non-terminals.

3. What are the terminals? Explain briefly!

   Answer. The terminals are a and b. By convention, and unless otherwise specified, the symbols that are not non-terminals are terminals. Epsilon is not a terminal. It does not correspond to a token

4. Give a parse tree for the input:

   a a b b b b b b a a



5. In the parse tree of

   a a b b b b b b a a

   the root node is labeled S and its children are labeled A, B and C from left to right. Which parts of the input correspond to the children of S in the parse tree?

   Answer The part of input that corresponds to A is aabb

   The part of input that corresponds to B is b

   The part of input that corresponds to C is bbbaa

Problem 3. Consider the grammar

$A \rightarrow X\ Y \mid Z\ X$

$X \rightarrow a\ B \mid Y$

$B \rightarrow b\ B \mid X \mid \varepsilon$

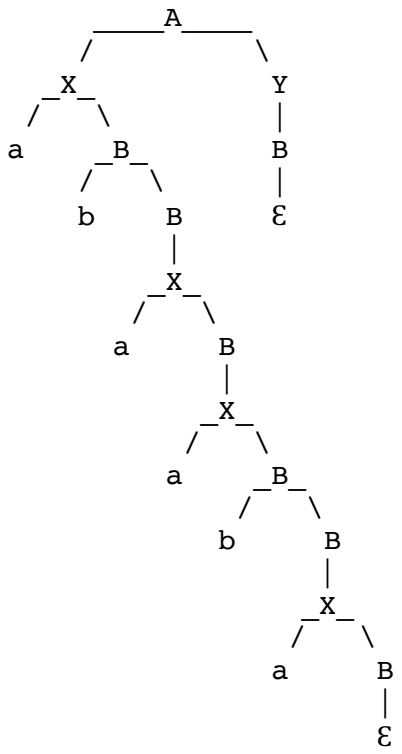
$Y \rightarrow a\ Y\ b \mid B$

$Z \rightarrow a\ Z\ b \mid X$

where A, B, X, Y and Z are non-terminal, A is the start symbol and a and b are tokens. Remember that ε represent the empty string. $Y \rightarrow \varepsilon$ means that Y does not have to match any tokens.

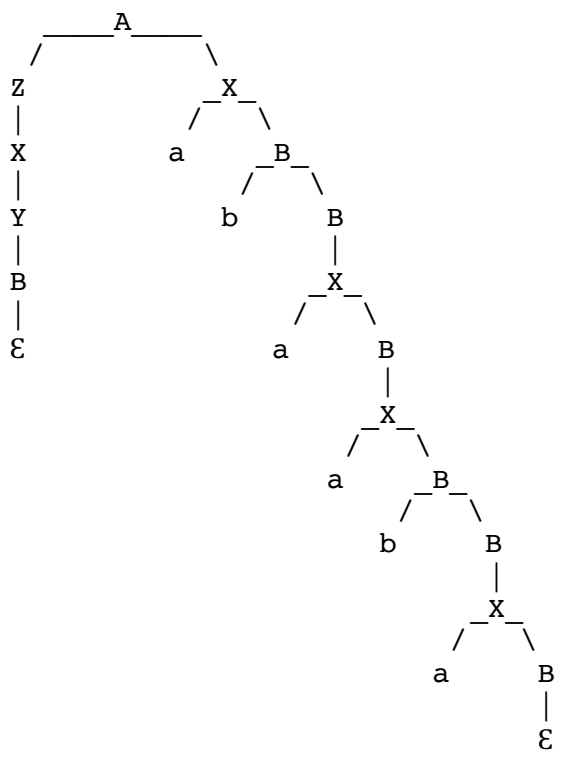
1. Give a parse tree for the sequence of tokens: a b a a b a

Answer



2. Give a another (different) parse tree for the sequence of tokens: a b a a b a

Answer



3. Is this grammar ambiguous? Why?

Answer This grammar is ambiguous because the input a b a a b a has two different parse trees