University of Central Florida

**Department Computer Science**

**COP3402: Systems Software**

**Summer 2021**

**Homework #2 (Lexical Analyzer)**

**(Solo or Team assignment – Same team members who implemented HW1)**

**Due June 25, 2021 by 11:59 p.m.**

***If you do not follow the specifications your grade will be zero***

**Goal:**

In this assignment you have to implement a lexical analyzer for the programming language PL/0. Your program must be capable to read in a source program written in PL/0, identify some errors, and produce, as output, the source program lexeme table and a list of lexemes. ***For an example of input and output refer to Appendix A****.*

**Lexical Conventions for PL/0:**

***A numerical value is assigned to each token (internal representation) as follows:***

oddsym= 1, eqlsym = 2, neqsym = 3, lessym = 4, leqsym = 5, gtrsym = 6, geqsym = 7, modsym = 8, multsym = 9, slashsym = 10, plussym = 11, minussym = 12,

lparentsym = 13, rparentsym = 14, commasym = 15, periodsym = 16, semicolonsym = 17, becomessym = 18, beginsym = 19, endsym = 20, ifsym = 21, thensym = 22, elsesym = 23, whilesym = 24, dosym = 25, callsym = 26, writesym = 27, readsym = 28, constsym = 29, varsym = 30, procsym = 31, identsym = 32, numbersym = 33.

***Reserved Words:*** const, var, procedure, call, if, then, else, while, do, begin, end, read, write, odd.

***Special Symbols* ::= ==** **<> < <= > >= % \* / + - ( ) , . ; :=**

***Identifiers:*** identsym **::=** letter (letter | digit)\*

***Numbers:*** numbersym **::=** (digit)**+**

***Invisible Characters:*** tab, white spaces, newline

***Comments denoted by:*** /\* . . . \*/

*Refer to* ***Appendix B*** *for a declaration of the token symbols that may be useful.*

**Constraints:**

***Input:***

1. Identifiers can be a maximum of 11 characters in length.
2. Numbers can be a maximum of 5 digits in length.
3. Comments should be ignored and not tokenized.
4. Invisible Characters should be ignored and not tokenized.

**Important Note: Input files may NOT be grammatically valid PL/0 code.**

***Output:***

1. In your output's Token List, identifiers must show the token and the variable name separated by a space.
2. In your output's Token List, numbers must show the token and the value separated by a space.
3. The token representation of the Token List will be used in the Parser (Project 3). So, PLAN FOR IT!

**Detect the Following Lexical Errors:**

1. Variable does not start with letter.
2. Number too long.
3. Name too long.
4. Invalid symbols.
5. Neverending comment.

**Submission Instructions**:

***Submit to Webcourse:***

1. Source code named lex.c
2. Instructions to use the program in a readme text file.
3. This is a team assignment (the same team members who worked together in HW1) or a solo assignment.
4. Only one submission per team.
5. The name of all team members must be written in all source code header files, in a comment on the submission, and in the readme.
6. Include comments in your program.
7. Same policy on late submission as in HW1. If there is an extension, no late submissions accepted
8. Output should print to the screen and should follow the format in Appendix A. A deduction of 5 points will be applied to submissions that do not print to the screen.
9. The input file should be given as a command line argument. A deduction of 5 points will be applied to submissions that do not implement this.

**Hints:**

* You could create a transition diagram (DFS) to recognize each lexeme on the source program and once accepted, generate the token otherwise emit an error message.
* Use the C function iscntrl() to check for whitespace rather than hardcoding acceptable control characters. Your program should function regardless of what control characters are present in the input file.
* Use the C functions isalpha() and isdigit() to check for letters and digits respectively.
* Reading in each character line by line can cause lots of problems. Instead read in character by character with fgetc().
* The only guaranteed whitespace is the whitespace that separates identifiers, numbers, and reserved words. All other whitespace is optional. Assuming no length errors:
  + If an identifier is followed by a number with no whitespace, it is an identifier.
  + If an identifier is followed by a reserved word with no whitespace, it is an identifier.
  + If an identifier is followed by an identifier with no whitespace, it is an identifier.
  + If a number is followed by an identifier with no whitespace, it is an invalid identifier error.
  + If a number is followed by a reserved word with no whitespace, it is an invalid identifier error.
  + If a number is followed by a number with no whitespace, it is a number.
  + If a reserved word is followed by an identifier with no whitespace, it is an identifier
  + If a reserved word is followed by a number with no whitespace, it is an identifier
  + If a reserved word is followed by a reserved word with no whitespace, it is an identifier.

**Error Handling:**

* When your program encounters an error, it should print out an error message and stop.
* When you are reading in a token that begins with a letter, read in characters until you reach one that not alphanumeric (at which point you check for reserved word and move on to the next token) or you reach the twelfth consecutive alphanumeric character (at which point you print the Excessive Identifier Length error and stop).
* When you are reading in a token that begins with a number, read in characters until you reach on that is not alphanumeric (at which point you tokenize the number and move on) OR you reach the sixth consecutive number (at which point you print the Excessive Number Length error and stop) OR you reach a letter (at which point you print the Invalid Identifier error and stop).
* When you are reading in a token that starts with an invalid symbol, print out the error message and stop.

**For this assignment, we are providing you with a skeleton. You must implement the function lexeme \*lexanalyzer(char \*input) in lex.c. You may add as many helper functions and global variables as you desire. This function takes the input file contents as a string and should return the lexeme list unless there is an error, in which case it should return NULL. lex.c also includes two printing functions: printerror which will print the error message and free the lexeme list, and printtokens which will print the lexeme table and token list and mark the end of the lexeme list. Make sure to call one of these functions before you return from the lexanalyzer function. For these functions to work, you must have the list and list index as global variables. We also provide driver.c which handles reading in the input file and calling lexanalyzer as well as compiler.h which includes the lexeme struct and token type enumeration. Additionally a make file, two examples, and a bash script for testing are included.**

**Rubric**

-100 – Does not compiles

15 – Compiles

20 – Produces some entries to list/table before segfaulting or looping infinitely

5 – Follows IO specifications (takes command line argument for input file name and prints output to console)

5 – README.txt containing author names

10 – Prints both the list and table formats

15 – Prints out message and stops execution after encountering error

5 – Is not context dependent

5 – Is not whitespace dependent

10 – Supports all four errors

10 – Supports all thirty-three symbols

**Appendix A:**

***If the input is:***

const==var<>procedure<begin>end<=if>=then.else;while(do)call:=read,write+124-jalapeno\*/\*comment\*//

***The output should be:***

Lexeme Table:

lexeme token type

const 29

== 2

var 30

<> 3

procedure 31

< 4

begin 19

> 6

end 20

<= 5

if 21

>= 7

then 22

. 16

else 23

; 17

while 24

( 13

do 25

) 14

call 26

:= 18

read 28

, 15

write 27

+ 11

124 33

- 12

jalapeno 32

\* 9

/ 10

Token List:

29 2 30 3 31 4 19 6 20 5 21 7 22 16 23 17 24 13 25 14 26 18 28 15 27 11 33 124 12 32 jalapeno 9 10

***If the input is:***

12345abcdefghijklmnop$/\*

***The output should be:***

Lexical Analyzer Error: Invalid Identifier

**Appendix B:**

***Declaration of Token Types:***

typedef enum {

oddsym = 1, eqlsym, neqsym, lessym, leqsym, gtrsym, geqsym,

modsym, multsym, slashsym, plussym, minussym,

lparentsym, rparentsym, commasym, periodsym, semicolonsym,

becomessym, beginsym, endsym, ifsym, thensym, elsesym,

whilesym, dosym, callsym, writesym, readsym, constsym,

varsym, procsym, identsym, numbersym,

} token\_type;

***Token Definitions:***

oddsym 1 odd

eqlsym 2 ==

neqsym 3 <>

lessym 4 <

leqsym 5 <=

gtrsym 6 >

geqsym 7 >=

modsym 8 %

multsym 9 \*

slashsym 10 /

plussym 11 +

minussym 12 -

lparentsym 13 (

rparentsym 14 )

commasym 15 ,

periodsym 16 .

semicolonsym 17 ;

becomessym 18 :=

beginsym 19 begin

endsym 20 end

ifsym 21 if

thensym 22 then

elsesym 23 else

whilesym 24 while

dosym 25 do

callsym 26 call

writesym 27 write

readsym 28 read

constsym 29 const

varsym 30 var

procsym 31 procedure

identsym 32 identifiers

numbersym 33 numbers