CS4100/5100 COMPILER DESIGN
PROJECT LANGUAGE TOKENS

SPRING 2022

The tokens for the compiler project language, referred to as **PL22**, are described below. **Note that no token may exceed the end of an input line.**

1. **Identifiers:** The language is NOT case-sensitive, so all identifiers, including reserved words, should be recognized without regard to any capitalization. Identifiers consist of uppercase and lowercase letters, digits, underscores, and vertical bar signs ('|'). They must begin with a letter, and have a length limitation of 20. Identifiers longer than 20 characters are to be truncated to 20 **after they are fully read**, and a warning message is to be produced, but no fatal error should be generated. Identifiers cannot be broken across input line boundaries. The token code for identifiers is 50.

2. **Numeric constants:** A numeric value consists of at least 1 digit, followed by 0 or more digits, an optional period, and 0 or more digits, maximum length 15. Exponential notation may optionally be applied only to floats, using 'E' [or 'e'] followed by an optional '+' or '-', and at least 1 digit. The token code for integers is 51, and for floating-point values it is 52. The Regular Expression is:

```
        <digit>+[.<digit>*[ E|e [+|-]<digit>+]]
```

3. **String constants:** A string consists of a double-quote char ("), any characters **except** a line-terminator (carriage return, line feed), and a terminating double-quote. As with all of these tokens, a string may not exceed the end of a line. The token code for strings is 53.

4. **Comments:** may be delimited with *either* a paired (* and *) sequence *or* paired '{', '}' ('curly braces'). The two styles cannot be mixed in a single comment (the comment starting and ending delimiters must match), and, as one might logically assume, the outermost comment overrides any comments within it. A comment is **not,** technically, a token in itself, and its delimiters and contents are **completely ignored** by the lexical analyzer. A comment may extend to multiple lines and include line terminators. There is no token code for comments, and Lexical treats them essentially as white space.

5. **Reserved words:** The following are the PL22 reserved words, and their integer token code values:

0	GO_TO	4	IF
1	INTEGER	5	THEN
2	TO	6	ELSE
3	DO	7	FOR

8	OF
9	PRINTLN
10	READLN
11	BEGIN
12	END
13	VAR
14	DOWHILE
15	PROGRAM
16	LABEL
17	REPEAT
18	UNTIL
19	PROCEDURE
20	DOWNTO
21	FUNCTION
22	RETURN
23	FLOAT
24	STRING
25	ARRAY

6. **Other tokens:**

The following numerical values are to be used for the 1- and 2-character tokens below.

30	/
31	*
32	+
33	-
34	(
35	)
36	;
37	:=
38	>
39	<
40	>=
41	<=
42	=
43	<>
44	,
45	[
46	]
47	:
48	.
99	Used for any other input characters which are not defined elsewhere here