The University of Newcastle

School of Electrical Engineering and Computing

COMP3290 Compiler Design

Semester 2, 2019

Project Part 1 A Scanner for CD19 (15%) Due: August 30th

Change Notes: additions and updates will be highlighted, while redundant sections of text are ~~struck out~~. Document updates will result in a change of version number – this is v1.01

Part 1 - Scanner

You are to write a **Scanner** for the language CD19; the lexical description for CD19 follows, notice that you do not need to know anything of how the tokens are used once your scanner has recognised them. Make sure that you *do not do syntactic processing* in the scanner.

White Space:

CD19 is a *free format language*. Whitespace characters such as spaces and tabs are lexical delimiters in all cases except within comments and strings. Newline characters are also whitespace, except that they delimit comments and also except that a newline character within a string is a lexical error. When reading from a text file, you may also need to specifically handle the carriage-return character as a whitespace character, depending upon which editor was used to produce your CD19 source file.

Keywords:

These are: **CD19 constants types is arrays main begin end array of function void const integer real boolean for repeat until if else input print printline return not and or xor true false.**

The keywords of CD19 are reserved words and so cannot be used as identifier names. They are *not case sensitive* (so *Begin, BEGIN* and *BEgin* are the same as the keyword *begin*). The programming convention in CD19, however, is to use keyword variants which only have lower case characters, except for the CD19 keyword which is uppercase.

Delimiters and Operators:

The following characters and character sequences are used to identify particular language elements. The definitive list can be found in the CD19 Programming Language Specifications. CD19 also has comment delimiters as outlined below.

semicolon (**;**) leftbracket (**[**) rightbracket (**]**) comma (**,**) leftparen (**(**) rightparen (**)**) equals (=) plus (+) minus (**-**) star (*****) slash (/) percent (**%**) carat (^) less (**<**) greater (**>**) quote (**"**) colon (**:**) dot (**.**). Some of these are combined to form operators such as: **<=**, **>=**, **!=**, **==**, **+=**, **-=**, ***=** **/=**. These composite operators may NOT contain embedded whitespace characters, i.e. *equals-space-equals* will be returned as *two equals tokens*.

Comments:

Single line comments may begin with **/--** whereupon they continue until the end of the current line (they are delimited by the next newline character). CD19 does not allow multi-line comments.

Identifiers and Reserved Keywords:

Identifiers begin with a letter and contain any number of letters and digits. Keywords such as *CD19, constants, types, arrays*, etc are reserved and may not be used as identifier names. All identifiers are CaSE senSItivE, which means, for example, that xModule and Xmodule are different identifiers.

Integer Constants:

Integers contain any number of digits (and therefore can have leading zeroes). Please note that the value associated with the integer token cannot be negative (the string –3 should be returned to the parser as 2 successive tokens *minus* and *3*).

Floating Point Constants:

These follow the usual fixed point structure of <*integer*>.<*integer*>.

String Constants:

String constants are sequences of characters enclosed in double-quotes ("……") but may not contain newline characters. A string constant which is terminated by a newline character is an undefined (or error) token. There is *no provision* to cater for special characters using a mechanism such as escaping as used in C/C++ and Java.

Structure of the Scanner:

Constant declarations for the tokens and a string array that allows these values to be printed easily as strings are available via Blackboard under Project Specs.

The tokens returned will minimally be tuples **(tokNo, lexeme)** where **tokNo** is the token value and **lexeme** is usually null, but is a reference to a lexeme string for identifiers, integer constants, floating point constants, and string constants.

Other data that could be useful to error reporting and debugging would be the line number and column number of the token within the input file. Addition of a line and column number makes the token a 4-tuple **(tokno, lexeme, line, col)**.

The scanner should be a single object, driven by a very small driver of the style:

```
while not scanner.eof( ) do {
   token = scanner . gettoken( );
   scanner . printtoken(token);
}
```

A single message to a scanner object is the best structure because it will become the input processor for your parser and later parts of the project. If you do not follow this design style, it is highly likely that you will need to rewrite major parts of your scanner for **Project Part 2** (the Parser).

Output of the Scanner:

The output of your scanner must be a stream of tokens. You will need to write a special (henceforth useless) debug routine which will print the tokens as they are produced. This debug routine will print to standard output.

The token values should be printed as ascii strings, i.e. print the token value *TCD19* as the string **TCD19**, the *TPLUS* token value as **TPLUS**, etc. If you use Java, there will be some declaration code released that you may use if you wish. It contains a list of token numbers and an associated array of String constants that will print what is required for token values, you will note that these Strings are all 6 characters in length and contain trailing space characters. The end-of-file token *TEOF* will be the last token output (as **TEOF**).

Each line of output, in the absence of errors, will exceed 60 characters in length. Once any line of output has exceeded 60 characters then you should terminate that output line.

For identifiers, integers, reals, strings: print the token value followed by the lexeme for the id, integer, real, or string (for strings, output the double-quote characters, even though they are not part for the string itself). This second field is rounded up in length to the next multiple of 6 characters, trailing space filled, and must contain at least one trailing space.

In other words, if you have a row of only tokens, it will extend to 66 characters and then wrap. Or if you have a row that is currently up to 60 characters in length, you will print the next token (and any arguments/values) before wrapping.

For lexical errors: Print the token value **TUNDF** followed by the sequence of characters that constitutes the *undefined token* and then proceed to find the *next valid token* to be returned to the caller. See **Error Handling**.

Eventually the scanner will have to produce the program listing (it will be the only part of the compiler which will know about comments, for example), so you may like to start thinking about how to do this. It is not required but you'll have to do it eventually. If you decide to produce a special listing file then make sure you dump it out before your program exits. At the end of these specifications you will find some design hints on how to do this in a way that will best benefit later Parts of the project.

OUTPUT FORMAT IS IMPORTANT.

Error Handling:

Errors found (e.g. *a hash character which is not within a comment*) will be output as if they are undefined tokens (as outlined above), but they should also be reported separately as an error message from the compiler. This error reporting will survive into later project phases, where it will be augmented with other types of error messages as other errors are found.

Note that this is only the first stage of error reporting - misspelt keywords will be returned as valid identifiers, etc. and these errors may not be found until the parser. Misspelt identifiers may not be found until semantic checking is completed.

Also note that your scanner does not recognise sequences of valid items as being incorrect, even in cases such as <<= (which would be **`TLESS`** returned, and then **`TLEQL`** returned by the next request for a token).

A sequence such as /--= would be ignored, with the = being recognised as part of an inline comment (with the rest of that source line also being ignored as comment).

Your scanner only reports lexical errors. When a lexical error is found, an error message is to be printed on a line by itself (the next valid token gets printed on the line following the error report, beginning in col.1), after the **`TUNDF`** token is printed, itself on a new line:

```
TDOTT TIDNT seven TLPAR TRPAR
TUNDF
lexical error ?@@#
TIDNT next  TIDNT tokens      TIDNT here  TDOTT
```

For invalid strings or characters print out **lexical error** *X* (where *X* is the offending character or string). When a lexical error is found then the associated "undefined token" incorporates all characters up to but not including the next space, tab, newline, alphanumeric or operator character.

Note that a sequence such as `123abc` could be detected as a lexical error, but for this project, you will follow the "return next valid token" semantics and therefore a string such as this would be returned as two tokens – the integer constant `123`, and the identifier `abc`.

Restrictions:

As a formality, it must be mentioned that you are to write your own Scanner and NOT use any form of *third-party compiler tool* or *library* to achieve this. Additionally you are NOT to use *regular expressions* to match your *keywords*, *numerics*, or other *glyphs*.

Testing Your Scanner:

You are responsible for making up sufficient data files which will adequately test your scanner. See Part 2.

There may be a class suite of standard test programs released later, but this may not be exhaustive for the purposes of testing your scanner.

Note that you do not need to know what the grammatical structure of the language is in order to do this project, you only need to know what constitutes a valid lexical item in the language. If you find yourself consulting the syntax specification of CD19, you are probably going outside these specifications.

It is recommended that you plan out your attack on this project, don't write the whole thing and then go looking for bugs – you will finish up with a mess, impossible to read, understand and extend later. A short while writing (henceforth useless) debug routines will probably save you lots of time later.

Submission:

Project Part 1 (*Scanner*), is due on **Friday August 30th at 23:59pm** (end of Week 5) – please zip up all your files and submit them via the **Part 1 Submission Point** within the assessments tab on **Blackboard**.

Use a file name that contains your student number and "**`pt1`**" (e.g. **`c9876543_pt1.zip`**), and put your name into the associated comment field for the submission. Remember to incorporate an assignment cover sheet into your submission.

Place your java **source files in the root** of your zip archive, along with your **coversheet** (as a *PDF*).

Please ensure that your project can be compiled on the standard *University Lab Java environment*, using the command **`javac A1.java`**, and executed similarly with the command **`java A1 source.txt`**, where source.txt will be specified by the end user (note also, it *may or may not be a txt file*); **do not hardcode this filename**.

Please do not deviate from these folder structures and standards.

If you have enjoyed Part 1 of the Project and want to work ahead …

Listing:

Your scanner responds to a **`nextToken()`** request by returning the *next valid token* and by reporting any errors it finds along the way. The first extension to this is to re-produce the input which has been scanned as a separate listing file. This is just a separate text file which mirrors your input file except that it should have line numbers added at the start of each line. It can also output errors associated with any line after the line has been produced on the listing, or it can save up any errors messages until the end of the program and report them all (with their associated line numbers).

This is best done by sending messages to a separate output object and this output object can be used as a single place for the control of the output of the listing and the output of any error messages. In my compiler I call this object my *OutputController* and it responds to messages to *print a source code character*, *report an error*, etc.

Please note that this is separate to any token output required for Part 1 (which goes directly to standard output). If you work ahead with this, you may dump your listing file out to standard output before your program exits.

And even more …

Symbol Table:

Start to think about how specific identifier values, integer and real literal values, and string values should be stored. They will not be *needed* until Part 4 of the project, but they will have to be remembered for then and if they are not remembered now, then they can't be resurrected later. If we declare an identifier X, and then refer to it later, then we will have to tell that it is the *same* X. This will be done using a *Symbol Table*, which is another stand alone object (or set of objects) which will allow these lexeme values to be inserted and looked up later. For now it is best to have a hash table that can insert and look up string values as keys to a simple record/object structure which records the line and column number of where a lexeme is first found and then increments a counter each time it appears in the CD19 source program.

If you build a symbol table then do *not* include it in your Part 1 submission.

DB

v1.01 issued : *2019-07-31*