# CSE 3341 Project 1 - Core Scanner

## Overview

The goal of this project is to build a scanner for a version of the Core language, a pretend language we will be discussing in class.

For this project you are given the following:

- "3341 Project 1.pdf" - This handout. Make sure you read it completely and handle all requirements in your implementation. You are encouraged to post any questions on Piazza.
- "ScannerOutline.pdf" - These are some lectures slides which outline my recommended approach towards implementing the scanner.
- "Main.java", "Core.java", "Scanner.java" - I have outlined the project in this files and give you some of the code you will need. Make no changes to to "Core.java" or "Main.java".
- You may create additional files to contain any additional classes or methods you want to create.
- "tester.sh" - This is a script I wrote to help you test your project. It is very similar to the script that will be used to grade your project, so if your project works correctly with this script you are probably doing well. The only guarantee I give is that this script will work on stdlinux.
- Folder "Correct" - This contains some correct inputs and their expected outputs. The "tester.sh" script will test your code against these examples.
- Folder "Error" - This contains some inputs that should generate error messages. The "tester.sh" script will test your code against these examples.

The following are some constraints on your implementation:

- Do not use scanner generators (e.g. lex, flex, jlex, jflex, ect) or parser generators (e.g. yacc, CUP, ect)
- Use only the standard libraries of Java.

Your submission should compile and run in the linux environment the department provides (coelinux). I will leave it up to you to decide what IDE you will use or if you will develop your code locally or remotely, but as a final step before submitting your code please make sure it works on coelinux. **Use the "module avail" command to add openjdk**. The graders will not spend any time fixing your code - **if it does not compile on coelinux, your project will not be graded and you will get a 0**.

## Your Scanner

You are responsible for writing a scanner, which will take as input a text file and output a stream of "tokens" from the Core.java enumeration. You scanner must implement the following methods:

- Scanner: These functions open the file, find the first token, and release memory when we are done scanning.
- currentToken: This function should return the token the scanner is currently on, without consuming that token.
- nextToken: This function should advance the scanner to the next token in the stream (the next token becomes the current token).
- getId: If the current token is ID, then this function should return the string value of the identifier. If the current token is not ID, behavior is undefined.
- getConst: If the current token is CONST, then this function should return the value of the constant. If the current token is not CONST, behavior is undefined.
- getString: If the current token is STRING, then this function should return the string value. If the current token is not STRING, behavior is undefined.

All of these functions will be necessary for the parser you will write in the second project. You are free to create additional functions.

## Input

The input to the scanner will come from a single ASCII text file. The name of this file will be given as a command line argument to the main function.

The scanner should process the sequence of ASCII characters in this file and should produce the appropriate sequence of tokens. There are two options for how your scanner can operate:
(1) the scanner can read the entire character stream from the file, tokenize it, stores all the tokens in some list or array and calls to currentToken and nextToken simply walk through the list
or
(2) the scanner reads from the file only enough characters to construct the first token, and then later reads from the file on demand as the currentToken or nextToken functions are called.
Real world scanners typically work as described in (2). In your implementation, you can implement (1) or (2), whichever you prefer.

Once your scanner has scanned the entire file, it should return the EOS token (End Of Stream).

## Invalid Input

Your scanner should recognize and reject invalid input with a meaningful error message. The scanner should make sure that the input stream of characters represents a valid sequence of tokens. For example, characters such as '_' and '%' are not part of a valid sequence of tokens. If your scanner encounters a problem, it should print a meaningful error message to standard out (please use the format "ERROR: Something meaningful here") and return the ERROR token so the main program halts.

## The Language

The Core language consists of 4 kinds of strings, which you will need to tokenize:

- **Keywords:**
  and begin case do else end for if in integer is
  new not object or print procedure read return then

- **Identifiers:**
  Begins with a letter (uppercase or lowercase) followed by zero or more letters/digits.
  Refer to this regular expression once we cover regular expressions:
  (a|...|z|A|...|Z)(a|...|z|A|...|Z|0|1|...|9)*

- **Constants:**
  Integers from 0 to 1000003 (inclusive)

- **Symbols:**
  + - * / = == < : ; . , ( ) [ ] { }

- **Strings:**
  Any sequence of symbols contained within single quotes

Your scanner walk through the input character stream, recognize strings from the language, and return the appropriate token from the enumeration in "Core.java" or "Core.py". If there is any situation in which it is unclear to you which token should be returned, please ask for clarification on Piazza.

Write your scanner with these rules in mind:

1. The language is case sensitive, and the keywords take precedence over the identifiers. For example, "begin" should produce the token BEGIN (not ID), but "bEgIn" should produce the token ID.

2. Strings in the language may or may not be separated by whitespaces. For example the character stream may contain the string "x=10" or the string "x = 10", and both of these should generate the token sequence ID ASSIGN CONST.

3. Always take the greedy approach. For example, the string “forfor” should produce an ID token instead of two FOR tokens, string “123” should produce a single CONST token, and string “==” should produce EQUAL.

4. Keyword/identifier strings end with either whitespace or a non-digit/letter character. For example:

   (a) the string “for (” and the string “for(” should both result in the FOR and LPAREN tokens.

   (b) the string “for 12” should result in the FOR and CONST tokens, but the string “for12” should result in the ID token.

5. Constant strings end with any non-digit character. For example:

   (a) the string “120for” or “120 for” should result in the CONST and FOR tokens.

   (b) When applying the greedy rule to constants, you may assume a continuous sequence of digit characters is meant to be a single constant, then report an error if it does not form a valid constant. For example, for strings like “007” or “1234567890” your scanner should report an error rather than trying to break them up into valid constants.

6. Symbols may or may not be separated from other strings by whitespace. For example:

   (a) String “++while<= =12=” should result in the token sequence ADD ADD ID LESS ASSIGN ASSIGN CONST ASSIGN.

7. There is no support for escape characters. So a string like ”0+’123\’+’4 5 6’+” should result in the token sequence CONST ADD STRING ADD STRING ADD.

Let me know if you think of any situations not covered here.

## Testing Your Project

I have provided some test cases. For each correct test case there are two files (for example 4.code and 4.expected). On stdlinux you can redirect the output of the main program to a file, then use the diff command to see is there is any difference between your output and the expected output. For an example of how to do this you can take a look at the script file ”tester.sh”.

The test cases are weak. You should do additional testing with your own test cases. Feel free to create and post additional test cases on piazza.

## Project Submission

On or before 11:59 pm January 24th, you should submit to the Carmen dropbox for Project 1 a single zip file containing the following:

- All your .java files.
- An ASCII text file named README.txt that contains:
  - Your name on top
  - The names of all files you are submitting and a brief description stating what each file contains
  - Any special features or comments on your project
  - Any known bugs in your scanner

If the time stamp on your submission is 12:00 am on January 25th or later, you will receive a 10% reduction per day, for up to three days. If your submission is more than 3 days late, it will not be accepted and you will receive zero points for this project. If you resubmit your project, only the latest submission will be considered.

## Grading

The project is worth 100 points. Correct functioning of the scanner is worth 65 points. The handling of errors is worth 20 points. The implementation style and documentation are worth 15 points.

## Academic Integrity

The project you submit must be entirely your own work. Minor consultations with others in the class are OK, but they should be at a very high level, without any specific details. The work on the project should be entirely your own; all the design, programming, testing, and debugging should be done only by you, independently and from scratch. Sharing your code or documentation with others is not acceptable. Submissions that show excessive similarities (for code or documentation) will be taken as evidence of cheating and dealt with accordingly; this includes any similarities with projects submitted in previous instances of this course.

Academic misconduct is an extremely serious offense with severe consequences. Additional details on academic integrity are available from the Committee on Academic Misconduct (see http://oaa.osu.edu/coamresources.html). If you have any questions about university policies or what constitutes academic misconduct in this course, please contact me immediately.