Programming Assignment II

Due Friday, September 24, 2004

1. Overview

Programming assignments II-V will direct you to design and build a compiler for Cool. Each assignment will cover one component of the compiler: lexical analysis, parsing, semantic analysis, and code generation. Each assignment will ultimately result in a working compiler phase which can interface with other phases.

For this assignment you are to write a lexical analyzer, also called a scanner, using a lexical analyzer generator (the Java tool is called `jlex`). You will describe the set of tokens for Cool in an appropriate input format and the analyzer generator will generate the actual Java code for recognizing tokens in Cool programs. You must work in a group for this assignment (where a group consists of one or two people).

2. Files and Directories

To get started, create a directory where you want to do the assignment and execute one of the following commands in that directory. You should type:

```
gmake -f ~icom4029/cool/assignments/PA2J/Makefile source
```

(notice the “J" in the path name). This command will copy a number of files to your directory. Some of the files will be copied read-only (using symbolic links). You should not edit these files. In fact, if you make and modify private copies of these files, you may find it impossible to complete the assignment. See the instructions in the README file. The files that you will need to modify are:

- `cool.lex`

This file contains a skeleton for a lexical description for Cool. You can actually build a scanner with this description but it does not do much. You should read the jlex manual to figure out what this description does do. Any auxiliary routines that you wish to write should be added directly to this file in the appropriate section (see comments in the file).

- `test.cl`

This file contains some sample input to be scanned. It does not exercise all of the lexical specification but it is nevertheless an interesting test. It is not a good test to start with, nor does it provide adequate testing of your scanner. Part of your assignment is to come up with good testing inputs and a testing strategy. (Don't take this lightly - good test input is difficult to create, and forgetting to test something is the most likely cause of lost points during grading.)

You should modify this file with tests that you think adequately exercise your scanner. Our `test.cl` is similar to a real Cool program, but your tests need not be. You may keep as much or as little of our test as you like.

- `README`

This file contains detailed instructions for the assignment. You should also edit this file to include the write-up for your project. You should explain design decisions, why your code is correct, and why your test cases are adequate. It is part of the assignment to clearly and concisely explain things in text as well as to comment your code.

Although these files as given are incomplete, the lexer does compile and run (gmake lexer). There are a number of useful tips in the README file.

3. Scanner Results

You should follow the specification of the lexical structure of Cool given in Section 10 and Figure 1 of the CoolAid. Your scanner should be robust. It should work for any conceivable input. For example, you must handle errors such as an EOF occurring in the middle of a string or comment, as well as string constants that are too long. These are just some of the errors that can occur; see the manual for the rest.

You must make some provision for graceful termination if a fatal error occurs. Core dumps or uncaught exceptions are unacceptable.

Programs tend to have many occurrences of the same lexeme. For example, an identifier generally is referred to more than once in a program (or else it isn't very useful!). To save space and time, a common compiler practice is to store lexemes in a string table. We provide a string table implementation. See the following sections for the details.

All errors will be passed along to the parser. The Cool parser knows about a special error token called ERROR, which is used to communicate errors from the lexer to the parser. There are several requirements for reporting and recovering from lexical errors:

- When an invalid character (one which can't begin any token) is encountered, a string containing just that character should be returned as the error string. Resume lexing at the following character.
- When a string is too long, or contains invalid characters, that should be reported. Lexing should resume after the end of the string.

- If a string contains an unescaped newline, report that, and resume lexing at the beginning of the next line – we assume the programmer simply forgot the close-quote.
- If a comment remains open when EOF is encountered, report that. Do not tokenize the comment's contents simply because the terminator is missing. (This applies to strings as well.)
- If you see “*)" outside a comment, report this as an unmatched comment terminator, rather than tokenzing it as * and).
- Do not test whether integer literals fit within the representation specified in the Cool manual - simply create a Symbol with the entire literal's text as its contents, regardless of its length.

There is an issue in deciding how to handle the special identifiers for the basic classes (Object, Int, Bool, String), SELF TYPE, and self. However, this issue doesn't actually come up until later phases of the compiler. The scanner should treat the special identifiers exactly like any other identifier.

Your scanner should maintain the variable curr_lineno that indicates which line in the source text is currently being scanned. This feature will aid the parser in printing useful error messages.

Your scanner should convert escape characters in string constants to their correct values. For example, if the programmer types these eight characters:

"ab\ncd"

your scanner would return a token whose semantic value is these 5 characters:

ab¶cd

In this example, we use ¶ to represent the ascii code for newline. In JLex, you can produce this code by typing \n.

Finally, note that if the lexical specification is incomplete (some input has no regular expression that matches) then the scanners generated by jlex do undesirable things. Make sure your specification is complete.

4. Notes for the assignment

- The codes for all tokens are defined in the file `TokenConstants.java`. Each call on the scanner returns the next token and lexeme from the input. The semantic value or lexeme, can be obtained by calling the function `yytext()`, which returns a Java

String object. All of the single character tokens are listed in the grammar for Cool in the CoolAid.

- Each action for the lexemes in the last section of `cool.lex` should return an object of type `Symbol` with its `value` field containing the semantic value or lexeme unless it is not necessary (for example, a state-related action). The only parameter of the constructor for `Symbol` specifies the token type (integer constant, object id, etc) as specified in `TokenConstants.java` and in the API.

  Example:
  ```
  Symbol ret = new Symbol(TokenConstants.BOOL_CONST);
  ret.value = boolvalue;
  return ret;
  ```

- We provide you with a string table implementation, which is discussed in detail in A Tour of the Cool Support Code and documentation in the code. The type of string table entries is `Symbol`. Before returning a string constant, integer constant or an identifier, you must add it to the provided string tables in `AbstractTable.stringtable`, `AbstractTable.inttable` and `AbstractTable.idtable`, respectively.

  Example:
  ```
  AbstractSymbol lex_val = AbstractTable.idtable.addString(yytext());
  Symbol ret = new Symbol(TokenConstants.OBJECTID);
  ret.value = lex_val;
  ```

- When a lexical error is encountered, the action should return a Symbol of type `TokenConstants.ERROR`. Its `value` field should contain a `StringSymbol` with the error message. See previous section for information on what to put in error messages.

  Example:
  ```
  String err_msg = new String("Unterminated string");
  StringSymbol error = new StringSymbol(err_msg, err_msg.length(), 0);
  Symbol ret = new Symbol(TokenConstants.ERROR);
  ret.value = error;
  ```

- Additional information, including a Java API specification of the classes available, is available at the UC Berkeley CS 164 webpage (http://inst.eecs.berkeley.edu/~cs164/) when the Cool information becomes available again. The professor is working on that and an e-mail will be sent to the e-mail forum regarding this.

5. Testing the Scanner

There are at least two ways that you can test your scanner. The first way is to generate sample inputs and run them using `./lexer` which prints out the line number and the lexeme of every token recognized by your scanner. The other way, when you think your scanner is working, is to try running `./mycoolc` to invoke your `lexer` together with all other compiler phases (which we provide). This will be a complete Cool compiler that you can try on the sample programs and your program from Assignment I.
You can also look at the reference lexer's output and try to match yours to that:

```
    ~icom4029/cool/bin/lexer test.cl
```

Your output should be the same as the reference lexer. Do not output any unnecessary code or debug strings or your lexer will not work with the other phases of the compiler.

6. Turning In the Assignment

1. Make sure your code is in `cool.lex` and that it compiles and works.
2. Your test cases should be in `test.cl`.
3. Include any other relevant comments in the `README` file and answer any questions that appear in it.
4. Make sure everything (`cool.lex`, `test.cl`, and `README`) is in a directory called `PA2`.
5. Create a tar-gzipped file `PA2.tar.gz` containing the PA2 directory:
   ```
   tar -czf PA2.tar.gz PA2
   ```
6. Submit the file (before September 24 11:59pm):
   ```
   ~icom4029/submit/submit 2 PA2.tar.gz
   ```
   You can submit multiple times; if you do so, any previous submissions will be overwritten (until after the project's deadline).