dOvs - lexical analysis

Group 9
Miran Hasanagić - 20084902
Jakob Graugaard Laursen - 20093220
Steven Astrup Sørensen - 201206081

September 16, 2015

1 Introduction

This report describes our approach to develop the lexer, which is the first component in the compilation process.

First, it will provide an overview of the concrete solutions to specific problems, such as the handling of nested comments, escape codes and multi-line strings.

Additionally, it provides an overview of problems encountered and interesting experiments during the development process. Finally, five tiger programs are provided as test cases for the various elements of the lexer.

2 How did we implement the lexer

This lexer is implemented by using the existing tool ML-Lex, like it was suggested by the book. With this tool it is possible to specify which regular expression should map to which specific tokens. At first the support for all the basic constructs in the Tiger language were implemented. For example this includes the ability to recognise the keywords. Also the support for non-nested comments was implemented in this first version of the lexer.

2.1 Nested comments

In order to implement nested comments, we used ML-Lex's ability to have states. In this case we made a state called `COMMENT`. Additionally, a counter was used in order to handle nested comments. Once the counter reaches zero, we jump out of the `COMMENT` state. This is accomplished by the following code:

```
<INITIAL>"/*" => (commentLevel := ((!commentLevel)+1); YYBEGIN
COMMENT; continue());
<COMMENT>"/*" => (commentLevel := ((!commentLevel)+1); continue());
<COMMENT>"*/"=> (commentLevel := ((!commentLevel)-1);
 if !commentLevel < 1 then YYBEGIN INITIAL else (); continue());

<COMMENT>"\n" => (handleNewline yypos; continue());
<COMMENT>.    => (continue());
```

This code basically shows that for each `/*` we meet, the counter is incremented by one. Afterwards, for each `*/` the counter is decremented by one. When the counter is at zero, we jump out of the `COMMENT` state. Additionally, this counter value is used in order to detect unclosed comments at EOF, and raise an error. Namely, if the counter is not zero at EOF, we raise an error.

2.2 Escape codes

The escape codes where implemented when the lexer recognises that it is in the `STRING` state. Afterwards when a backslash is found, it check the following character/sequence in order to recognise if it is legal inside a string according to the tiger language specification.

When inside the `STRING` state, the three basic chars are handled by:

```
< STRING > "\\n"|"\\t"|"\\\\"|"\\\"" => (addToCurString yytext;continue());
```

Additionally, the three decimals after backslash are handled the following way:

```
< STRING > "\\"{digit}{3} => (addToCurString (digitEsc yypos yytext);continue());
```

`digitEsc` will convert the escape character to its standard representation, such converting 097 to a. It will raise an error if the number `ddd` is not in the correct range of ASCII values.

There is also some error handling for the decimal in order to ensure that there are exact three digits after the backslash.

```
< STRING > "\\"{digit}{1,2} => (ErrorMsg.error yypos (yytext ^ " is an illformed
ascii decimal escape code. ascii decimal escape code must be of the form
\\ddd, with 0 <= ddd <=" ^ (Int.toString maxAsciiCode) ^ " , and d a digit
between 0 and 9"); continue());
```

At last the escape control chars where handled in order to convert these to their respective escape characteres. This achieved by recognising the character `\^`. Then a function handles to find if the control character is a valid one, and provide the correct mapping using the function `handleCtrl`:

```
< STRING > "\\^". => (addToCurString (handleCtrl yytext yypos);continue());
```

2.3 Multiline strings

For strings also the possibility to split the string over multiple lines is supported by the implemented lexer. This is handle by having a state called `MULTILINE`, which is entered from the before mentioned state `STRING`, when only a "\" is typed inside a string, followed by either tab, space or newline. This change happens in the following code:

```
< STRING > "\\"(" "|"\t") =>(inMultiline := true; YYBEGIN MULTILINE; continue());
```

Inside this `MULTILINE` state we only allow space, tab and newline to be used in order to format the string. Afterwards, this state is left when the next "\" is received at the input of the lexer.

```
< MULTILINE > "\\" => (inMultiline := false; YYBEGIN STRING; continue());
```

Basically this functionality enables the support of multiline strings inside the lexer. There is, however, some additional error handling for multiline strings.

3 Problems experienced

One of the problems of the more interesting nature is how to handle EOF. While it is rather straightforward to make the lexer jump between relevant states, it quite another matter when wanting to query about it's current state. By looking through the documentation, there does not seem to be any built in way to do so.

This means that we decided to include boolean variables, that we can set when we enter and exit a state, to use at EOF to detect if, for example, a string is unclosed. We also use the commentLevel counter to detect if we still are in the COMMENT state.

Also it shall be noted that the expression */ was chosen to be treated as a close comment in all cases, and not a multiple and divide token.

A bizarre bug occurred when using a newline inside a comment. It caused everything to be considered a comment. The fix was going from

```
< INITIAL COMMENT MULTILINE >"\n" => (handleNewline yypos; continue());
```

to

```
< INITIAL >"\n" => (handleNewline yypos; continue());

< MULTILINE >"\n" => (handleNewline yypos; continue());

< COMMENT >"\n" => (handleNewline yypos; continue());
```

4 5 tiger programs

These test show five tiger programs, which where used in order to test the lexer. It shall be noted that the main focus of these test are to test if error are reported correctly in different cases. Additionally, it is used in order to test if the lexer correctly jumps between the states described above. For this reason the tests are made simple. However, in order to test for more complicated tiger programs, the provided test cases can be used.

Note there is more than the 5 test cases described in this report in the test folder. They are merely artefacts of the development process. They merely served to test earlier and incomplete implementations of the lexer.

4.1 own_test01.tig: Simple Comment and String

Here it is tested if the `COMMENT` state is entered correctly, and that the `STRING` state gets entered when necessary.

```
/* In this part we test a simple comment and string */

"Hello from a String"
```

The lexer behaves as expected, in that it ignores the comment and makes a string token.

4.2 own_test02.tig: Nested Comment and unclosed string

In this test a nested comment is tested, and that an error is issued because of the unclosed string.

```
/* A nested
/* comment */ */

"This is a unclosed string
```

The lexer behaves as expected. It ignores the nested comment, and issues an error for the unclosed string.

4.3 own_test03.tig: Nested comment, with error

In this part it is check if en error is issued when a nested comment is not closed correctly.

```
/* A nested
/* comment with an error */
```

The lexer behaves as expected, and gives an error for the unclosed nested comment.

4.4 own_test04.tig: Escape Characters and Errors

In this part escape characters are tested, and if the necessary errors are given for illegal characters. Lines 8-9 test the multiline string functionality.

```
"Testing tabs, backslash and newline: \t\n\\"
"Test ctrl char: \^@ \^[ \^? \^G \^H \^I \^J \^K \^L \^M"
"Test ill-formed ctrl char: \^A \^:"
"Testprintable ! # $ % ' ( ) * + , - . / : ;"
"Test digit escape: \0000 \027 \127 \007 \008 \009 \010 \011 \012 \013 u"
"Test printables again: < = > ? @ [ ] ^ _ ` { | } ~"
"Test illformed ascii code: \65 \9"
"\      \hi\
\ we only see a space between hi and we"
""
```

4.5 own_test05.tig: Switching between the different states

This last test just check if the switching between the different states is correct.

```
/* Test
Some more text for the hell of it
correct
*/
let function hello() = print("Hello"/*a comment /* going */ deep */) in
hello end
```

The lexer parses this tiger program as expected

Additionally, we have a interesting test inside `string2.tig`, which shows the multiline string error.

5 Conlusion

This report showed the implementation of the lexer. Furthermore some intersting problem encountered where described. Finally, some tests provided confidence that this lexer is working correctly.