Compiler Project

1. Introduction

The purpose of the project was to learn about the concepts of a compiler in an abbreviated and practical manner. Briefly, a compiler reads in a high level code written by a human and makes sure that all the symbols are valid. It will then pass along those symbols and check on the validity of groupings of symbols. The groupings are then passes along and checked against predetermined statement lists that define the language. Lastly, these groupings are then translated into code that a machine can execute at the register level.

This project was part of the Compilers class, which was a requirement in the completion of the degree of Bachelor of Science in Computer Science.

The goal of this project, to create a compiler for the Little programming language, was achieved in several steps. A brief introduction to each of these steps will be made in the next section. However, the purpose of the rest of this paper is to explain the details of this achievement in more detail.

2. Background

A compiler is a program that takes a text file written in a given high level language (i.e., Java or Python) and creates byte-code for the written program that can be executed on some target machine.

The purpose of a compiler is to allow programmers to develop coded projects using higher-level, more powerful languages that are easier to read and write than lower-level languages (i.e.: C or assembly).

Compilers are partitioned into four general components to reduce coupling and simplify the compilation process. These four components are:

* Scanner or Lexer
* Parser (syntax and semantic analysis)
* Symbol Table and Intermediate Code Generation
* Code Generation and Optimization

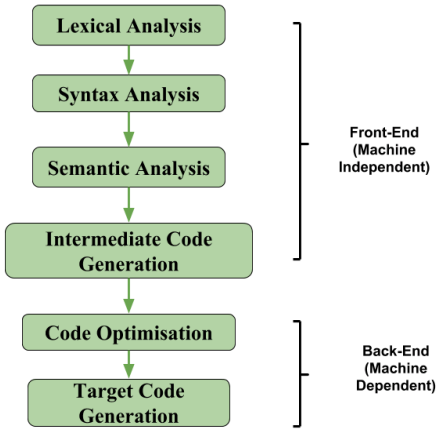


Figure 1: Parts of a compiler
https://www.geeksforgeeks.org/intermediate-code-generation-in-compiler-design/

Examples of compilers include gcc (C and C++), clang (for languages C, C++, Objective C/C++, OpenCL, CUDA, and RenderScript [https://clang.llvm.org/]), javac (for Java), go

(for the Go language), and Microsoft Visual Studio (for C#) (https://stackoverflow.com/questions /31180056/example-of-compiler-interpreter-and-both).

3. Methods and Discussion

A. ANTLR Setup

ANTLR is a program which generates a scanner and parser for the user given a configuration file (of the file format .g4). G4-files contain 2 sections: the lexer and the parser. This program is provided by ANTLR in the form of a jar named 'antlr-4.7.1-complete.jar'. This jar contains multiple programs, including one that we used to generate a lexer and a parser.

As part of the process of setting up ANTLR, we followed a tutorial on using ANTLR provided by ANTLR at www.antlr.org. The tutorial provided the following .g4-file named Hello.g4:

```
grammar Hello;
r  : 'hello' ID ;
ID : [a-z]+ ;
WS : [ \t\r\n]+ -> skip ;
```

After downloading and placing ANTLR's 'antlr-4.7.1-complete.jar' file-path into our environment variables, we were able to invoke the 'org.antlr.v4.Tool' program in this jar with the Hello.g4 file and an arbitrary string as input. This program scanned the string using the scanner defined in Hello.g4, and it determined if it was syntactically correct according to Hello.g4's grammar.

B. Scanner

A scanner (also known as a tokenizer or lexer) is a program that reads in a text file that is a series of characters representing a program of a specific language. The scanner then produces a series of tokens that are passed on to the parser to analyze. A token is the most basic of components of a programming language. Characters are separated into different types of tokens based on their functions (keywords, operators, identifiers, separators, and reserved words). These tokens are dependant on the rules of the specific programming language that is being used.

After installing and testing ANTLR, a partial grammar for the language called Little was created from the text file describing the structure of the language. The grammar file included rules for identifiers, keywords, and operators. Based on the grammar, ANTLR was used to generate the basics of the scanner in Java (see figure below). Java was decided upon based upon the common experiences of the group members.

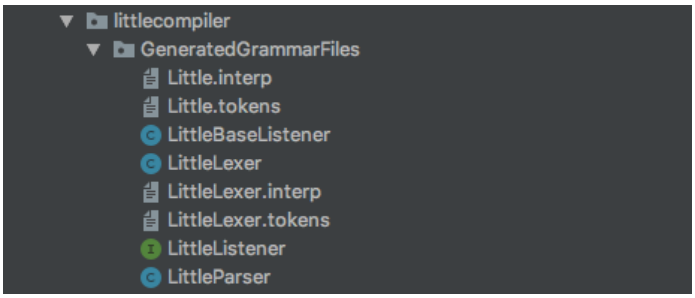


Figure 2: List of automatically generated files.

There were four files created in addition to those that were auto generated. The driver class was used as a start point. A factory class was written to help with dependency injection and to make unit testing of written files easier. The factory pattern was used to keep class coupling to a minimum. This type of coupling made the project more flexible and less fragile. The compiler class was written to aggregate all the tools used for the entire Compiler class project. The compiler class was created with two dependences, the lexer (or scanner) and the parser. The implemented method is for the scanner and was used to generate a file for holding the token output to be used later by the parser. The method for the parser was left as a stub since it was not needed for this part of the project. The last hand written class was token visualizer. This class was used to generate a string of all the tokens from a given program and then compared to a known output. There were five different programs used to test the operation of the scanner output and each difference was return with a value of zero. The two auto generated classes used were LittleLexer.java and LittleParser.java. These names were automatically given from the name of the grammar used and the function of the class. Lastly, script files were made to aid in running the appropriate files to recreate the ANTLR files and then use the hand written driver file to automatically scan test files for grading.

There were a few difficulties in accomplishing this first part of the compiler project. The first came from learning the nuances of the regular expressions that ANTLR will accept. For example, in the figure below the regular expressions for a string literal and use the tilde symbol (~) as NOT. This was not listed at either http://www.rexegg.com/regex-quickstart.html or https://www.shortcutfoo.com/app/dojos/regex/cheatsheet and was eventually arrived at through trial and error. So, ~[‘ ‘] means ‘NOT nothing’ or ‘anything’ is acceptable as a string literal and a comment.

```
6   D:          [0-9];              // digit
7   AN:         [a-zA-Z0-9];        // alphanumeric
8   ID:         [a-zA-Z]AN*;        // identifier
9   INTLIT:     D+;                 // integer literal
10  FLOATLIT:   D*.D+;              // float literal
11  STRINGLIT:  '"' ~['']* '"';     // string literal
12  COMMENT:    '--'~['']* '\n';    // comment
13  WS:         [ \t\r\n]+ -> skip; // skip spaces, tabs, carrage returns, and newlines
```

Figure 3: Identifiers from the grammar Little.g4

Another hurdle to get past was learning the ANTLR API. This took time to learn the new tool but was made a little easier from experience using other development environments and APIs. The automatically generated files were not exactly human readable and took a while to wade through along with searching for and reading through ANTLR documentation. This in turn helped to solve another challenge, how to pass files to the parser and then, in turn, pass tokens to the parser. The ANTLR libraries provided options and methods to help keep the hand written code simpler and easier to manage. The last item to work through was how ANTLR handled scanning and parsing. In class this was handled as two different but related steps in building a compiler. ANTLR, however, generates both of these components at the same time, which made it more difficult to distinguish what exactly need to be done for this part of the project.

C. Parser

According to Merriam-Webster the definition of parsing a grammar is "to divide (a sentence) into grammatical parts and identify the parts and their relations to each other" (https://merriam-webster.com). A parser is a program that acts as the interpreter part of a compiler that analyses a list of tokens by breaking them into smaller pieces and finds meaning in them that conforms to a set of rules defined by the grammar. Meaning can be made of these tokens by creating a parse tree of all of the tokens and their groupings. There are two stages involved in the parsing process: syntactical analysis, and semantic parsing. Syntactic analysis looks for a meaningful expression from the tokens generated by the lexer. The definition of the algorithmic procedures to create components comes from the usage of a context-free grammar. The placement of tokens must happen in a meticulous order and each of the rules in the grammar work to create an expression that has the proper syntax. This does not mean that the accepted string will have any meaning. The semantic stage of parsing determines the meaning and implications of a valid expression. There are two ways in which data from the input string of tokens can be derived from the start symbol: top-down parsing, and bottom-up parsing. Bottom-up parsing takes the entire list of tokens and tries to assemble a parse tree in reverse, working from the leaves to the start symbol. Top-down parsing, which is the way ANTLR works (see figure below), begins with the start symbol and works from left to right to find all the expressions that fit the given grammar. (https://www. techopedia.com)

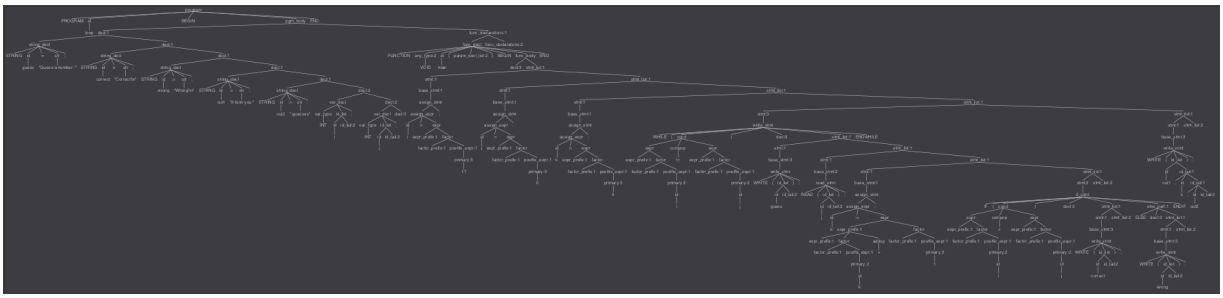


Figure 4: Generated parse tree.

There were a few changes to our code base for this stage of the project. First of all we needed to create a method generate a string of valid tokens that wasn't printed out this time. This token stream was then to be fed into the parser as input. Before we could actually feed this stream to the parser, the LittleGrammar.g4 class needed to have all the productions added to it. This was a rather involved process since order and naming of non-terminals matters. A small sample of these productions can be seen below.

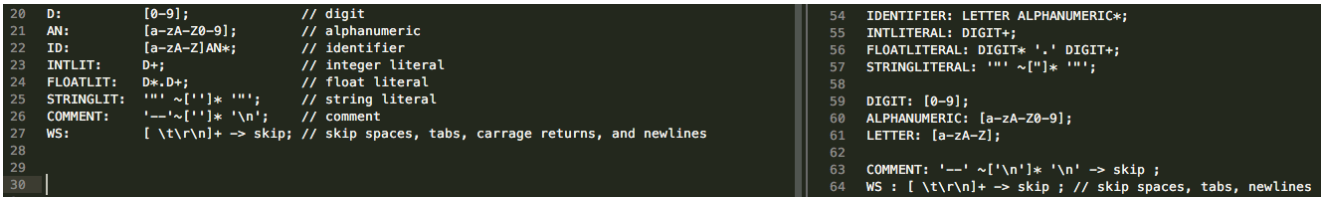
```
3   program: 'PROGRAM' id 'BEGIN' pgm_body 'END';
4   id: IDENTIFIER;
5   pgm_body: decl func_declarations;
6   decl: string_decl decl | var_decl decl |;
7
8   string_decl: 'STRING' id ':=' str ';';
9   str: STRINGLITERAL;
10
11  var_decl: var_type id_list ';';
12  var_type: 'FLOAT' | 'INT';
13  any_type: var_type | 'VOID';
14  id_list: id id_tail;
15  id_tail: ',' id id_tail |;
```

Figure 5: Sample of grammar productions.

We could then make the necessary addition to the driver to call the parser with the output string of tokens from the lexer. The sole output was to print that the input was either accepted or rejected. In order to accomplish this task we added a new class for error identification and handling. The factory class was also updated to accommodate this new dependency on the error strategy we selected.

Again, there were several roadblocks to overcome in finishing this part of the compiler project. First off, with the addition of productions to our grammar file, we found that there were needed changes to the original identifier list (seen in the figure below on the left). One identifier was added to the file and the names were changed to become more human readable. This also made it easier to add these to the different productions without needing to lookup what abbreviation belonged to each of the identifiers. The last change was in how to identify a comment. Originally a comment could contain anything in it and ended with a new line identifier (\n). This became problematic since the first new line encountered at the end of a line of code was turned into a comment. Simply excluding the

new line symbol from the body of the comment identifier changed this undesirable behavior.



```
20  D:          [0-9];              // digit
21  AN:         [a-zA-Z0-9];        // alphanumeric
22  ID:         [a-zA-Z]AN*;        // identifier
23  INTLIT:     D+;                 // integer literal
24  FLOATLIT:   D*.D+;              // float literal
25  STRINGLIT:  '"' ~['"]* '"';     // string literal
26  COMMENT:    '--'~['']* '\n';    // comment
27  WS:         [ \t\r\n]+ -> skip; // skip spaces, tabs, carrage returns, and newlines
28
29
30
```

```
54  IDENTIFIER: LETTER ALPHANUMERIC*;
55  INTLITERAL: DIGIT+;
56  FLOATLITERAL: DIGIT* '.' DIGIT+;
57  STRINGLITERAL: '"' ~["]* '"';
58
59  DIGIT: [0-9];
60  ALPHANUMERIC: [a-zA-Z0-9];
61  LETTER: [a-zA-Z];
62
63  COMMENT: '--' ~['\n']* '\n' -> skip ;
64  WS : [ \t\r\n]+ -> skip ; // skip spaces, tabs, newlines
```

Figure 6: Changes to the identifiers in the grammar.

The next challenge came in trying to change the non-terminal names of some of the productions to make them more readable. However, this quickly cascaded into chaos and the only option was to go back to the naming convention given in the reference text file for the grammar. Even though it had been mentioned in class, we found that disordering productions caused havoc when trying to some of the test files. The challenge was trying to determine a proper order. This was mostly done by carefully reading through each of the productions and a little bit of trial and error. The last obstacle was perhaps the largest. In order to decide whether or not the input would be accepted or rejected we had to find some way to handle errors in parsing. There didn't seem to be any way of catching errors since the lexer output a string of all tokens, valid or not. The parser would also work its way through the entire string fed to it even though there were marked errors in the output. After trying to catch these errors for both the lexer and parser and failing, it was decided to look for some hints in outside resources. This proved fruitful. The solution was to override some methods in one of the ANTLR libraries to do the error handling (https://stackoverflow .com/questions/39533809/antlr4-how-to-detect-unrecognized-token-and-given-sentence-is-invalid). Time to explore this solution was necessary to understand exactly what errors were being caught and how. The solution was finally modified to catch only the error we needed to either accept or reject the given input.

D.Symbol Table

A compiler creates and maintains a data structure called a symbol table so that information can be stored about occurrences of various items in the language "such as variable names, function names, objects, classes, interfaces" and the like (https://www. tutorialspoint.com/compiler_design/compiler_design_symbol_table.htm). A compiler will use this data structure in both synthesis and analysis. Depending on the language used, this table may serve several different purposes. These include variable declaration verification, type checking (semantic correctness of expressions and assignments), scope resolution of a name, and centralized and structured names of all entities. This symbol table can take the form of a linear list, linked list, binary search tree or a hash table. For small languages a linear or linked list is adequate since traversing this list wouldn't take too much time or

computing power. For larger languages a hash table or binary search tree is necessary for reducing time (O(1) and O($\log_2$ n) respectively) and cutting down on the usage of the CPU (https://www. geeksforgeeks.org/symbol-table-compiler/). The information stored in the table about each symbol is its name, type, and attribute and can be formatted as follows:

<symbol name, type, attribute>

The first decision to be made was to use either the default setting (listener pattern) for ANTLR or choose the visitor pattern. The decision was made based on the pros and cons of each pattern. The visitor pattern is good for using the parser output directly for interpretation. Tree traversal is in full control of the user, meaning only one branch is visited in conditionals and ***n*** number of visits can be made in loops. The visitor pattern also seems to be more flexible. Both pattern have their virtues when the input is translated to a lower level such as instructions on a virtual machine. In his book, *The Definitive ANTLR 4 Reference* (Terence Parr 2012, https://media.pragprog .com/titles/tpantlr2/listener.pdf), Terence Parr states:

> *"The biggest difference between the listener and visitor mechanisms is that listener methods are called by the ANTLR-provided walker object, whereas visitor methods mush walk their children with explicit visit calls. Forgetting to invoke visit() on a node's children means those subtrees don't get visited."*

Another important deciding factor between these two patterns is that the visitor uses the *call stack* and the listener used an *explicit stack*. The explicit stack is "allocated on the heap" and managed by the walker provided by ANTLR (https://stackoverflow.com/questions/20714492/ antlr4-listeners-and-visitors-which-to-implement#30762056). This means that for large inputs the visitor pattern could run into stack overflow issues but the listener pattern wouldn't have a problem. The main deciding factor came with the amount of management it would take to use the visitor pattern. With the listener pattern a programmer is only interacting with the provided tree walker. This seemed like the least amount of added headache for the most gain even though the language is small. The second decision was to come to a consensus on what type of data structure to use in making the symbol table. Since this is a little language a stack was decided upon for this purpose. Three changes to the grammar file became necessary at this stage in the project. As seen below in Figure 1, *func_declarations:* recursively refers to itself or an empty string. The caused a new function declaration to be created after an existing declaration that was empty. The second line in the figure shows the change to fix this issue.

```
21 func_declarations: func_decl func_declarations |;
22 func_declarations: func_decl func_declarations | func_decl;
```

Figure 7: Change to func_declarations.

The second change to the grammar file was for the *if_stmt*. An *if* statement doesn't necessarily need to have an *else* block to follow it. The change made was so that the *else* block would be optional, as seen in Figure 2.

```
47 if_stmt: 'IF' '(' cond ')' decl stmt_list else_part 'ENDIF';
48 else_part: 'ELSE' decl stmt_list |;
47 if_stmt: 'IF' '(' cond ')' decl stmt_list else_part 'ENDIF' | 'IF' '(' cond ')' decl stmt_list 'ENDIF' ;
48 else_part: 'ELSE' decl stmt_list;
```

Figure 8: Change to if_stmt and else_part.

The third change was in the *pgm_body* declaration. Since one of the test files contained a body declaration but no function declarations, an adjustment to the grammar file was necessary here (see Figure 3).

```
5 pgm_body: decl func_declarations;
5 pgm_body: decl func_declarations | decl ;
```

Figure 9: Change to pgm_body.

There were six total new classes used for part three of the project. The first was Symbol.java to represent a symbol held in a table. The next was SymbolTable.java; this holds all the symbols with a single scope. The class SymbolTables.java links all the tables in their relative scopes. The last of the table classes was SymbolTableVisualizer.java. This class recursively traversed all of the symbol tables in a tree structure while printing each of the table names and contents. Two enumerator classes were also added to support the symbol classes. The first held the possible attributes of a symbol (int, float, string, or void). The second was for the two possible types of a symbol (var or procedure).

The difficulties for this portion of the project came in the form of a few smaller hurdles to get past. One of the example output files listed the attribute for a procedure to be an *int* but it didn't seem to be an integer. The issue ended up being reading the return type from the procedure. For the Little language a procedure can either return an integer value (*int*) or return nothing (*void*). So either will show up depending on the whether the procedure returns anything or not. Another small issue that needed to be hunted down was that all child tables added to the *global* table and beyond were added twice. This turned out to be a minor oversight. It was thought that children needed to be added manually but when the SymbolTable.java class was called with a new symbol it automatically added it as well. Once found, this was a simple matter to correct. Another thing that was realized at this point in the project was that not all the methods in the LittleBaseListener.java class needed to be filled out. Only the methods that dealt with variables, procedures, integers, floats, strings, and voids.

D.Code Generation

Although optimizations can still be applied, the generation of assembly language or machine code "can be considered as the final phase of compilation" (https://www.tutorials point .com/compiler_design/compiler_design_code_generation.htm). Optimizing the generated code wasn't a part of the absolute requirement for this project it is often seen as part of the code generation step and process. From the first step to the last, the entire process should minimally have the following properties:

- The value of the source code should be exactly represented.
- The management of memory and usage of the target CPU ought to be efficient.

Keeping the two points above in mind, one type of optimization is called peephole optimization. This technique is easiest to apply locally within a basic block of code. Several statements will be analyzed and checked for several possible optimizations. One is the elimination of redundant instructions. There may be multiple instructions for loading and storing values that end up having the same meaning even when some are removed. A second way to optimize is the elimination of code that's unreachable. If, for example, a method is to print a certain line but breaks or returns before that statement is reached then it can be removed from the generated code. Optimizing flow control is another way to improve performance. For example, the following figure has unnecessary jumps that accomplish nothing as far as the compiler is concerned.

```
...
MOV R1, R2
GOTO L1
...
L1 :   GOTO L2
L2 :   INC R1
```

Figure 10: Example of code with an unnecessary jump.

It is easy to see here that the first *GOTO* statement can easily be changed to go directly to *L2* and completely remove the intermediate jump. The simplification of algebraic expressions is another place to make simple changes to streamline the code generation process. Adding zero to a value is an unnecessary process altogether and adding or subtracting by one can be replace by a simple *increment* or *decrement* statement. Strength reduction involves replacing operations with ones that will manufacture the same result but burn up less space and time to process. Also, the instructions held on the target machine can deploy instructions that are more sophisticated and be much more efficient in performing specific operations. This will not only produce results in a more proficient manner but also improve code quality. Additionally, the registers in the processor that variable must be allocated to should be taken into consideration. If the compiler is to truly

be useful to a developer then the generation of debug data will also be required (https://en.wikipedia.org/wiki/Code_generation_%28compiler%29).

An optional yet extremely useful part of code generation is the development and use of a syntax tree. This is actually just a parse tree but in a condensed form. The syntax tree takes out all the names and types of variables and declarations and creates an easy to read tree with just the essentials in it (see Figure 11).

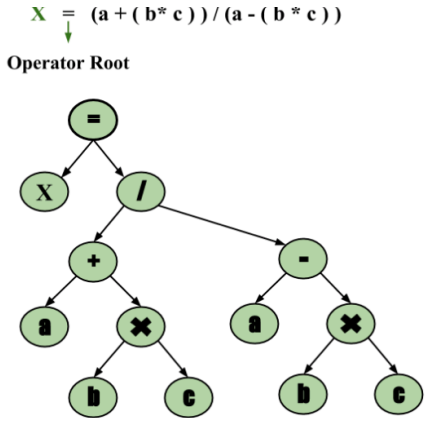


Figure 11: Syntax tree representing the assignment of an algebraic expression to *X*.

From the syntax tree it is much easier to generate three-address code working from the leaves back up to the root node. Three-address code is a linear representation of the code received from the semantic analyzer. This involves making a statement that contains three or fewer references, one for the result and two for the operands. Occasionally a statement will have less than three references in it but will still fall under the name of a three-address statement. There are two forms that a three-address code can be represented as: quadruples and triples. For example, take a statement such as *a = b + c * d;* and turn it into the following three-address statements.

```
r1 = c * d;
r2 = b + r1;
a = r2
```

Figure 12: Three-address statement for *a = b + c * d;*

The quadruple representation of this is separated into four different fields: the operator, two arguments, and the result (see Figure below).

Op	arg_1	arg_2	result
*	c	d	r1
+	b	r1	r2
+	r2	r1	r3
=	r3		a

Figure 13: Quadruple representation of three-address statements.

Triples representation, as the name implies, only has three fields: the operator and two arguments. Triples, however, "face the problem of code immovability while optimizing, as the results are positional and changing the order for position of an expression may cause problems" (https://www.tutorialspoint.com/compiler_design/compiler_design_intermediate_code_generations.htm).

Changes made include minor edits to the Little.g4 grammar file to make the three-address code a little easier to generate from the nodes in the abstract syntax tree. Thirty-two node classes were made to actually generate the appropriate intermediate three-address code. All of these classes inherited from one ASTNode.java class. Two Singleton classes were written to keep track of the labels for *if*, *else*, and *while* blocks of code and the temporary registers (*$T4*) for storing values to. The third Singleton class to be created was ParameterRegisterHandler.java to keep track of value registers that would be passed to and from functions. A fourth Singleton class, FunctionCodeGenerator.java, was written to make sure functions each had their own identifier and could be kept track of. The AST.java class more or less sits at the top of this part of the project, keeping track of nodes and children in the abstract syntax tree. The class TACLine.java takes all of the individual pieces of the three-address code generated in each of the node classes and concatenates them into strings that will then be translated into Tiny code. One working class and one test class were also created to take the three-address code and convert it to Tiny code as the last step.

Difficulties overcome in this part of the project are as follows. Nested and recursive statements needed to be changed and debugged so that the three-address code would generate properly and keep the correct register number throughout the life of the variable. Mapping from the given *.micro files to three-address code was a bit different from the examples in class and ended up taking a little while to see how the loops were labeled and taken care of. Going from three-address code to Tiny code also had similar difficulties because there was not a one to one mapping between the two codes. Another challenge was matching out put generator values to given outputs and freeing the proper registers for reuse. Also, learning all the syntax differences between conditional branching instructions was not trivial.

F.Full-fledged Compiler

All of the parts of this project built upon one another in a fairly seamless manner. Overall project planning was able to be accomplished since the over view of the entire project was explained early on in the class and was reiterated for each of the different parts. This allowed for forethought to be made for the next portion of the project while designing and writing the current portion. The factory pattern was used for the first two parts of the project to create the token stream. A logic and code was built onto this base for the second

two parts. Before any real programming began, it was decided that GitHub would be used to bring all the parts written by each of the group members together and as excellent version control tool. As far as day-to-day communication between members, a group text message was started in the first week of class and maintained throughout the semester.

4. Conclusion and Future Work

While each of the sections of the project was completed by each deadline and in working order, time was not a commodity in surplus. Had this not been a constraint better coding practices would have been employed. This would include common code reuse, more opportunities for inheritance and better use of design patterns. Optimization was discussed in class but was optional for the project. This was a welcome relief as the project was done just on time with the requirements as they were. However, it would have been a constructive challenge to use something like peephole optimization and see how much the code could have been reduced and what the clock cycles may possibly have been brought down to. Along with this would be to see what minimum number of registers could be used and still meet the requirements of three-address code generation. An exploration of many of the coding possibilities with the Tiny language and what that might have done to the compiler would be well worth the time to investigate along with what actual programs could be written in this language.