University of Brasília

九　尾　狐　の　語

K Language

Polymorphism in a C-like programming language

Diogo César Ferreira

11/0027931

December 10, 2019

Contents

1 The language

1.1 Introduction

Programming languages are languages that can be converted into sets of instructions to be executed by a computer. This conversion process is called translation (or compilation) and is done by software known as compilers. As programming languages evolved, we observe an increasing variety of abstractions to accommodate for new programming paradigms and techniques, which also bring them closer to the domain of the problems they are set out to solve and away from the specificities of architecture set implementations [1].

One such abstraction is polymorphism. In typed languages, symbols (named identifiers for entities such as variables or functions) are constrained by their data type. If polymorphism is available, we are able to interact with different data types through a single interface, for instance, by allowing multiple types to be assigned to a given symbol. Polymorphism improves code readability and helps keeping the namespace clean by allowing functions to be identified by their expected behavior rather than contrived by the types of their arguments. Polymorphism is also one of the key features in object-oriented programming [1, 2].

Aside from the possibility to perform arithmetical operations over a few different data types (integers, floating point and pointers), the C programming language does not support polymorphism. In this work we seek to provide polymorphism in the form of function overloading - where polymorphic functions may have multiple definitions according to their argument types - to a simplified subset of the C programming language. The grammar of the language is presented in the following subsection.

1.2 Lexical features

The following table summarizes what word patterns are part of the language we are designing. These patterns should be recognized by the scanner, and therefore are provided as an input file for the FLEX [3] scanner generator (see attached file "language.l"):

Data types	void	char	int	float			
Keywords	if	else	do	while	return	asm	
Arithmetical operators	+	−	*	/	%	++	−−
Comparison operators	<	>	<=	>=	==	!=	
Logical operators	&&	||	!				
Assignment operator	=						
Character/string delimiters	'	"					
Array indexers	[	]					
Scope delims./init lists	{	}					
End of statement mark	;						
Function args/calls	(	)					
List element separator	,						
Comment delimiters	/*	*/	//				

Character constants	text delimited by single quotation marks (')
String literals	text delimited by double quotation marks (")
Integer decimal constants	e.g. 1, 2, 100, -5, +7
Integer hexadecimal constants	e.g. 0x01, -0xA, +0x123
Floating-point constants	g.g. 0.0, -1.234, +3.14
Comments and white space	are ignored.

1.3 Language syntax

Our language's grammar contains elements taken from the one presented by Harbison III and Steele [4], which is a compilation from the many versions of the C grammar specified by the ISO C Standard along the years. Our grammar follows the general structure of the C language's grammar. We aim to provide a simplified version of the C language with support for polymorphic function definitions, that is, which allows the programmer to define multiple functions with the same name but different argument types. The following grammar is also provided as an input file for the GNU Bison [5] parser generator (see attached file "language.y").

1. declaration-list → declaration
   | declaration-list declaration

2. declaration → function-definition
   | init-declarator **;**

3. init-declarator → var-declarator
   | var-declarator **=** assignment-expression
   | arr-declarator **= STRING-LITERAL**

4. var-declarator → type **IDENTIFIER**

5. arr-declarator → type **IDENTIFIER** [ ]

6. function-definition → function-declarator **{** statement-list **}**
   | function-declarator **{ }**

7. function-declarator → type **IDENTIFIER (** argument-list **)**
   | type **IDENTIFIER (VOID)**

8. argument-list → argument
   | argument-list **,** argument

9. argument → type **IDENTIFIER**
   | type **IDENTIFIER** [ ]

10. compound-statement → **{ }**
    | **{** statement-list **}**

11. statement-list → statement
    | statement-list statement

12. statement → **;**
| init-declarator **;**
| assignment-expression **;**
| conditional-statement
| iteration-statement
| compound-statement
| return-statement **;**
| inline-asm **;**

13. inline-asm → **ASM (STRING-LITERAL)**

14. conditional-statement → if-statement else-statement

15. if-statement → **IF** (assignment-expression) compound-statement

16. else-statement → **ELSE** compound-statement
| ε

17. iteration-statement → **WHILE** (assignment-expression) compound-statement
| **DO** compound-statement **WHILE** (assignment-expression) **;**

18. return-statement → **RETURN**
| **RETURN** assignment-expression

19. assignment-expression → logical-or-expression
| postfix-expression = logical-or-expression

20. logical-or-expression → logical-and-expression
| logical-or-expression || logical-and-expression

21. logical-and-expression → equality-expression
| logical-and-expression **&&** equality-expression

22. equality-expression → relational-expression
| equality-expression == relational-expression
| equality-expression **!**= relational-expression

23. relational-expression → additive-expression
| relational-expression < additive-expression
| relational-expression > additive-expression
| relational-expression <= additive-expression
| relational-expression >= additive-expression

24. additive-expression → multiplicative-expression
| additive-expression + multiplicative-expression
| additive-expression − multiplicative-expression

25. multiplicative-expression → unary-expression
| multiplicative-expression ***** unary-expression
| multiplicative-expression / unary-expression
| multiplicative-expression % unary-expression

26. unary-expression → postfix-expression
| **!** unary-expression
| − unary-expression
| − − unary-expression
| ++ unary-expression

27. postfix-expression → primary-expression
| **IDENTIFIER ( )**
| **IDENTIFIER (** argument-call-list **)**

28. primary-expression → **IDENTIFIER**
| **CONSTANT-INT**
| **CONSTANT-HEX**
| **CONSTANT-FLOAT**
| **CONSTANT-CHAR**
| **STRING-LITERAL**
| **(** assignment-expression **)**

29. argument-call-list → assignment-expression
| argument-call-list **,** assignment-expression

30. type → **VOID**
| **INT**
| **FLOAT**
| **CHAR**

1.3.1 Grammar changes

Rule 1: *declaration-list* is the start symbol once more. We now use Bison's `%destructor` to capture the syntax tree's root.

Rules 3, 4 and 5: the *init-declarator* has been broken down into rules 4 and 5 to avoid the need to write many possible combinations of variables in these rules. Initializer lists were removed from this because integer and float arrays have not yet been implemented (the only implemented arrays are strings).

Rules 6 and 7: *function-definition* was broken down into rules 6 and 7 to avoid the need to write many possible combinations of variables in these rules. Function without arguments now need to be declared with "void" as the sole argument.

Rules 12 and 13: new statement *inline-asm* (rule 13) was created and included in the list of possible *statement*s (rule 12). That new rule allows for direct translation of string literals into the generated code (see Section 2.6).

Rules 14, 15 and 16: *conditional-statement* has been broken down into its *if-statement* and *else-statement* counterparts. This was done to resolve a reduce/reduce type conflict that would arise due to a mid-rule action. The *else-statement* also produces the empty string (ε), so that it is still possible to use an *if* without an *else*.

Rule 27: postfix array indexing operator has been left out from this version since array indexing are not yet implemented.

1.3.2 Input / output

Input and output operations are not specified in our language's grammar. However, these are be provided in the form of a small standard I/O library composed of special polymorphic functions which are always available to the user. The available functions have the following signatures:

```
void write(int i){...}    // outputs an integer to the standard output
void write(char c){...}   // outputs a character to the standard output
void write(char[] s){...} // outputs a string to the standard output
void write(float f){...}  // outputs a float to the standard output
void write(void){...}     // outputs a new line to the standard output

int read(int i){...}     // reads an integer from the standard input and returns it
char read(char c){...}   // reads a character from the standard input and returns it
float read(float f){...} // reads a float from the standard input and returns it
```

The `write` functions output the value provided to them as argument. They do not otherwise modify the provided value. Example use:

```
// output 3 to the standard output
write(3);
```

Since there are no pointers in the language, the `read` functions require an argument in order to resolve which data type will be read and returned (that is, which function will be used to read from the standard input). However they do not in any way use or modify the provided argument. Example use:

```
// read an integer from the standard input and
// store it in variable a. The value of a is
// not used within the function read.
int a = read(a);
```

As with any other function, `read` and `write` are polymorphic and the users are allowed to define their own versions of these functions, provided that they do not have the same parameters as the ones already defined. We simulate linking of the library to user code by creating a temporary file in which the library is concatenated with the input file.

1.4 Semantics

1.4.1 Scope

Except in the case of initializer lists, scopes are enclosed by curly braces ('{' and '}'). Every program should have at least one scope, which is the global scope, where the programmer is

only allowed to declare and initialize (but not assign) global variables and to define functions. Functions can be defined only in the global scope.

The body of a function is treated as a compound statement itself and compound statements can be nested. Within compound statements, there can be variable declaration, initialization and assignment, arithmetic and logical expressions, flow control statements (conditional and loops) and jump statements (return). Since variable assignment can only be done within compound statements and these are not allowed in the global scope, then a program must have at least one function for variables to be able to be assigned. Variables and functions must always be declared, initialized or defined before they can be used.

Declaration of variables with the same name in a given scope is not allowed. However, a variable declaration is allowed to shadow another variable in a parent scope. When a variable must be used, a declaration of that variable will be first searched in the in the current scope. If not found, it will be searched in the current scope's parent scope and so forth up to the global scope.

Unlike in the C programming language, definition of functions with the same name is allowed, provided that the functions have different argument types. Definition of functions with the same name and same argument types is not allowed, even if the return type differs. The argument variable names pertain to the outermost scope enclosed by the function's body.

1.4.2 Implicit type conversion

Our language is comprised of five basic types: `void` (no type), `char` (characters), `int` (integers), `float` (floating point numbers), and pointers in the form of arrays of one of the basic types: `char[]` (strings), `int[]` (arrays of integers), `float[]` (arrays of floating point numbers). Notice however, that among the array types, only strings have been implemented in this version (see Section 2.6). The rules for conversions are as follows:

- **Same type** when values are of the same type, no conversion is needed;
- `void` cannot be implicitly converted to another type;
- `char` can be implicitly converted to `int` or `float`;
- `int` can only be implicitly converted to `float`;
- `float` cannot be implicitly converted to another type;
- **Arrays** cannot be implicitly converted to another type.

Essentially, a value can be implicitly converted to another type with larger storage size. For function calls, the same rules should apply after the function's return value has been resolved. However, types are not implicitly converted on the arguments of a function call.

1.4.3 Explicit type conversion

Since there are no casts in the language, there can be no explicit type conversion.

1.4.4 Expressions and static evaluation

Most of the language's operators are left-associative with the exception of unary operators and the assignment operator, which are right-associative. Expressions are first type checked for incompatibilities. If no type errors are found and the expressions are composed only of constants, then they will be statically evaluated.

1.5 Use-cases

We present some use cases in which polymorphic functions would be useful. Functions with the same behavior, but different argument types:

```
int   i = -1;
float f = -1.0;
abs(i); // return the absolute value of i
abs(f); // return the absolute value of f
        // In C, abs does not accept a floating-point argument
        // we must use fabs instead
```

In this simple example, the abs function could compute the absolute of any numeric value passed as argument. The type constrains in C, however, do not allow the abs function to be called with a floating-point argument.

Functions with different behavior, but the same "intuitive meaning" for the programmer

```
int  i[3] = { 9, 2, 5 };
char w[][9] = { "word", "vocable", "locution" };

sort(i); // The user implements a sorting function for integer vectors
sort(w); // The user implements a lexicographic or alphabetic
         // sorting function for string vectors
```

The implementation of a sorting function for integers and strings would be considerably distinct. However, intuitively a programmer might expect some notion of ordering to be present in arrays of strings. This can also be applied to functions with different number of arguments:

```
// Should return the lowest of either x or y
int min(int x, int y) {
    if (x <= y) { return x; }
    else { return y; }
}

// Should also return the lowest of either x, y or z
int min(int x, int y, int z) {
    if (x <= y) {
        if (x <= z) { return x; }
        else { return z; }
    } else {
        if (y <= z) { return y; }
        else { return z; }
    }
}
```

2 The compiler

In this section we will describe the compiler's implementation for the K Programming Language, which has the responsibility to read the input text and ultimately convert it into an executable file. The first step is to define the patterns which are to be recognized by the scanner. Then, the Fast Lexical Analyzer (FLEX) [3] software is used to generate the lexical analyzer's main

functionality. Moreover we provide the language's grammar to the GNU Bison parser generator [5] which generates the syntax analyzer. Additional routines for input, output, error handling and data structures to build the syntax tree and symbol table were also implemented and will be explained in detail later.

2.1 Program usage

The program takes one argument which should be a path to the input file.

```
$ ./kyu <input_file>
```

A file named k.tac will be created in the same directory as the program, which contains the set of translated instructions from the input file. The program will also show the symbol table produced from the input file if it is part of the language. Should there be errors, the program will instead list them along with their positions in the text and an output file will not be created.

2.1.1 Compilation instructions

1. (Optional) Generate the lexical analyzer .c source code through FLEX

   ```
   $ flex language.l
   ```

2. (Optional) Generate the syntax analyzer .c source code through GNU Bison

   ```
   $ bison language.y
   ```

3. Compile the program

   ```
   $ gcc -okyu -std=c18 -m64 -Isrc src/main.c src/node.c src/node-list.c src/table.c src/
       table-stack.c src/arg-list.c src/parser.c src/scanner.c src/actions.c src/typechecker/
       base.c src/typechecker/unary.c src/typechecker/binary.c src/misc.c src/lines.c src/
       error.c src/generator.c
   ```

4. Alternatively, a makefile has also been provided which will run all commands

   ```
   $ make
   OR
   $ make flex    # for step 1
   $ make bison   # for step 2
   $ make program # for step 3
   ```

System specifications

Compiler version: gcc (Debian 9.2.1-17) 9.2.1 20191102

FLEX version: flex 2.6.4

GNU Bison version: bison (GNU Bison) 3.4.2

2.2 The symbol table

We use a single hash table with linked lists as buckets, defined in `table.h` and `table.c`, to implement the symbol table. If performance becomes an issue due to high load factor, the table can be rehashed to decrease its load factor. Each entry in the symbol table is comprised of a key and its attributes. The following attributes are currently being stored:

- **Symbol type**: indicates if the symbol is a variable, a function or some other abstract structure such as a scope enclosed by a compound statement;
- **Return type**: stores a function's return type or a variable's data type;
- **Value**: an union that stores a symbol's current value, if applicable;
- **Argument list**: a structure that stores a function's argument list, if there is one;
- **Pointer to node**: a pointer to a node in the syntax tree that represents a statement's body
- **Scope**: a pointer to this symbol's enclosing scope, which itself is an entry in some other symbol table.

As per current implementation, the data structure used for a symbol table is the same data structure used for a symbol itself. Therefore, each entry in the symbol table is also a symbol table. The reason for this is to make the implementation of some auxiliary data structures simpler: there would be a pointer to a new symbol table in the attributes of a symbol anyway, since some entries such as functions and compound statements have their own child scope.

2.2.1 Polymorphic functions

Whereas variables are included into the table based solely on their names, the insertion of a function in the table uses a two-step hashing scheme: first, the function is hashed by its name and inserted into the table. This symbol is now a new symbol table itself. A key that represents the function's argument types is then constructed from its argument list, hashed and inserted into this new symbol table. Again, this symbol is another symbol table, which now holds the scope of the function's body, where the argument variable names are added to. For example, the following code:

```
int min(int x, int y) {}
int min(int x, int y, int z) {}
```

Would generate the following entries in the symbol table:

```
min  [ s_type=FUNCTION, r_type=void, value=(void), code=, temp=0, args=v ]  # function name
   iii  [ s_type=FUNCTION, r_type=int, value=0, code=, temp=0, args=iii ]   # arg types
      a  [ s_type=VARIABLE, r_type=int, value=0, code=#0, temp=0, args=v ]  # arg 1
      b  [ s_type=VARIABLE, r_type=int, value=0, code=#1, temp=0, args=v ]  # arg 2
      c  [ s_type=VARIABLE, r_type=int, value=0, code=#2, temp=0, args=v ]  # arg 3
   ii  [ s_type=FUNCTION, r_type=int, value=0, code=, temp=0, args=ii ]     # arg types
      a  [ s_type=VARIABLE, r_type=int, value=0, code=#0, temp=0, args=v ]  # arg 1
      b  [ s_type=VARIABLE, r_type=int, value=0, code=#1, temp=0, args=v ]  # arg 2
```

Thus, a function call would look up for a function in the current scope's symbol table, then builds the argument types key from the types resolved by the type checker and looks up for such an entry in the function's symbol table. In the present version, the type checker is unable to determine the types of function calls. However, insertions and redefinition errors are being correctly handled.

2.2.2 Scope resolution

The symbol table is also responsible for scope resolution issues. Whenever a variable declaration is found, the analyzer looks up for its name in the current scope. If it is found, a symbol re-declaration error is thrown, otherwise, the variable is added to the current scope's table. Whenever a variable is used, the analyzer looks up for its name in the current scope. If it is not found, the analyzer searches the parent scope and then its parent all the way up to the global scope or until a declaration is found. If none is found, an undeclared variable error is thrown.

2.2.3 Scope management

Scopes are stored withing a stack structures which elements are symbol tables. The stack is initialized with a single element that corresponds to the program's global scope. Whenever the parser finds a rule that would begin a new scope, then a new symbol, with an appropriately given key is inserted into the current context (top of the stack) and then pushed onto the stack, becoming the new element on the top, thus the new current context. When the parser finds the terminal token that ends a scope the parser pops the context from the stack, leaving its parent on the top. These actions are handled by the parser and are defined in the GNU Bison input file in rules such as this:

```
compound_statement
    : ...
    | '{'    { begin(NULL); }    statement_list '}'    { $$ = $3  ; finish(); }
```

Here, `begin(NULL)` is a mid-rule action whose only purpose is to create a new context and push it to the stack and `finish()` pops the stack when the context is closed.

2.3 Syntax tree

The syntax tree is build using the data structures defined in `node.h` and `node.c`. Most of the non-terminals in the grammar are set as pointer to our tree node type, to allow for the tree to be built as the parser proceeds. Identifiers, types and argument lists are interpreted respectively as strings, integers and lists. They are currently being handled in such a way that they need not be included in the tree.

```
%union {
    struct node * node;
    int ival;
    const char * sval;
    struct arg_list * al;
}
```

In the current version, the scanner no longer creates tree nodes. The scanner now appropriately sets `yylval` and only the parser handles the management of the nodes. The parser is instructed to either promote or to build new nodes for most grammar productions, building the syntax tree (Listings 1 and 2).

```
declaration
    : function_definition                   { $$ = $1; }
    | init_declarator ';'                   { $$ = $1; }
```

Listing 1: Example of parser rules where nodes are promoted. This procedure is preformed for productions in which there is no need to create a new node, since no new information would be gained in doing so.

```
additive_expression
    : ...
    | additive_expression '-' multiplicative_expression
    {
        $$ = nl_push(node_list, node_init(OP_SUB, "-", $1, $3, ENDARG)); // create a new node
        assign_context($$);
        typecheck_lazy($$);
    }
```

Listing 2: Example of a parser rule where a new nodes is created. Here, we will need to type check and evaluate both operands of the expression, before we know what attributes will be assigned to the head of the rule.

2.3.1 Syntax tree annotations

As nodes are created, they may be assigned to an entry in the symbol table. Some of these entries are actual symbols derived from identifiers, while others are surrogate symbols that represent each node's current state. Currently, these symbols are only being used to store the annotations in the tree, but in the future they will used as actually temporary symbols for the production of the 3-address code. Most of the annotations are currently being done by the type checker, as will be shown in Section 2.4. It is importante to note, however, that the type-checker prunes the nodes it has already analyzed from the tree. Therefore, while it is still possible to visualize the tree after the program has finished processing the input, it will not reflect the actual structure of the input text anymore.

2.4 Type checker

Every time the parser performs a reduction for a statement rule and creates a node in the tree, it is also type-checked (See Listing 2 for example). Type-checking is performed by the functions declared in the header "typechecker.h", and it is responsible for the following actions:

1. Verify whether operations and operands have compatible types;
2. If possible, perform static evaluation: if all operands in an expression are constants, evaluate it;
3. If not possible to perform static evaluation, generate the corresponding code;

The rules for type conversion during either static evaluation or code generation are the same as the ones presented in Section 1.4.2: `char` can be promoted to `int`, `int` can be promoted to `float` and other types cannot be part of expressions. In the middle of expressions, if either operand is able to be promoted it will be so. Therefore, unless the user tries to perform an operation with a `void` type or array, type errors will only be reported on assignments and function calls (arguments of function calls are never converted).

The type checker links the node it just analyzed to an entry in the symbol table, setting its attributes accordingly. However, in order to release resources, the type checker also prunes

a node's children from the tree after it has been successfully analyzed, because we no longer require the data held by the children at this point: either they have been statically evaluated or the corresponding code has already been generated.

2.5 Static evaluation

Static evaluation is performed by the functions expanded from the preprocessor macros defined in the header "evaluator.h". These macros define the expected behavior and appropriate type conversions for each arithmetic, logical and relational operators present in the language. The evaluator also sets flags for type errors that still may occur inside expressions, such as operations with `void` or arrays.

The program uses an internal operation stack for static evaluation. The operands are popped from the stack and, if both are constants, the operation is evaluated and the result is pushed back into the stack, which will then be available for the next time the parser reduces a rule and calls the type checker again.

2.6 Code generation

Our compiler's target language is the 3-address-code (TAC) implemented by Santos [6] which is itself an intermediate code based on the one presented by Alfred et al. [1]. Code is generated every time an operation cannot be statically evaluated, such as flow control statements or when variables are involved, because the compiler performs only a single pass through the input. Code generation for expressions also uses the same operation stack used by static evaluation, to assist in the management and reuse of temporaries created throughout the expression. We'll now present how each statement translates from our language to TAC instructions.

Statement	**K Lang code**	**TAC**
Inline assembler is copied directly into the output	`asm("// copy me");`	`// copy me`
Global variables are added to the TAC's symbol table uninitialized with the same name as the variable	`int globalvar;`	`.table` `int globalvar`
Strings are added to the TAC's symbol table initialized with a unique name.	`char string[] = "awooo";`	`.table` `char str_tNkEFwgQ1[] = "awooo"`

Function definition creates a label with the function's name and parameters.	`int f(int x) {`
	`  asm("// function body");`
	`  return x;`
	`}`
		`f_i:`
		`// function body`
		`  return #0`

Statement	K Lang code	TAC

Function parameters are assigned to a parameter identifier in the order in which they are declared	`int f(int a, int b) {`
	`  return a;`
	`  return b;`
	`}`
		`f_ii:`
		`  return #0`
		`  return #1`

Variable declaration the next available temporary in the scope is reserved to the variable	`void f(void) {`
	`  asm("// reserve $0 to a");`
	`  int a;`
	`  asm("// reserve $1 to b");`
	`  int b;`
	`}`
		`f_v:`
		`// reserve $0 to a`
		`// reserve $1 to b`
		`  return`

Initialization and assignment moves value to the reserved temporary.	`void f(void) {`
	`  int a = 10;`
	`  int b = 20;`
	`  a = 30`
	`}`
		`f_v:`
		`  mov $0, 10`
		`  mov $1, 20`
		`  mov $0, 30`
		`  return`

Function call push parameters, calls function and pops return value into a new temporary.	`int x = f(1, 2);`	`  push 1`
		`  push 2`
		`  call f_ii, 2`
		`  pop $1`
		`  mov $0, $1`

Expressions performs type conversion, if needed, then apply the equivalent corresponding.	`float f(int x) {`
	`  float y = 2.0*(x-1);`
	`  return ++y;`
	`}`
		`f_i:`
		`  sub $1, #0, 1`
		`  inttofl $1, $1`
		`  mul $1, 2.000000, $1`
		`  mov $0, $1`
		`  add $0, $0, 1.000000`
		`  return $0`

if creates two new unique labels and sets branch to the exit label if condition is false	`void f(int condition) {`
	`  if (condition) {`
	`  asm("// if-body");`
	`  }`
	`}`
		`f_i:`
		`  brz if_ZcM_XRpED_end0, #0`
		`if_ZcM_XRpED:`
		`// if-body`
		`if_ZcM_XRpED_end0:`
		`  return`

if-else creates three new unique labels and sets branch to the else label if condition is false. Also set jump to the exit label at the end of the if block	`void f(int condition) {`
	`  if (condition) {`
	`  asm("// if-body");`
	`  } else {`
	`  asm("// else-body");`
	`  }`
	`}`
		`f_i:`
		`  brz if_V6d_jeoe__end0, #0`
		`if_V6d_jeoe_:`
		`// if-body`
		`  jump if_V6d_jeoe__end1`
		`if_V6d_jeoe__end0:`
		`// else-body`
		`if_V6d_jeoe__end1:`
		`  return`

Statement	K Lang code	TAC

while creates two new unique labels and sets branch to the exit label if condition is false. Also sets jump back to the start label at the end of the block.	`void f(int condition) {`
	`  while (condition) {`
	`  asm("// while-body");`
	`  }`
	`}`
		`f_i:`
		`while_U0rdyI0Bj:`
		`  brz while_U0rdyI0Bj_end, #0`
		`// while-body`
		`  jump while_U0rdyI0Bj`
		`while_U0rdyI0Bj_end:`
		`  return`

do-while creates a new unique label and sets branch back to it if condition is true at the end of the block.	`void f(int condition) {`
	`  do {`
	`  asm("// do-body");`
	`  } while (condition);`
	`}`
		`f_i:`
		`do_F46jXJJv2:`
		`// do-body`
		`  brnz do_F46jXJJv2, #0`
		`  return`

2.7 Error handling

As the program reads the input file, the parser requests tokens from the scanner, who splits the input text into tokens and relays them back to the parser. If an error is found, a message is sent to `stderr` indicating the position (line and column) in the text where it was found, and whether it was detected by the scanner, the parser or the semantic analyzer.

2.7.1 Lexical errors and warnings

If the scanner finds an error, it outputs a message and returns a special error token to the parser. Following suggestions, identifiers with more than 32 characters are no longer considered an error and only output a warning if they are found, which does not block compilation. There is also a warning if an identifier begins with double underscores, as such identifiers are reserved for the compiler and may cause conflicts. The scanner recognizes the following errors (Table 1).

Table 1: Lexical errors and their associated tokens. The tokens are returned to the parser and the program continues analyzing the rest of the text.

Error type	Token type
Unrecognized token	`UNRECOGNIZED_TOKEN`
Malformed character constant	`INVALID_CHAR_CONST`

2.7.2 Syntax errors

Syntax errors are thrown from the parser when it is unable to complete a shift or reduce operation. To allow the parser to continue reading the input, the built in `error` token is included in some of the grammar's rules. Particularly, we have included such rules in order to capture:

- Malformed blocks of code (the parser continues after the next semicolon or enclosing curly brace)

```
declaration
    : ...
    | error ';'
    | error compound_statement
    ;
```

- Malformed statements (the parser continues after the next semicolon)

```
statement
    : ...
    | error ';'
```

- Malformed expressions (the parser continues after the next semicolon or enclosing parenthesis)

```
primary_expression
    : ...
    | error
```

- Malformed argument lists (the parser continues after the next enclosing parenthesis)

```
function_definition
    : ...
    | type IDENTIFIER '(' error ')'
```

2.7.3 Semantic errors and warnings

Semantic errors are thrown from the semantic analyzer when there are constructs in the input text that violate the language's semantics as described in the Section 1.4. These errors do not abort the compilation process before the parser has finished, as most of the semantic analysis is performed simultaneously with the parsing. Semantic errors are handled by the functions declared in the header "error.h".

Re-declaration of variable: This error is issued if a variable is declared or initialized more than once with the same name in a given scope. It is **not** an error to declare of initialize a variable with the same name of one that exists in a parent scope, but the new variable will shadow the old one, that is, the variable in the parent scope will be inaccessible.

Redefinition of function: This error is issued if a function is defined more than once with the same name **and** the same argument types. This language aims to implement polymorphic functions, therefore, defining functions with the same name and different argument types is allowed.

Variable use without declaration: This error is issued if a variable is used before being declared or initialized either in the current scope or any of its parent scopes up to the global scope. Variable declaration or initialization must always precede its use.

Function call without definition: This error is issued if a function is called but no function with that name has been defined. Functions must always be defined before they are used.

Function call with invalid argument types: This error is issued if a previously defined function is called with arguments types that are not present in any of the previous definitions of that function.

Incompatible types: This error is issued if there is an expression with operands of incompatible types, as described on Section 1.4.2.

Assignment to lvalue: This error is issued if the user tries to assign a value to a constant.

Integer division by zero: This error is issued if the user tries to assign perform an integer division by zero.

Function without return (warning): This warning is issued if the user does not define a top-level return statement inside a non-void function. A return instruction with a default value (0 for integers and characters, 0.0 for floats) is included in such case.

Program without main function (warning): This warning is issued if the user does not define any function with the name "main" in the program. The program will compile but the entry point will be undefined.

2.7.4 Remarks about error handling

These strategies allows us to display a reasonable amount of errors before the program is unable to proceed and terminates. In fact the parser should be able to finish (albeit not correctly) the syntax tree even when errors are found in several situations. The error messages are displayed as they are found, and since the scanner returns a token that is never used by the parser, lexical errors also generate syntax errors. To avoid encumbering the output when errors are found, only error messages and the symbol table are displayed on invalid input.

2.8 Tests

Eight sample test files are provided in the folder ./tests.

2.8.1 Valid input files

The folder ./tests/valid contains valid code that was used mostly for debugging during development. These files do not produce meaningful output, but can be used to verify the symbol table and generated TAC.

The other valid files are:

./tests/valid-sqrt.c numeric method to calculate the square root of positive numbers;

./tests/valid-strings.c tests for string initialization and output;

./tests/valid-sums.c tests for polymorphic function definition and function calls;

2.8.2 Invalid input files

error1.c is invalid and contain the following errors:

Line 1, column 25: syntax error, unexpected '*', expecting IDENTIFIER
Line 4, column 17: semantic error: undeclared symbol 'argc'
Line 4, column 22: lexical error, unrecognized token: '?'
Line 4, column 22: syntax error, unexpected UNRECOGNIZED_TOKEN
Line 4, column 26: lexical error, unrecognized token: ':'
Line 4, column 28: lexical error, invalid character constant.

error2.c is invalid and contain the following errors:
Line 1, column 1: lexical error, unrecognized token: '#'
Line 1, column 1: syntax error, unexpected UNRECOGNIZED_TOKEN
Line 1, column 16: lexical error, unrecognized token: ''
Line 6, column 9: warning, identifier exceeds 32 characters.
Line 6, column 52: warning, identifier exceeds 32 characters.
Line 8, column 9: warning, identifiers beginning in '___' (double underscore) are reserved and may c
Line 11, column 15: semantic error: redefinition of 'abcde'
Line 11, column 17: warning, identifier exceeds 32 characters.
Line 11, column 57: semantic error: undeclared symbol 'a234567890b234567890c234567890d2345678
Line 14, column 14: lexical error, invalid character constant.
Line 14, column 14: syntax error, unexpected INVALID_CHAR_CONST
Line 15, column 13: semantic error: incompatible types, 'void' return type 'int'

error3.c is invalid and contain the following errors:
Line 3, column 11: semantic error: redefinition of 'x'
Line 4, column 11: semantic error: redefinition of 'y'
Line 16, column 7: semantic error: undeclared symbol 'a'
Line 19, column 1: warning: non-void function 'min' does not have a top-level return statement
Line 21, column 23: semantic error: redefinition of 'min' with argument types 'ii'
Line 23, column 7: semantic error: undeclared symbol 'c'
Line 25, column 20: semantic error: undeclared symbol 'max'
Line 26, column 20: semantic error: there is no definition of 'min' with argument types 'iiii'
This file may also cause a segmentation fault while trying to release memory at the end of the program.

error4.c is invalid and contain the following errors:
Line 9, column 10: semantic error: cannot implicitly cast from 'int' to 'char'
Line 10, column 10: semantic error: cannot implicitly cast from 'float' to 'char'
Line 12, column 10: semantic error: cannot implicitly cast from 'float' to 'int'
Line 19, column 13: semantic error: incompatible types, 'char' return type 'int'
Line 23, column 1: warning: non-void function 'main' does not have a top-level return statement

error5.c is because it contains integer divisions by zero and invalid declarations, however the program is unable to finish parsing the file due to a segmentation fault.

Lines and columns are 1-indexed and the columns should point to where the error begins. Note, however, that tabs (\t) count as if they had a width of four characters.

2.9 Known issues

- "return" inside "if" or "else" causes the returned value to be 0

```
if (...) { return 1; } // actually returns 0
else { return 2; } // actually returns 0
```

Workaround:

```
int ret;
if (...) { ret = 1; }
else { ret = 2; }
return ret;
```

- Function call as the last parameter of another function call causes the the function not to be correctly identified in the symbol table

  ```
  f(a, b, c, g(x));
  ```

  Workaround:

  ```
  int y = g(x); // assign the call to an auxiliary variable
  f(a, b, c, y);
  ```

- Function call not at the beginning of an expression causes an assertion error (which, if removed will cause a segfault)

  ```
  int x = expr + func2(...);
  ```

  Workaround:

  ```
  int x = func2(...) + expr;
  ```

- Comparison between different types causes an assertion error (which, if removed will cause a segfault)

  ```
  int x = 1;
  float y = 2.0;
  x < y;
  ```

  Workaround:

  ```
  int x = 1;
  float y = 2.0;
  float xf = x; // declare an auxiliary variable of the correct type before comparison
  x < y;
  ```

- Some syntax errors might cause a segfault at the end of the program
- Division by zero semantic error causes a segfault during evaluation

3 References

[1] V Aho Alfred et al. Compilers, Principles, Techniques & Tools. Second Edition. Pearson Education, 2007.

[2] C Strachey. "Fundamental concepts in programming languages." In: 13 (2000), pp. 11–49.

[3] Paxson Vern et al. The Fast Lexical Analyzer. URL: `https://github.com/westes/flex` (visited on 10/19/2019).

[4] S.P. Harbison III and G.L. Steele. C, A Reference Manual. Fifth Edition. Prentice-Hall, 2002.

[5] The Free Software Foundation. GNU Bison. URL: `https://www.gnu.org/software/bison/` (visited on 10/19/2019).

[6] Luciano Santos. TAC - Simple three address code virtual machine. URL: `https://github.com/lhsantos/tac` (visited on 12/10/2019).