



an interpreted imperative programming
language that wants to believe

Lorenzo Loconte

Introduction

Shrimp is a very simple and didactic imperative programming language designed during the course of *Formal Methods for Computer Science* at Università degli Studi di Bari Aldo Moro. **Shrimp** uses an *eager evaluation strategy*. In order to ensure that, the interpreter executes the code using the *call by value* method.

Software Modules

The program is composed by three main components:

- The **parser**
- The **optimizer**
- The **interpreter**

The **parser** takes in input the source code and convert it into an intermediate representation. The intermediate representation have the structure of a n -ary tree having the non-terminals of the grammar as internal nodes and commands, identifiers and integer constants on the leaves.

The **optimizer** takes in input the intermediate representation given by the parser. The result of the optimizer is an *optimized* intermediate representation. It evaluates the constant expressions (both arithmetic and boolean) that might be present in the source code and replace them with the resulting constants. The optimizer also checks for empty command blocks and useless branch statements and removes (or optimize) them.

The **interpreter** execute the semantics present in an intermediate representation. The basic idea is to use a **state** (or environment) that collects the values of the variables during the execution of the program. The result of the interpretation is the resulting state, that is a set of ground integer assignments to the variables.

The Language Syntax

The syntax for the **Shrimp** programming language is a context-free grammar. So, it can be denoted using EBNF (Extended Backus Naur Form) as following:

```
1 Integer ::= [0-9]+
2 Identifier ::= [a-zA-Z_]+
3 Program ::= Header "shrimp" Block
4 Header ::= [VariableDecl]*
5 Block ::= [Command]*
6
7 VariableDecl ::= "let" Identifier "as" Type
8 Type ::= {"int" | "bool" | "array" "[" Integer "]" }
9 Command ::= {Assignment | Branch | Loop}
10
11 Assignment ::= Identifier
12               {"=" ArithmeticExpr | ":" BooleanExpr |
13               "[" ArithmeticExpr "]" "=" ArithmeticExpr} ";"
14
15 Branch ::= "if" "(" BooleanExpr ")" "then"
16           Block ["else" Block] "end if" ";"
17
18 Loop ::= "while" "(" BooleanExpr ")" "do"
19         Block "end while" ";"
20
21 ArithmeticExpr ::= ArithmeticTerm
22                  [{"+" | "-"} ArithmeticTerm]*
23 ArithmeticTerm ::= ArithmeticFactor
24                  [{"*" | "/" | "%"} ArithmeticFactor]*
25 ArithmeticFactor ::=
26   Integer
27   | Identifier
28   | Identifier "[" ArithmeticExpr "]"
29   | "-" ArithmeticExpr
30   | "(" ArithmeticExpr ")"
31
32 BooleanExpr ::= ArithmeticTerm ["or" ArithmeticTerm]*
33 BooleanTerm ::= ArithmeticFactor ["and" ArithmeticFactor]*
34 BooleanFactor ::=
35   "true"
36   | "false"
37   | Identifier
38   | "not" BooleanExpr
39   | ArithmeticExpr "eq" ArithmeticExpr
40   | ArithmeticExpr "neq" ArithmeticExpr
41   | ArithmeticExpr "lt" ArithmeticExpr
42   | ArithmeticExpr "gt" ArithmeticExpr
43   | ArithmeticExpr "leq" ArithmeticExpr
44   | ArithmeticExpr "geq" ArithmeticExpr
45   | "(" BooleanExpr ")"
```

Some of the non-terminals of this context-free grammar are reported directly in Haskell. That is, I defined an *abstract syntax tree* that also represents the intermediate representation of a program. This intermediate representation will be the result of the parser. Moreover, the presence of an intermediate representation permits us to apply optimizations *at prior* respect to the interpretation step. The following code snippet contains the definition of the *abstract syntax tree*.

```

1 type Program = (Header, Block)
2 type Block = [Command]
3 type Header = [Variable]
4
5 data ArithmeticExpr
6   = Constant Int
7   | IntegerVar String
8   | ArrayVar String ArithmeticExpr
9   | Add ArithmeticExpr ArithmeticExpr
10  | Sub ArithmeticExpr ArithmeticExpr
11  | Mul ArithmeticExpr ArithmeticExpr
12  | Div ArithmeticExpr ArithmeticExpr
13  | Mod ArithmeticExpr ArithmeticExpr
14  | Neg ArithmeticExpr
15  deriving (Eq, Show)
16
17 data BooleanExpr
18   = Truth Bool
19   | BooleanVar String
20   | Not BooleanExpr
21   | Or BooleanExpr BooleanExpr
22   | And BooleanExpr BooleanExpr
23   | Equal ArithmeticExpr ArithmeticExpr
24   | NotEqual ArithmeticExpr ArithmeticExpr
25   | Less ArithmeticExpr ArithmeticExpr
26   | Greater ArithmeticExpr ArithmeticExpr
27   | LessEqual ArithmeticExpr ArithmeticExpr
28   | GreaterEqual ArithmeticExpr ArithmeticExpr
29   deriving (Eq, Show)
30
31 data Command
32   = Skip
33   | ArithmeticAssignment String ArithmeticExpr
34   | BooleanAssignment String BooleanExpr
35   | ArrayAssignment String ArithmeticExpr ArithmeticExpr
36   | Branch BooleanExpr Block Block
37   | Loop BooleanExpr Block
38   deriving (Eq, Show)
39
40 data Variable
41   = IntegerDecl String
42   | BooleanDecl String
43   | ArrayDecl String Int
44   deriving (Eq, Show)

```

The Parser

The parser can be viewed as a function from a string to a list of pairs of values and strings (Graham Hutton).

```

1 newtype Parser a = Parser {unwrap :: String -> [(a, String)]}

```

Note that the parser have a special function called `unwrap` that takes the function out from the parser. The parser is implemented in a “monadic” way. That is, I implemented the following interfaces: *monad plus* and *monad alternative* as suggested in *Monadic*

Parsing in Haskell (Hutton & Meijer). In order to implement the monad interface I also implemented the *functor*, *applicative* and *monad* interfaces.

```
1 instance Functor Parser where
2   fmap f p = Parser (\cs ->
3     [(f a, cs') | (a, cs') <- unwrap p cs])
```

The functor implementation for the parser implements the `fmap` function, that is the application of a function on a wrapped parser.

```
1 instance Applicative Parser where
2   pure a = Parser (\cs -> [(a, cs)])
3   p <*> q = Parser (\cs -> concat
4     [unwrap (fmap a q) cs' | (a, cs') <- unwrap p cs])
```

Moreover, the applicative implementation for the parser introduces both the function `pure` and the operator `<*>`. The `pure` function takes a simple value and wraps it into a parser. The `<*>` operator takes in input a function wrapped in a parser and another parser. The result is the application of the wrapped function onto the parser.

```
1 instance Monad Parser where
2   return a = pure a
3   p >=> f = Parser (\cs -> concat
4     [unwrap (f a) cs' | (a, cs') <- unwrap p cs])
```

The `bind` operator `>=>` takes as input a parser and a function that returns a parser. The result of this operator is the application of the function to the “unwrapped” parser. The application of multiple `bind` nested operations can be abbreviated using the `do` construct in Haskell.

Moreover, two other *custom* interfaces are implemented, needed for making the parser: `MonadPlus` and `MonadAlternative`. The `plus` operator defined on parsers concatenates the result of each one. Moreover, the `MonadAlternative` give us an operator `<|>` useful for combining parsers in a mutually exclusive way. The classes definition can be found in the following code snippet.

```
1 class (Monad m) => MonadPlus m where
2   zero :: m a
3   plus :: m a -> m a -> m a
4
5 class (MonadPlus m) => MonadAlternative m where
6   (<|>) :: m a -> m a -> m a
7   many :: m a -> m [a]
8   many m = some m <|> return []
9   some :: m a -> m [a]
10  some m = liftA2 (:) m (many m)
11  chain :: m a -> m (a -> a -> a) -> m a
12  chain p o = do a <- p; rest a
13  where
14    rest a = (do f <- o; a' <- p; rest (f a a')) <|> return a
```

Note that the function `liftA2` is the composition of `fmap` and the applicative operator `<*>`. The functions `many` and `some` are called *combinators* and are used to define the concept of repeated parsing. The `many` function refers to zero to any number of applications of a parser while the `some` function refers to at least one to any number of applications of a parser. The implementation for the parser of the `bind` operator and the other functions described before is the following. The `chain` function is another combinator that it’s used for left-associative recursion on parsers, as described by Hutton & Meijer.

```

1 instance MonadPlus Parser where
2   zero = Parser (const [])
3   p `plus` q = Parser (\cs -> unwrap p cs ++ unwrap q cs)
4
5 instance MonadAlternative Parser where
6 (<|>) p q = Parser (\cs ->
7   case unwrap (p `plus` q) cs of
8     [] -> []
9     (x : _) -> [x])

```

The `zero` function defines what is an empty parser. Note that in this context an empty parser is a *failed* parser, that is a parser that results from a syntactical error. The `plus` function concatenates the results of two individual parsers. Using the monadic definition of parser, it permits us to easily build a parser for each structure of the programming language, without having to handle the eventual parsing errors individually. For example, it's possible to define a parser for symbols, identifiers and integers as in the following code snippet. Moreover, since the parser implements the `MonadAlternative` interface described above, the combination of multiple parsers is straightforward and easily *parallelizable* by the compiler.

```

1 item :: Parser Char
2 item = Parser (\case " " -> []; (c : cs) -> [(c, cs)])
3
4 satisfy :: (Char -> Bool) -> Parser Char
5 satisfy p = do c <- item; if p c then return c else zero
6
7 token :: Parser a -> Parser a
8 token p = do space; v <- p; space; return v
9
10 space :: Parser String
11 space = many $ satisfy isSpace
12
13 identifier :: Parser String
14 identifier = token $ some $ satisfy isLetter
15
16 constant :: Parser Int
17 constant = read <$> token (some $ satisfy isDigit)
18
19 char :: Char -> Parser Char
20 char c = satisfy (c ==)

```

That is, I firstly defined a `item` function that reads a character from the string. Then I defined a function called `satisfy` that apply a constraint to the character read by the parser. This function returns an `zero` parser if the constraint is not satisfied. In the end, using both the combinators `many` and `some`, I built parsers for identifiers, constants and also spaces that are the main blocks of the context-free grammar already defined. In the same way, it's possible to define a parser for other constructs, such as keywords and symbols.

```

1 keyword :: String -> Parser String
2 keyword cs = token $ word cs
3
4 word :: String -> Parser String
5 word [c] = do char c; return [c]
6 word (c : cs) = do char c; word cs; return (c : cs)
7

```

```

8 symbol :: Char -> Parser Char
9 symbol c = token $ char c

```

In order to build parsers for both arithmetic and boolean expression the `<|>` operator between parsers is used. So, I defined the parser for arithmetic expressions using other sub-parsers that are used in order to maintain the operators precedence. Moreover, I used the `chain` combinator in order to build the *abstract syntax tree* by the left-associative operators of the grammar.

```

1 arithmeticExpr :: Parser ArithmeticExpr
2 arithmeticExpr = chain arithmeticTerm op
3   where
4     op =
5       do symbol '+'; return Add
6       <|> do symbol '-'; return Sub
7
8 arithmeticTerm :: Parser ArithmeticExpr
9 arithmeticTerm = chain arithmeticFactor op
10  where
11    op =
12      do symbol '*'; return Mul
13      <|> do symbol '/'; return Div
14      <|> do symbol '%'; return Mod
15
16 arithmeticFactor :: Parser ArithmeticExpr
17 arithmeticFactor =
18   do Constant <$> constant
19   <|> do
20     d <- identifier
21     do
22       symbol '['
23       k <- arithmeticExpr
24       symbol ']'
25       return (ArrayVar d k)
26   <|> return (IntegerVar d)
27   <|> do symbol '-'; Neg <$> arithmeticExpr
28   <|> do symbol '('; a <- arithmeticExpr; symbol ')'; return a

```

The `chain` combinator is used in both `arithmeticExpr` and `arithmeticTerm` in order to permits to concatenate multiple addition/subtraction operators and multiple multiplication/division/modulus operators. If we didn't used the `chain` combinator but simple plain tail recursion we would have that the order of operations would be from right to left instead of left to right. The parser for boolean expressions is similar to the parser for arithmetic expressions and so it's omitted in this documentation.

In the end we have a parser for every command described in the grammar. So, I defined a parser for the following commands: `skip`, `assignment`, `if-then-else` and `while-do`. It's important to notice that the grammar also allows for `if-then` statements (i.e. without the `else` command block). A parser that works on both `if-then-else` and `if-then` statements is implemented by combining the two individual parsers using the `<|>` operator. Moreover, using the `many` combinator and the `<|>` operator, the parsers for both multiple commands and a single command are straightforward.

```

1 assignment :: Parser Command
2 assignment = do
3   d <- identifier
4   do
5     symbol '='
6     a <- arithmeticExpr
7     symbol ';'
8     return (ArithmeticAssignment d a)
9   <|> do
10    symbol ':'
11    b <- booleanExpr
12    symbol ';'
13    return (BooleanAssignment d b)
14  <|> do
15    symbol '['
16    k <- arithmeticExpr
17    symbol ']'
18    symbol '='
19    a <- arithmeticExpr
20    symbol ';'
21    return (ArrayAssignment d k a)
22
23 branch :: Parser Command
24 branch = do
25   keyword "if"
26   symbol '('
27   b <- booleanExpr
28   symbol ')'
29   keyword "then"
30   c1 <- block
31   do
32     keyword "else"
33     c2 <- block
34     keyword "end if"
35     symbol ';'
36     return (Branch b c1 c2)
37  <|> do
38     keyword "end if"
39     symbol ';'
40     return (Branch b c1 [Skip])
41
42 loop :: Parser Command
43 loop = do
44   keyword "while"
45   symbol '('
46   b <- booleanExpr
47   symbol ')'
48   keyword "do"
49   c <- block
50   keyword "end while"
51   symbol ';'
52   return (Loop b c)
53
54 skip :: Parser Command
55 skip = do keyword "skip"; symbol ';'; return Skip

```

Another parser that is needed for the grammar, is the parser for the variable declarations. That is, in this programming language we can declare variables only on the top of the program, and before the `shrimp` keyword. The variables can be integers, boolean or array of integers. For the declaration of array of integers, the size of the array must be fixed and a constant integer value.

```

1 variable :: Parser Variable
2 variable = do
3   keyword "let"
4   d <- identifier
5   keyword "as"
6   do
7     keyword "int"
8     symbol ';'
9     return (IntegerDecl d)
10  <|> do
11    keyword "bool"
12    symbol ';'
13    return (BooleanDecl d)
14  <|> do
15    keyword "array"
16    symbol '['
17    n <- constant
18    symbol ']'
19    symbol ';'
20    return (ArrayDecl d n)

```

The parser for the entire program is defined as in the following code snippet. That is, the parser for a block of commands is defined using the `many` combinator on the parser for a `command`. In the same way, the parser for a header is defined using the `many` combinator on the parser for a `variable`. The parser for the entire parser is defined using both the parser for the header and the command block. Note that the keyword `shrimp` separates these parts of the program.

```

1 program :: Parser Program
2 program = do
3   h <- header
4   keyword "shrimp"
5   b <- block
6   return (h, b)
7
8 block :: Parser Block
9 block = many command
10
11 header :: Parser Header
12 header = many variable
13
14 parse :: String -> Result (Program, String)
15 parse cs = case unwrap program cs of
16   [] -> Error EmptyProgram
17   [(p, cs)] -> Ok (p, cs)

```


The Optimizer

The *optimization* process is an intermediate step between the parsing and the interpretation of the program itself. Currently, the main optimization step is related to the execution of constant values. That is, if an expression in a loop is defined only on constant values, it's better to optimize the computation of that expression by replacing it with the result. This procedure is done before the interpretation of the program. For example, consider the following arithmetic expression, expressed in intermediate representation, that we wish to optimize.

```
1 let expr = Div (  
2   (Mul (Identifier "x") (Sub (Constant 5) (Constant 3)))  
3   (Add (Constant 9) (Constant 1)))
```

If we apply the optimization step to this arithmetic expression, we obtain the equivalent but more efficient arithmetic expression.

```
1 let expr' = Div (Mul (Identifier "x") (Constant 2)) (Constant 10)
```

The implementation of the *optimization* process in Haskell is straightforward, due to simple recursion functions. A very similar optimization process is also implemented on boolean expressions. The *optimization* process also includes a basic optimization on commands such as **skip**, **if-then-else** and **while-do**. First of all, all the **skip** commands are removed from the intermediate representation. Moreover, if the condition of a **if-then-else** command is always *true* then the entire statement is replaced with the first block of commands. In a similar way, if the condition is always *false* then the entire statement is replaced with the second block of commands. Furthermore, this approach is also used for **while-do** commands. That is, if the condition of a **while-do** command is always *false* then the entire statement is completely removed. However, if the condition is always *true* then an exception named *Infinite Loop* is raised. In other words, the optimizer is capable of detecting trivial infinite loops and prevents the interpretation of such programs.

The Interpreter

Finally, the obtained intermediate representation is interpreted by the interpreter. The interpreter scan the intermediate representation using *depth-first traversal* on the *n*-ary tree that compose the intermediate representation itself. Before introducing the execution of the interpreter on an intermediate representation, some useful data types are defined in order to handle errors.

```
1 data Exception  
2   = EmptyProgram  
3   | InfiniteLoop  
4   | DivisionByZero  
5   | UndeclaredVariable String  
6   | MultipleVariable String  
7   | TypeMismatch String  
8   | OutOfBound String Int  
9   | InvalidSize String  
10  
11 instance Show Exception where  
12   show EmptyProgram = "Empty Program"
```

```

13 show InfiniteLoop = "Infinite Loop"
14 show DivisionByZero = "Division By Zero"
15 show (UndeclaredVariable d) = "Undeclared Variable" ++ ": " ++ d
16 show (MultipleVariable d) = "Multiple Variable" ++ ": " ++ d
17 show (TypeMismatch d) = "Type Mismatch" ++ ": " ++ d
18 show (OutOfBounds d i) = "Out Of Bound" ++ ": " ++ d ++ " at " ++ show
    i
19 show (InvalidSize d) = "Invalid Size" ++ ": " ++ d
20
21 data Result a = Ok a | Error Exception
22
23 instance Functor Result where
24   fmap f (Ok v) = Ok (f v)
25   fmap _ (Error e) = Error e
26
27 instance Applicative Result where
28   pure v = Ok v
29   (<*>) (Ok f) (Ok v) = Ok (f v)
30   (<*>) (Error e) _ = Error e
31   (<*>) _ (Error e) = Error e
32
33 instance Monad Result where
34   (>=>) (Ok v) f = f v
35   (>=>) (Error e) _ = Error e
36
37 exception :: Exception -> a
38 exception e = errorWithoutStackTrace $ show e

```

As one can see, the `Result` data type is a polymorphic type that can be either a `Ok` or a `Error`. The `Error` type also encapsulate an exception, one of the listed above. Moreover functor, applicative and monad interfaces are implemented in order to apply the needed operators directly on intermediate results. The main function that executes a block (i.e. a list of command) is reported in the following code snippet.

```

1 execute :: State -> Block -> State
2 execute s [] = s
3 execute s (Skip : cs) = execute s cs
4 execute s ((Assignment d a) : cs) =
5   case evalArithmetic s a of
6     Ok v -> execute s' cs
7     where
8       s' = insert d v s
9       Error e -> exception e
10 execute s ((Branch b cs' cs'') : cs) =
11   case evalBoolean s b of
12     Ok True -> execute s (cs' ++ cs)
13     Ok False -> execute s (cs'' ++ cs)
14     Error e -> exception e
15 execute s (c@(Loop b cs') : cs) =
16   execute s (Branch b (cs' ++ [c]) [Skip] : cs)

```

The implementation of the execution of a `while-do` resembles how it's evaluated in operational semantics, i.e. by “wrapping” it in a `if-then-else` command statement.

The implementation of the evaluation of arithmetic expressions make *heavily* use of the methods exposed by the applicative interface. Moreover, the `liftA2` function is used in order to make the evaluation functions implementation more compact.

```

1 evalArithmetic :: State -> ArithmeticExpr -> Result Int
2 evalArithmetic _ (Constant v) = Ok v
3 evalArithmetic s (Identifier d) =
4   case search d s of
5     Just v -> Ok v
6     Nothing -> Error (UndeclaredVariable d)
7 evalArithmetic s (Add a1 a2) = liftA2 (+) v1 v2
8   where
9     v1 = evalArithmetic s a1
10    v2 = evalArithmetic s a2
11 evalArithmetic s (Sub a1 a2) = liftA2 (-) v1 v2
12   where
13     v1 = evalArithmetic s a1
14     v2 = evalArithmetic s a2
15 evalArithmetic s (Mul a1 a2) = liftA2 (*) v1 v2
16   where
17     v1 = evalArithmetic s a1
18     v2 = evalArithmetic s a2
19 evalArithmetic s (Div a1 a2) = seqM2 safeDiv v1 v2
20   where
21     v1 = evalArithmetic s a1
22     v2 = evalArithmetic s a2
23 evalArithmetic s (Mod a1 a2) = seqM2 safeMod v1 v2
24   where
25     v1 = evalArithmetic s a1
26     v2 = evalArithmetic s a2
27 evalArithmetic s (Neg a) = negate <$> v
28   where
29     v = evalArithmetic s a

```

In this implementation, I also used the functions `safeDiv`, `safeMod` and `seqM2` which implementation is showed in the following code snippet. In other words, the `seqM2` is a custom function that implements binary sequencing, similarly to the `bind` operator.

```

1 safeDiv :: Int -> Int -> Result Int
2 safeDiv _ 0 = Error DivisionByZero
3 safeDiv u v = Ok (div u v)
4
5 safeMod :: Int -> Int -> Result Int
6 safeMod _ 0 = Error DivisionByZero
7 safeMod u v = Ok (mod u v)
8
9 seqM2 :: (Monad m) => (a -> b -> m c) -> m a -> m b -> m c
10 seqM2 f x y = join $ liftA2 f x y
11
12 join :: (Monad m) => m (m a) -> m a
13 join m = m >>= id

```

The implementation of the evaluation of boolean expressions is very similar to the one for arithmetic expressions, so it's omitted.

Conclusion

The monadic implementation of the parser give us a very simple way of concatenating and combining multiple parsers of sub-grammars. Also, it's very easy to extend the grammar of the language in order to include other commands or statements. Moreover, the use

of an intermediate representation permits us to apply post-processing and optimizations and make the interpretation itself straightforward and more efficient.

Future works may include the introduction of other types for variables, such as boolean types. Another extension of this work consists of adding useful information about the parsing errors, i.e. missing tokens and relevant row and column locations in the source code where the error occurred. However, other improvements can be done in the optimization step. That is, one can implement more “aggressive” optimizations of the intermediate representation based on the expected state of memory of the program during its execution.

Running an example

First of all, navigate to the project directory. Then, open `ghci` and load the needed modules as following.

```
1 :load app/Main.hs src/Shrimp/Exception.hs src/Shrimp/Grammar.hs
2   src/Shrimp/Optimizer.hs src/Shrimp/Utils.hs src/Shrimp/State.hs
3   src/Shrimp/Interpreter.hs src/Shrimp/Parser.hs
```

Finally, run the main program and insert the path of a source file in the `examples` directory (for example `factorial.shr`).

```
1 *Main> main
2 The Shrimp Interpreter
3 Insert the path of the source file:
4 examples/factorial.shr
5 Memory state:
6   i: 6
7   n: 5
8   x: 120
```