Decaf Language Reference

Mike Lam, James Madison University

Fall 2017

1 Introduction

Decaf is an imperative language similar to Java or C, but is greatly simplified compared to those languages. It will allow us to implement a compiler for the entire language in a single semester while still exploring all of the basic facets of a modern compiler.

This project is originally based on a project from course 6.035 at the Massachusetts Institute of Technology. However, significant changes have been made to the language. This document is the only authoritative reference to the version of Decaf used in CS 432 and CS 630 at James Madison University.

Here is an example program in Decaf:

```
// add.decaf - simple addition example

def int add(int x, int y)
{
    return x + y;           // add the two parameters
}

def int main()
{
    int a;
    a = 3;
    return add(a, 2);
}
```

Here is the output from running the reference compiler on the above program:

```
$ java -jar decaf-1.0.jar -i add.decaf
5
```

Here are some important points of difference between Decaf and more general-purpose procedural languages like C and Java:

- Decaf is not object-oriented.
- Function declarations must begin with the `def` keyword.
- Variable declarations may not include initializations.
- Variable declarations must all occur together at the top of their scope.
- Many syntactic sugar operators (such as `+= or ++`) are not included.
- Boolean expression evaluation is not guaranteed to be short-circuited.

2 Lexical Considerations

There are six token classes in Decaf:

Name	Description
`ID`	identifiers
`KEY`	keywords
`DEC`	decimal literals
`HEX`	hexadecimal literals
`STR`	string literals
`SYM`	symbols

Identifiers and keywords must begin with an alphabetic character and may contain alphanumeric characters and the underscore ('_'). A keyword can be thought of a special identifier that is "reserved" and cannot be used for actual variables or functions. Both keywords and identifiers are case-sensitive, and all keywords are lowercase. For example, `if` is a keyword, but `IF` is a variable name; `foo` and `Foo` are two different names referring to two distinct variables.

The following are all of the supported keywords in Decaf:

```
def if else while return break continue int bool void true false
```

In addition, all of the following words are reserved; they are not currently keywords but may be used in future versions of the language and thus should not be used as identifier names:

```
for callout class interface extends implements new this string float double null
```

Keywords and identifiers must be separated by white space, or a token that is neither a keyword nor an identifier. For example, `iftrue` is a single identifier, not two distinct keywords. If a sequence begins with an alphabetic character, then it and the longest sequence of alphanumeric characters and underscores following it forms a token (either a keyword or an identifier).

Symbols can be split into three types: 1) grouping operators: parentheses, brackets, curly braces, assignment (equals), commas, and semicolons, 2) binary operators (`BINOP`) and 3) unary operators (`UNOP`). There are a variety of binary and unary operators in Decaf, including both arithmetic operators (e.g., plus and minus) and boolean operators (e.g., less-than and boolean OR). The following is a list of all supported symbols in Decaf:

```
( { [ ] } ) , ; = + - * / % < > <= >= == != && || !
```

Integer literals are unsigned and may be written in either decimal or hexadecimal form. The latter will always begin with the sequence `0x`, and may contain either lower- or upper-case letter digits. Neither decimal nor hexadecimal literals may be zero-padded at the beginning.

String literals must be enclosed in quote marks and may contain four kinds of escaped characters: newlines ('`\n`'), tabs ('`\t`'), quotes ('`\"`'), and backslashes ('`\\`'). ASCII is the only supported character set.

Comments are started by "`//`" and are terminated by the end of the line. "Whitespace" may appear between any lexical tokens, and consists of one or more spaces, tabs, newlines, or carriage returns. Comments and whitespace have no effect on a Decaf program's execution, and both should be discarded during the lexical analysis phase of compilation.

Decaf programs should be stored in files with the "`.decaf`" extension, and a single Decaf program may not span multiple files.

3 Syntax

The following is the Decaf grammar (in EBNF form):

$$
\begin{array}{rcl}
\textit{Program} & \rightarrow & \big(\ \textit{Var} \mid \textit{Func}\ \big)^* \\
\\
\textit{Var} & \rightarrow & \textit{Type}\ \texttt{ID}\ \big(\ \text{`}\texttt{[}\text{'}\ \texttt{DEC}\ \text{`}\texttt{]}\text{'}\ \big)?\ \text{`}\texttt{;}\text{'} \\
\\
\textit{Func} & \rightarrow & \texttt{def}\ \textit{Type}\ \texttt{ID}\ \text{`}\texttt{(}\text{'}\ \textit{ParamList}?\ \text{`}\texttt{)}\text{'}\ \textit{Block} \\
\textit{ParamList} & \rightarrow & \textit{Type}\ \texttt{ID}\ \big(\ \text{`}\texttt{,}\text{'}\ \textit{Type}\ \texttt{ID}\ \big)^* \\
\\
\textit{Block} & \rightarrow & \text{`}\texttt{\{}\text{'}\ \textit{Var}^*\ \textit{Stmt}^*\ \text{`}\texttt{\}}\text{'} \\
\\
\textit{Stmt} & \rightarrow & \textit{Loc}\ \text{`}\texttt{=}\text{'}\ \textit{Expr}\ \text{`}\texttt{;}\text{'} \\
\textit{Stmt} & \rightarrow & \textit{FuncCall}\ \text{`}\texttt{;}\text{'} \\
\textit{Stmt} & \rightarrow & \texttt{if}\ \text{`}\texttt{(}\text{'}\ \textit{Expr}\ \text{`}\texttt{)}\text{'}\ \textit{Block}\ \big(\ \texttt{else}\ \textit{Block}\ \big)? \\
\textit{Stmt} & \rightarrow & \texttt{while}\ \text{`}\texttt{(}\text{'}\ \textit{Expr}\ \text{`}\texttt{)}\text{'}\ \textit{Block} \\
\textit{Stmt} & \rightarrow & \texttt{return}\ \textit{Expr}?\ \text{`}\texttt{;}\text{'} \\
\textit{Stmt} & \rightarrow & \texttt{break}\ \text{`}\texttt{;}\text{'} \\
\textit{Stmt} & \rightarrow & \texttt{continue}\ \text{`}\texttt{;}\text{'} \\
\\
\textit{Expr} & \rightarrow & \textit{Expr}\ \texttt{BINOP}\ \textit{Expr} \\
\textit{Expr} & \rightarrow & \texttt{UNOP}\ \textit{Expr} \\
\textit{Expr} & \rightarrow & \text{`}\texttt{(}\text{'}\ \textit{Expr}\ \text{`}\texttt{)}\text{'} \\
\textit{Expr} & \rightarrow & \textit{Loc} \\
\textit{Expr} & \rightarrow & \textit{FuncCall} \\
\textit{Expr} & \rightarrow & \textit{Lit} \\
\\
\textit{Type} & \rightarrow & \texttt{int} \mid \texttt{bool} \mid \texttt{void} \\
\textit{Loc} & \rightarrow & \texttt{ID}\ \big(\ \text{`}\texttt{[}\text{'}\ \textit{Expr}\ \text{`}\texttt{]}\text{'}\ \big)? \\
\\
\textit{FuncCall} & \rightarrow & \texttt{ID}\ \text{`}\texttt{(}\text{'}\ \textit{ArgList}?\ \text{`}\texttt{)}\text{'} \\
\textit{ArgList} & \rightarrow & \textit{Expr}\ \big(\ \text{`}\texttt{,}\text{'}\ \textit{Expr}\ \big)^* \\
\\
\textit{Lit} & \rightarrow & \texttt{DEC} \mid \texttt{HEX} \mid \texttt{STR} \mid \texttt{true} \mid \texttt{false}
\end{array}
$$

The following is a key to some of the meta-notation used above:

Example	Description
Foo	non-terminal
`foo`	keyword
`FOO`	token class (see Section 2)
‘`a`’	symbol consisting of character ‘`a`’
x?	zero or one occurences of x (i.e., optional)
x*	zero or more occurences of x
$\big(\ \big)$	used for grouping
$\mid$	designates alternatives

Although this is not reflected in the reference grammar on the previous page, all binary operations in Decaf are left-associative. In addition, there are seven levels of operator precedence, shown in the table below from highest to lowest:

Operators	*Comments*
`-`	unary negation
`!`	unary boolean negation (NOT)
`* / %`	integer multiplication, division, and remainder
`+ -`	integer addition and subtraction
`< <= >= >`	boolean ordinal relation
`== !=`	boolean equality
`&&`	boolean conjunction (AND)
`||`	boolean disjunction (OR)

As written above, this grammar is neither LL(1) nor LR(1); however, it can be transformed (using standard CFG transformations) into a grammar that is both. The process of performing these transformations is a useful exercise for the reader and a vital component of building a parser for the language.

Note that variables may be declared `void`; such declarations are allowed by the grammar but will be flagged during the static analysis phase.

Because the provided grammar is not yet suitable for directly building a parser, there are many possible correct LL(1) and LR(1) grammars, and thus many correct parse trees for any given Decaf program. Thus, it is important to have a clearly-defined *abstract syntax tree* format that is independent of any specific parsing implementation. Here is a list of standardized Decaf AST objects (arranged in a tree format to indicate inheritance relationships):

```
Node
    Program                     (contains Variables and Functions)
    Variable
    Function                    (contains a Block)
    Block                       (contains Variables and Statements)
    Statement
        Assignment              (contains a Location and an Expression)
        VoidFunctionCall        (contains Expressions)
        Conditional             (contains an Expression and either one or two Blocks)
        WhileLoop               (contains an Expression and a Block)
        Return                  (contains an optional Expression)
        Break
        Continue
    Expression
        Location
        FunctionCall            (contains Expressions)
        Literal
        BinaryExpr              (contains two Expressions)
        UnaryExpr               (contains one Expression)
```

4 Semantics

A Decaf program consists of a series of variable and function declarations.

4.1 Variables

All Decaf variables are statically typed, and there are two basic types: 32-bit signed integers (`int`) and booleans (`bool`). Variables may be declared outside a function; such variables are considered "global" variables and are accessible from all functions. Variables declared inside a function or block are only visible inside the corresponding lexical scope (see Section 4.3 for details).

Decaf supports simple, fixed-length, one-dimensional, static arrays. Arrays must be declared global and their size is specified at their declaration time. Arrays are indexed from 0 to $N-1$, where N is the size of the array (which must be greater than zero). The usual bracket notation is used to index arrays. Because arrays have a compile-time fixed size and cannot be declared as parameters (or local variables), there is no facility for querying the length of an array variable in Decaf. Arrays may be of either basic type (integers or booleans).

Assignment is only permitted for scalar values. Decaf uses value-copy semantics, and the assignment "`<loc> = <expr>`" copies the value resulting from the evaluation of `<expr>` into `<loc>`. If a variable is referenced before it is assigned, the result is undefined.

4.2 Functions

In Decaf, functions must be declared using the "`def`" keyword. Functions must either be "`void`" or have a single return value; they may take an unlimited number of parameters. Each parameter must have a formally-declared type. Arrays may NOT be passed as function parameters.

Every Decaf program must contain a function called `main` that takes no parameters and returns an `int`. Execution of the program begins at the `main` function. Functions may be called before their definition in the source code.

Decaf does not provide support for variadic parameters (e.g., "printf" in C).

The Decaf language does allow recursive functions, but the built-in interpreter currently does not support them.

4.3 Scope

Decaf uses very simple static scoping rules. There are at least two valid scopes at any point in a Decaf program: the global scope and the function scope. The global scope consists of names of variables and functions introduced at the top level of the source code. The function scope consists of names of variables and formal parameters introduced in a function declaration. Additional local scopes exist within each block in the code; these can come after `if` or `while` statements or anywhere there is a new block.

An identifier introduced in a function scope can *shadow* an identifier from the global scope. In this case, the identifier may only be used as a variable until the variable leaves scope. Similarly, identifiers introduced in local scopes shadow identifiers in less deeply nested scopes, the function scope, and the global scope.

No identifier may be defined more than once in the same scope. Thus, variable and function names must all be distinct in the global scope, and local variable names and formal parameters names must be distinct in each local scope.

4.4 Function Calls

Function invocation involves (1) passing argument values from the caller to the callee, (2) executing the body of the callee, and (3) returning to the caller, possibly with a result.

Argument passing is defined in terms of assignment: the formal arguments of a function are considered to be like local variables of the function and are initialized, by assignment, to the values resulting from the evaluation of the argument expressions. The arguments are evaluated from left to right.

The body of the callee is then executed by executing the statements of its function body in sequence.

A function that is declared with a `void` return type can only be called as a statement, i.e., it cannot be used in an expression. Such a function returns control to the caller when return is called (no result expression is allowed) or when the textual end of the callee is reached.

A function that returns a result may be called as part of an expression, in which case the result of the call is the result of evaluating the expression in the return statement when this statement is reached. It is illegal for control to reach the textual end of a function that returns a result.

A function that returns a result may also be called as a statement. In this case, the result is ignored.

4.5 Control Structures

The `if` statement has the same semantics as in standard procedural programming languages. First, the conditional expression is evaluated. If the result is true, the "true" block is executed. Otherwise, the "else" block is executed, if it exists. Since Decaf requires that the "true" and "else" blocks be enclosed in braces, there is no ambiguity in matching an "else" block with its corresponding `if` statement.

The `while` statement also has the same semantics as in standard procedural languages. First, the guard expression is evaluated. If the result is true, the loop body is executed. After the loop body executes one iteration, the guard expression is re-evaluated. If the guard expression evaluates to false at any point at which it is evaluated, the loop terminates and execution resumes at the statement following the loop.

Decaf provides support for the standard `continue` and `break` statements, which immediately transfer control to the beginning or end of the loop (respectively). In the case of the former, the guard expression should be re-checked before the loop body is executed again.

4.6 Expressions

Expressions follow the normal rules for evaluation. In the absence of other constraints, operators with the same precedence are evaluated from left to right. Parentheses may be used to override normal precedence.

A location expression evaluates to the value contained by the location. The semantics of this differ depending on whether the location is an l-value or an r-value. Array operations are discussed in Section 4.1.

Method invocation expressions are discussed in Section 4.4.

Integer literals evaluate to their integer value. Boolean literals are evaluated to the integer values 1 (true) and 0 (false). String literals do not evaluate to anything; since variables cannot store strings, their only valid use is as an argument to predefined functions (e.g., the `print_str` function; see Section 4.7).

The arithmetic operators (arith op and unary minus) have their usual precedence and meaning, as do the relational operators (rel op). % computes the remainder of dividing its operands.

Relational operators are used to compare integer expressions and may not be used on boolean variables. The equality operators (`==` and `!=` are valid for both `int` and `boolean` types, and can be used to compare any two expressions having the same type. The result of a relational operator or equality operator has type boolean. Boolean expression evaluation is not guaranteed to be short-circuited.

4.7 Input and Output

Decaf supports only very primitive I/O. There is currently no support for input; all input values must be hard-coded in the source. Output is supported using the following predefined library functions:

```
void print_str(string value)
void print_int(int value)
void print_bool(bool value)
```

The first two operate as expected. The last one (`print_bool`) prints the internal representation of the boolean (i.e., 1 for true and 0 for false). None of these functions add a newline at the end of the input; if newlines are desired they must be printed explicitly using the `print_str` function.

These functions are not actually implemented in Decaf (they are provided by the Decaf standard library) and thus must be manually added to the global symbol table during static analysis to ensure that calls to them are properly type-checked. Here are some examples of their use:

```
def int main()
{
    print_str("Hello!");            // result: "Hello!"
    print_int(2+3);                 // result: "5"
    print_bool(5 < 9);              // result: "1"
    print_bool(5 < 2);              // result: "0"
    return 0;
}
```

5 Type Checking

Here is an incomplete list of basic type inference and type checking rules for Decaf. Each rule consists of a series of antecedents or premises (above the line) and a conclusion (below the line). In these rules, $\Gamma \vdash e : \tau$ means that "e has type τ in environment Γ". If τ is omitted, the statement simply means that "e is well-typed" (i.e., it has no type errors). Here, Γ (called the *typing context* or *type environment*) refers to the local symbol table (with the ability to look up symbols recursively through its parent table). The absence of Γ indicates that no symbol table is necessary to type-check e.

Expressions:

$$\text{TDec}\frac{}{\vdash \texttt{DEC} : \textbf{int}} \qquad \text{THex}\frac{}{\vdash \texttt{HEX} : \textbf{int}} \qquad \text{TStr}\frac{}{\vdash \texttt{STR} : \textbf{str}}$$

$$\text{TTrue}\frac{}{\vdash \texttt{true} : \textbf{bool}} \qquad \text{TFalse}\frac{}{\vdash \texttt{false} : \textbf{bool}} \qquad \text{TSubExpr}\frac{\Gamma \vdash e : \tau}{\Gamma \vdash \text{`(`}\ e\ \text{`)`} : \tau}$$

$$\text{TLoc}\frac{\texttt{ID} : \tau \in \Gamma}{\Gamma \vdash \texttt{ID} : \tau} \qquad \text{TArrLoc}\frac{\texttt{ID} : \tau[] \in \Gamma \qquad \Gamma \vdash e : \textbf{int}}{\Gamma \vdash \texttt{ID}\ \text{`[`}\ e\ \text{`]`} : \tau}$$

$$\text{TFuncCall}\frac{\texttt{ID} : (\tau_1, \tau_2, ..., \tau_n) \to \tau_r \in \Gamma \qquad \Gamma \vdash e_1 : \tau_1 \qquad \Gamma \vdash e_2 : \tau_2 \qquad ... \qquad \Gamma \vdash e_n : \tau_n}{\Gamma \vdash \texttt{ID}\ \text{`(`}\ e_1, e_2, ..., e_n\ \text{`)`} : \tau_r}$$

$$\text{TNot}\frac{\Gamma \vdash e : \textbf{bool}}{\Gamma \vdash\ \text{`!`}\ e : \textbf{bool}} \qquad \text{TNeg}\frac{\Gamma \vdash e : \textbf{int}}{\Gamma \vdash\ \text{`-`}\ e : \textbf{int}}$$

$$\text{TAdd}\frac{\Gamma \vdash e_1 : \textbf{int} \qquad \Gamma \vdash e_2 : \textbf{int}}{\Gamma \vdash e_1\ \text{`+`}\ e_2 : \textbf{int}} \quad \text{(similar for TSub (-), TMul (*), TDiv (/) and TMod (\%))}$$

$$\text{TLe}\frac{\Gamma \vdash e_1 : \textbf{int} \qquad \Gamma \vdash e_2 : \textbf{int}}{\Gamma \vdash e_1\ \text{`<`}\ e_2 : \textbf{bool}} \quad \text{(similar for TLeq (<=), TGe (>), and TGeq (>=))}$$

$$\text{TEq}\frac{\Gamma \vdash e_1 : \tau \qquad \Gamma \vdash e_2 : \tau}{\Gamma \vdash e_1\ \text{`==`}\ e_2 : \textbf{bool}} \quad \text{(similar for TNeq (!=))}$$

$$\text{TAnd}\frac{\Gamma \vdash e_1 : \textbf{bool} \qquad \Gamma \vdash e_2 : \textbf{bool}}{\Gamma \vdash e_1\ \text{`\&\&`}\ e_2 : \textbf{bool}} \quad \text{(similar for TOr (||))}$$

Statements:

$$\text{TIf}\frac{\Gamma \vdash e : \textbf{bool} \qquad \Gamma \vdash b}{\Gamma \vdash \texttt{if}\ \text{`(`}\ e\ \text{`)`}\ b} \qquad \text{TIfElse}\frac{\Gamma \vdash e : \textbf{bool} \qquad \Gamma \vdash b_1 \qquad \Gamma \vdash b_2}{\Gamma \vdash \texttt{if}\ \text{`(`}\ e\ \text{`)`}\ b_1\ \texttt{else}\ b_2} \qquad \text{TWhile}\frac{\Gamma \vdash e : \textbf{bool} \qquad \Gamma \vdash b}{\Gamma \vdash \texttt{while}\ \text{`(`}\ e\ \text{`)`}\ b}$$

$$\text{TAssign}\frac{\texttt{ID} : \tau \in \Gamma \qquad \Gamma \vdash e : \tau}{\Gamma \vdash \texttt{ID}\ \text{`=`}\ e\ \text{`;`}} \qquad \text{TArrAssign}\frac{\texttt{ID} : \tau[] \in \Gamma \qquad \Gamma \vdash e_i : \textbf{int} \qquad \Gamma \vdash e_v : \tau}{\Gamma \vdash \texttt{ID}\ \text{`[`}\ e_i\ \text{`]`}\ \text{`=`}\ e_v\ \text{`;`}}$$

$$\text{TVoidFuncCall}\frac{\texttt{ID} : (\tau_1, \tau_2, ..., \tau_n) \to \tau_r \in \Gamma \qquad \Gamma \vdash e_1 : \tau_1 \qquad \Gamma \vdash e_2 : \tau_2 \qquad ... \qquad \Gamma \vdash e_n : \tau_n}{\Gamma \vdash \texttt{ID}\ \text{`(`}\ e_1, e_2, ..., e_n\ \text{`)`}\ \text{`;`}}$$

$$\text{TBlock}\frac{\Gamma' \vdash s_1 : \tau_1 \qquad \Gamma' \vdash s_2 : \tau_2 \qquad ... \qquad \Gamma' \vdash s_n : \tau_n}{\Gamma \vdash \text{`\{`}\ s_1, s_2, ..., s_n\ \text{`\}`}} \qquad \text{where } \Gamma' = \Gamma \cup \{v_i : \tau_i \in Vars\}$$