

HamSkill: Run Haskell Anywhere
with ANTLR and Scala

CS252 Project Final Report

Zayd Hammoudeh
(zayd.hammoudeh@sjsu.edu)

March 30, 2016

Contents

1	Running in the Java Virtual Machine	3
2	Key Project Requirements	3
3	<i>HamSkill</i> 's Software Architecture	4
3.1	ANTLR	4
3.1.1	ANTLR Version 4 Grammar	4
3.1.1.1	Haskell Grammar	6
3.1.1.2	ScalaOutput Grammar	6
3.1.2	ANTLR Classes	7
3.1.2.1	HaskellTokensToScala	7
3.1.2.2	ScalaOutputTokensToHaskellFormat	8
3.2	Transpiler Output Language: Scala	8
3.2.0.1	<i>HamSkill</i> Standard	9
3.2.0.2	<i>HamSkill</i> +	10
4	Haskell Features Supported by <i>HamSkill</i>	10
4.1	Immutability in Scala	10
4.2	Lazy Evaluation	10
4.3	Supported Types	11

4.4	<code>main</code> Function	11
4.5	Conversion from Functions to Methods	11
4.6	Nested Function Calls	12
4.7	Defining Scope and Scala Object Name via <code>module</code>	12
4.8	Partially Applied Functions	13
4.9	Higher Order Function Support	13
4.10	Haskell Lambda Function to Scala Anonymous Functions	14
4.11	Maybe Monad Support	14
5	<i>HamSkill</i> Testing Overview	14
5.1	System Level and Regression Testing	14
5.2	Test Bench Implementation Overview	15
6	Conclusions	15

1 Running in the Java Virtual Machine

C is one of the most commonly used languages when the goal is maximum performance. However, C/C++’s “write once, compile anywhere” paradigm limits its portability. In contrast, the near ubiquity of the Java Virtual Machine (JVM) allows Java to be “write once, run anywhere.”

On many occasions, developers have leveraged the JVM’s “run anywhere” capability to run allow other languages. Examples include: JRuby for the Ruby programming language [2], Jython for the Python programming language [3], Renjin for the R programming language [4], and Scala [5].

Currently, there is no full implementation of Haskell in the JVM. One Haskell dialect that is runnable in Java is Frege [9].

In this project, I will implement, *HamSkill*, which is a transpiler from Haskell to Scala; *HamSkill* enables a dialect of Haskell to be runnable in the JVM.

2 Key Project Requirements

When designing and implementing this project, there were four primary goals:

1. **Runnable in the Java Virtual Machine** - As explained in section 1, Java’s Virtual Machine enables significant machine independence, which Haskell does not currently enjoy.
2. **Minimal JVM Requirements** - In addition to just running in the JVM, HamSkill was intended to be as standalone as possible. For example, for most applications, it was not expected that the user would have Scala installed on their machine. To achieve this, some advanced features may not be supported.
3. **Identical Input and Output Between Haskell and *HamSkill*** - In many scenarios, it may not be sufficient for a program to run in Haskell and in JVM. Rather, it may often be required that the generated output for the two programs be identical as well. As such, *HamSkill* will utilize an additional post processing step to ensure its output is identical to that generated by Haskell.
4. **Human Readable Output Code** - A transpiler is any program that takes source code from one programming languages and outputs as code in another programming language with a similar level of abstraction [7].

To enable increased reuse of the outputted code, HamSkill will use proper indenting, newlines, etc. to maximize readability the generated output. While this is not a requirement for the complete system to work properly, it can enhance the potential of the tool.

3 *HamSkill*'s Software Architecture

HamSkill is a transpiler that takes input Haskell code, converts it to Scala, and then runs it in the JVM. The *HamSkill* implementation consists of four major components. They are:

- ANTLR Lexer and Parser
- Haskell Antlr Grammar
- HaskellMain Java Class
- Scala Runtime Environment
- ScalaOutput Antlr Grammar
- ScalaOutput Java Class

The relationships between these four components are shown in figure 3.

The following subsections describe each of the components of this architecture.

3.1 ANTLR

ANTLR (ANother Tool for Language Recognition) is an adaptive left-to-right, left-most derivation (LL(*)) lexer and parser written in the Java programming language. ANTLR's primary function is to read, parse, and process structured text (e.g. Haskell code) [8].

3.1.1 ANTLR Version 4 Grammar

A grammar is a formal description of language; it is based off the concept of a context-free grammar in formal automata theory. ANTLR version 4 (v4) grammar files (denoted by the file extension *.g4*) explicitly define the how ANTLR

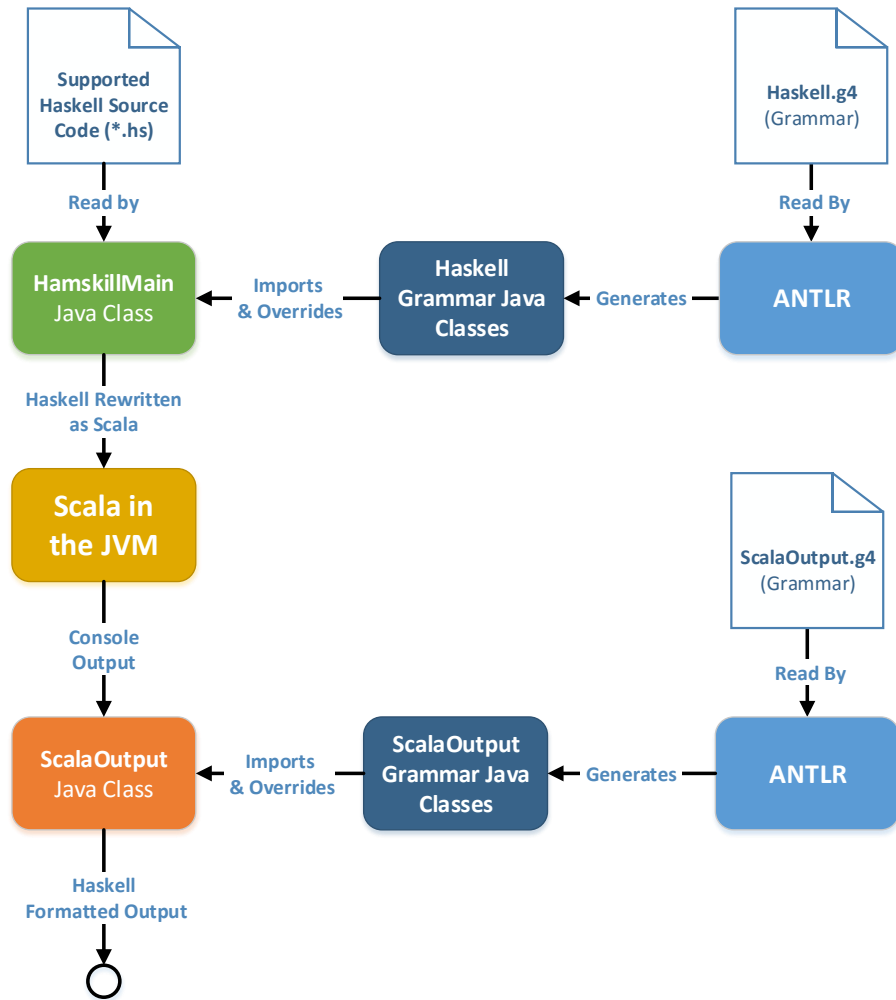


Figure 1: HamSkill Project Architecture

will parse the structured text. The grammar file may contain token definitions (which always start with a capital letter) and/or parser rules (which always start with a lowercase letter). The tokens are groups of characters that form a single object; the parser then uses these tokens to recognize sentence structure within the document.

This project utilizes two separate grammars namely: `Haskell` and `ScalaOutput`. They are described in the following two subsections.

3.1.1.1 Haskell Grammar

The `Haskell` parsing grammar is stored in a file named "`Haskell.g4`." This grammar defines both the tokens and abstract syntax tree for supported Haskell code. Some of Haskell language features that are supported by this grammar include:

- `case` Statement
- `if`, `then`, `else` Conditional
- `Maybe` Monad
- Currying
- Partially Applied Functions
- Higher Order Functions
- Lazy Evaluation

For the complete feature set as well as specific Haskell syntax requirements, see section 4.

3.1.1.2 ScalaOutput Grammar

As mentioned in section 2, one of the key requirements of this project was that the outputs from Haskell and HamSkill should be identical. Figure 2 and 3 show the code to print a list containing "2", "3", and "4" in Haskell and Scala respectively. Note that the subsequent lines which show the very different console outputs for lists that are identical.

```
Prelude> putStrLn $ show [2, 3, 4]
[2,3,4]
```

Figure 2: Printing a Three Element List in Haskell

```
scala> println(List(2,3,4))  
List(2, 3, 4)
```

Figure 3: Printing a Three Element List in Scala

The `ScalaOutput` grammar (contained in the file "`ScalaOutput.g4`") is used to correct for these types of differences by parsing the console output of the Scala program and converting it to a syntax that is more similar to that of Haskell.

3.1.2 ANTLR Classes

Programs that use ANTLR do not generally operate directly on the grammar. Rather, the grammar is converted by ANTLR into a set of Java classes; given an example grammar, *GrammarName.g4*, ANTLR would create the following files:

- *GrammarNameLexer.java* - This class is the the lexer definition for the input grammar file. It also extends ANTLR's base Lexer class.
- *GrammarNameParser.java* - Each rule in the original grammar constitutes a method in this class; it forms the parser class definition for the input grammar file.
- *GrammarName.tokens* - Assigns a token type (e.g. integer, identifier, floating point number etc.) to each token in the input grammar file.
- *GrammarNameListener.java* & *GrammarNameBaseListener.java* - ANTLR builds applies the grammar to a text input to build an Abstract Syntax Tree (AST). While walking the tree, ANTLR fires events that can be captured by a listener[8].

Most programs that want to use ANTLR override the functions in the file *GrammarNameBaseListener.java*.

3.1.2.1 HaskellTokensToScala

The class "`HaskellTokensToScala`" overrides some of the methods ANTLR auto-generated in the file "`HaskellBaseListener.java`"; note that all of these methods are derived from the grammar "`Haskell.g4`" described in section 3.1.1.1.

3.1.2.2 `ScalaOutputTokensToHaskellFormat`

Similar to the `HaskellTokensToScala` class, the class, “`ScalaOutput`” extends the ANTLR auto-generated class, “`ScalaOutputBaseListener`”.

The listeners in this class are primarily transforming for two different categories of Scala outputs. They are:

- **Lists** - As shown in figures 2 and 3, the console output of lists in Scala and Haskell are significantly different. The methods in this class coupled with the `ScalaOutput` grammar convert printed lists from Scala to Haskell format.
- **Maybe Monad** - The Scala equivalent of Haskell’s `Maybe` monad is named `Option`. While the syntax for the two are similar, they are not identical. For example, `Just` and `Nothing` in Haskell are `Some` and `None` in Scala. The conversion from Scala format to Haskell is handled by this class.

Any other Scala console outputs that do not fall into these two categories are passed through unchanged.

3.2 Transpiler Output Language: Scala

As mentioned previously, a transpiler takes source code from one programming language and converts it to code in another language. It was also previously explained that the primary criteria when deciding on the output language was that it needed to be runnable in the Java Virtual Machine. Other important criteria were:

- **Higher Order Function Support** - Any language that is devoid of higher order support would not be a good match with a purely functional language like Haskell.
- **Similar Syntactic Structure** - Closer alignment between the input and output languages will simplify the trans-compilation. Inevitably though, some amount of restructuring and reformatting will be required.
- **Personal Interest** - Increased personal investment generally leads to a more fulfilling outcome for the student and a better project overall.

The language that best fit all three of these criteria is Scala, which is a functional language which can be fully run inside the JVM [6]. What is more, it

natively supports many of Haskell’s core features including: functions written in a pattern matching style, immutability of objects, currying, partially applied functions, lazy evaluation, static typing, etc. One of the disadvantages of Scala is its weaker type inference in comparison to Haskell; due to this specific requirements were placed on the requirements of the Haskell dialect supported by HamSkill as described in section 4. It is also important to note that the author has had a personal predisposed interest in learning Scala given its extensive support by Apache Spark. As such, given the ”personal interest” requirement mentioned previously, the selection of Scala for this project became an even more obvious choice.

HamSkill implements two different schemes for running the transpiled Scala. These two approaches, *HamSkill* Standard and *HamSkill*+ are described in the follow subsections.

3.2.0.1 *HamSkill* Standard

Since Scala is a compiled language, it does not lend itself tremendously well to Java runtime compilation and execution. To achieve this functionality, I leveraged Twitter’s `util-eval` library, which allows runtime compilation and execution of Scala code entirely within a Java program [?]. The advantages and disadvantages of using this library are:

- **Advantages:**

1. **Reduced JVM Requirements** - Twitter’s `util-eval` library is a JAR file that only requires the presence of Scala compile and main library to run. Hence, this eliminates the need for the user to have Scala installed in their environment. This addresses one of the key requirements enumerated in section 2.
2. **Simplified Usage** - When HamSkill Standard is used, only the `main` method in the `HaskellMain` Java class needs to be called to import the Haskell code, transpile it to Scala, compile and run the Scala code, and then post-process the output. *HamSkill*+ requires three calls to perform the same action.

- **Disadvantages:**

1. **Reduced Feature Set** - The `util-eval` library does not support all Scala features. An example of a missing feature is `readLine` which takes console inputs.
2. **Reliant on Third Party Developers** - As mentioned previously, `util-eval` was developed by Twitter. Hence, it can be deprecated

at any time as was previously planned [?]. While the tool is open-source, picking up a dropped project introduces an additional set of complications.

3.2.0.2 *HamSkill+*

4 Haskell Features Supported by *HamSkill*

This section enumerates the Haskell features that are supported by HamSkill. It also describes any formatting requirements that may apply to these features.

4.1 Immutability in Scala

One of the key features of Haskell that allows it to achieve referential transparency is the immutability of data. While it would be possible to develop an infrastructure in a language like Python to assure immutability, it is an encumbrance. In cases such as this, it is almost always better to take advantage of a language's native features when possible. In Scala, the `val` construct ensure the immutability of an object without any user intervention. For example, the code in figure 4 would raise a runtime error since it is trying to change the value of immutable data.

```
val x = 5  
x = 3
```

Figure 4: Declaring Immutable Data in Scala

4.2 Lazy Evaluation

Another important aspect of Haskell is that it supports lazy evaluation. This entails that data's value is not calculated until it is not needed. Figure 5 is an example of lazy evaluation with Scala as when this code is run, it will print a negative elapsed time (which is clearly not correct).

```

def lazyTime(){
  lazy val t1 = System.nanoTime()
  val t2 = System.nanoTime()
  Thread.sleep(1000)
  println("Elapsed time is " + (t2-t1)/1000000 + "ms")
}

```

Figure 5: Lazy, Immutable Code in Scala

Scala also supports “call-by-name” to achieve laziness of function parameters. However, this feature is not truly lazy as it will recalculate the value each time the parameter is used in the function. This limitation often degrades the overall the performance; for this reason, I do not plan to implement laziness in *HamSkill* across functions.

4.3 Supported Types

HamSkill will only support a select subset of Haskell’s available types. The list of planned types are: **Bool**, **Integer** (i.e. bounded), and **List** (currently only finite lists are planned, but that may change depending on the complexity of the implementation).

While implementing floating point numbers would not add substantial complexity at a basic level, ensuring that the floating point behavior of *HamSkill* (i.e. Scala) and Haskell are identical is beyond the scope of this project.

4.4 main Function

4.5 Conversion from Functions to Methods

The function in Haskell to convert data to string is “**show**”. In contrast, the syntax in Scala to convert an object (e.g. “x”) to a string is “**x.toString()**”. Due to this, the ANTLR parser will need to be able to convert a prefix function to an object method call.

4.6 Nested Function Calls

Imperative languages (e.g. Java) are generally more verbose than functional languages; Haskell is no exception to this. Conciseness introduces significant challenges when writing a parser as the contextual information is reduced. For example, figure 6 is a simple line of Haskell code that prints to the screen the result of a function “`addTwoNumbs`” that takes two integers (e.g. “`x`” and “`y`”) and sums them.

```
putStrLn $ show $ addTwoNumbs x y
```

Figure 6: Simple Function Call in **Haskell**

Similar code in Java is shown in figure 7

```
System.out.println( addTwoNumbs(x, y) );
```

Figure 7: Simple Function Call in **Java**

The Java syntax explicitly shows that `addTwoNumbs` is a function since the parameters are inside parentheses and are comma separated. To simplify the parsing for this in Haskell, the *HamSkill* requires function arguments to be preceded and succeeded by triple parentheses “`((`” and “`))`”. Figure 8 shows the *HamSkill* version of the Haskell code in figure 8.

```
putStrLn $ show $ addTwoNumbs ((x y))
```

Figure 8: Simple Function Call in *HamSkill*

4.7 Defining Scope and Scala Object Name via module

A program in Haskell is composed of a set of “module” files. The “module” keyword is used to scope of functions (e.g. `public` or `private`) as well as for defining an abstract data type [1]. In *HamSkill*, I will use the Haskell module

to define whether the Scala methods are private (since by default functions are `public`) as well as the name of the Scala object.

4.8 Partially Applied Functions

Haskell supports partially applied functions. Figure 9 shows the `addTwoInts` function with a single argument (i.e. “5”) being stored in a variable “`addFive`”.

```
let addFive = addTwoInts 5
```

Figure 9: Partially Applied `addTwoInts` Function in **Haskell**

Partially applied functions in Scala have advantages and disadvantages in comparison to Haskell. One of these disadvantages is evident when figures 9 and 10 are compared. Note that the Scala function requires an underscore (“`_`”) for each missing argument as well as the type for that argument. This makes converting Haskell code to Scala problematic as the function prototype must be fixed and known at conversion time.

```
val addFive = addTwoInts(5, _ : Int)
```

Figure 10: Partially Applied `addTwoInts` Function in **Scala**

To simplify this, *HamSkill* will require that any partially defined functions are declared in the same file/module. I will investigate using a predefined list of functions, but this may not be feasible or support will be very limited due to the requirement to define the parameter type. What is more, partially applied functions will use need to use the “double parentheses style” described in section 4.6.

4.9 Higher Order Function Support

Scala and Haskell are both functional programming languages; one important consequence of this is that both support higher order functions. *HamSkill* will support functions as input parameters to functions. If time allows, I will also

investigate the ability to return functions from functions. The extent to which this is supported will be dependent on the extent to which partially applied functions are supported as defined in section ??.

4.10 Haskell Lambda Function to Scala Anonymous Functions

There is significant similarity between a Lambda function in Haskell and an anonymous function in Scala. One primary difference is that Scala requires the developer to specify the types of the parameters in the anonymous function while Scala does not. For this project, I will implement support for anonymous functions either as parameters to other functions or for support for an operation such as folding or filtering a list.

4.11 Maybe Monad Support

5 *HamSkill* Testing Overview

Software testing is important for detecting defects in software programs. What is more, there are different types of software testing including unit testing, module level testing, and system tests. Given that ANTLR relies heavily on the Listener pattern, it reduces the ease at which unit testing can be performed. What is more, given that *HamSkill* relies on the end to end operation of programs written in Haskell, ANTLR, Java, and Scala, it really emphasizes the importance of system level test.

5.1 System Level and Regression Testing

HamSkill's uses black-box system level testing. A set of use cases (i.e. Haskell files) were generated by me; these files are then run through both Haskell (to verify they are valid Haskell code) and through *HamSkill*; the resulting outputs are then compared. If the outputs are the same, the test is considered successful; in the case of any difference, the test is classified as a failure. The full set of *HamSkill* test cases are included with this submission in the folder `Test_Bench/test_cases`.

Regression testing verifies that new features do not affect/compromise existing ones. When developing *HamSkill*, I iteratively added new features. After each feature was added, I reran the entire test bench to ensure that no new

(detectable) issues were introduced. If an issue did arise, I would immediately address before moving on to new features to ensure that fundamental problems did not compound upon themselves. This formed a type of regression testing throughout my development.

5.2 Test Bench Implementation Overview

The *HamSkill* test architecture is written as a **bash** script (see the file in the submission `Test_Bench/test_bench.sh`. The function “`perform_hamskillStd_and_hamskillPlus_test`” is the most important in the script as it is used to perform most of the testing; the basic operational flow of this function is:

- **Step #1:** The function is provided the name of a Haskell program written in the dialect supported by *HamSkill*, but without the file extension.
- **Step #2:** The Haskell program is run in Haskell, and the program’s output is stored to a specific file on disk using the **bash** command “`>`”.
- **Step #3:** Run HamSkill in standard mode and store the results to a file using the **bash** command “`>`”.
- **Step #4:** Perform a file difference operation (using the **bash** `diff` command) on the output files from Haskell and HamSkill standard. If the two files are identical, then mark the test as passing; if they are different, the test fails.
- **Step #5:** Repeat steps # 5.2 and # 5.2 using HamSkill+.

There is an additional function in the **bash** test bench named “`print_final_results`” that prints the final results (i.e. number of passing tests versus the total number of tests).

6 Conclusions

HamSkill provides a foundation for running a dialect of Haskell inside the JVM. All of the original key features outlined in the proposal were implemented as well as additional features including Monads and the Scala output reformatter. While *HamSkill* does have limitation as all languages do (in particular those with a scope as limited as this), it is viewed by the development team as a very successful project that has exceeded our expectations.

Bibliography

- [1] A gentle introduction to haskell: Modules. <https://www.haskell.org/tutorial/modules.html>. (Accessed on 03/01/2016).
- [2] Home jruby.org. <http://jruby.org/>. (Accessed on 02/24/2016).
- [3] Platform specific notes. <http://www.jython.org/archive/21/platform.html>. (Accessed on 02/25/2016).
- [4] Renjin.org — about. <http://www.renjin.org/about.html>. (Accessed on 02/25/2016).
- [5] The scala programming language. <http://www.scala-lang.org/>. (Accessed on 02/25/2016).
- [6] What is scala? — the scala programming language. <http://www.scala-lang.org/what-is-scala.html>. (Accessed on 03/30/2016).
- [7] Remo H. Jansen. *Learning TypeScript*. Packt Publishing, 2015.
- [8] Terence Parr. *The Definitive ANTLR 4 Reference*. The Pragmatic Programmers, 2012.
- [9] Ingo Wechsung. Frege/frege: Frege is a haskell for the jvm. it brings purely functional programing to the java platform. <https://github.com/Frege/frege>. (Accessed on 02/25/2016).