Masaryk University
Faculty of Informatics

An Executable Formal Semantics of C++

Master's Thesis

Jan Tušil

Brno, Fall 2017

Replace this page with a copy of the official signed thesis assignment and a copy of the Statement of an Author.

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Jan Tušil

Advisor: Jan Strejček

Abstract

«abstract»

Keywords

C++ semantics k-framework

Contents

List of Tables

List of Figures

1 Introduction

Writing correct software is hard. Although formal methods for software verification are being developed, there are only a few high-quality tools on the market. In order to create a production-ready tool based on a particular set of formal methods and targeting a particular language, it is necessary to understand the formal methods, the precise semantics of the target language, and possibly some other things. Researchers and engineers often work in teams, and various teams often implement different techniques for the same language; because of that, a platform which would enable component reuse and separation of concerns might significantly improve the productivity of the tool developers.

𝕂 framework is one such platform. It is based on the idea that formal, executable language semantics can be used to derive a large variety of tools, including interpreters, debuggers, model checkers or deductive program verifiers [1] (see Figure 1.1). Tool developers skilled in a particular area of formal methods can work inside their area of expertise and focus on developing language independent tools, while leaving language details to someone else. Thus a separation of concerns is achieved.

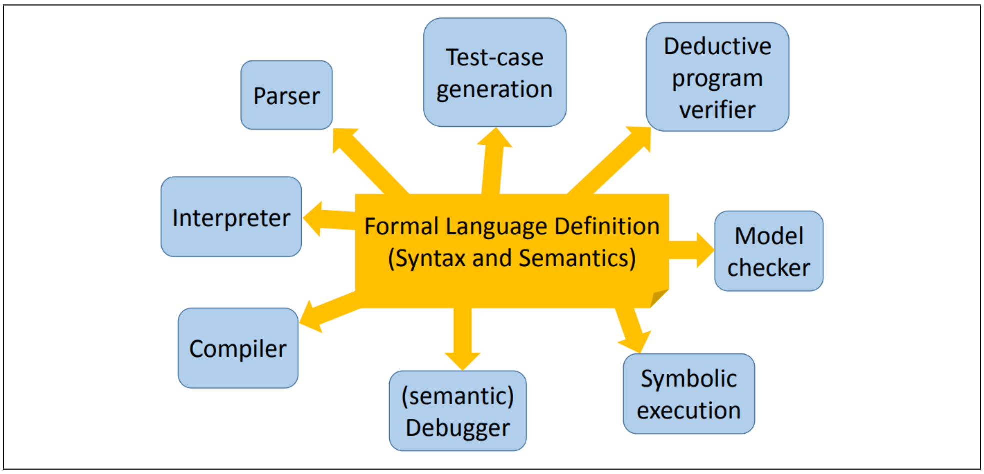

Figure 1.1: The idea behind 𝕂. Adopted from [2].

𝕂 framework has been successfully used to give formal semantics to various languages, including Java [3], Python, Javascript [4], Agda [5], and C [6, 7], all of which is publicly available. The C semantics has been used to create RV-Match, a „tool for checking C programs for undefined behavior and other common programmer mistakes" [8]. At the time of writing, the C semantics is being extended to support the C++ programming language. The C++ support is also the focus of this thesis. From this point on, when we write „Project", we mean this project of defining C/C++ language semantics in 𝕂.

The thesis aimed to implement two language features, *enumerations* and *compile-time evaluation of constexpr functions*, where the latter is an important part of a language trait called *constant expressions*. When we decided to implement the enumerations, we thought that it is a relatively simple language feature: they do not have any special runtime behavior, do not interfere much with other language features, and they exist in the language from its beginning. *Scoped enumerations*, introduced in C++11 [9], are even simpler both from language and user point of view (the legacy C-style enumerations still need to be supported, though). However, careful reading of the standard [10] revealed us that the enumerations behave in many ways, some of which we considered to be surprising, as we did not know them from common programming practice.

Constant expressions, on the contrary, had undergone a deep change in C++11, which allowed a restricted set of runtime computations to happen in the compilation time; the future revisions released the restrictions to the point that in C++17 [11], almost arbitrary side effect free computations can happen in the time of compilation. Because of that, we expected from the beginning that in order to implement the constant expressions (and *compile-time function evaluation* in particular) a more fundamental change to the semantics would have to be made.

Both features were successfully implemented. The implementation of enumerations is production ready and was largely merged into the upstream; the compile-time function evaluation will be probably merged later, when the precise semantics of constant expressions is implemented. In addition, we fixed several bugs in the project, including one hard-to-debug nondeterministic behavior, and implemented *zero initialization* for objects of classes types.

The purpose of this text is to describe the implementation of the aforementioned features. It is organized as follows. First, the $\mathbb{K}$ framework is described in such level of detail, which is enough for the reader to understand the implementation of the features and which enables him to experiment with simple language definitions. Second, basic concepts of the C++ language are outlined, followed by more in-depth discussion of the selected language features. Third, the general architecture of the Project is described; fourth, the implementation of the language features is described. The final sections of the text contains an evaluation of the implementation, with a discussion of possible future work.

2 K framework

This chapter intends to give a brief overview of the $\mathbb{K}$ framework. The level of detail here is necessary to understand the description of the Implementation section and does not go much deeper; an inquisitive reader is encouraged to go through the K tutorial.

In the $\mathbb{K}$ framework, languages are described in a style commonly known as *operational semantics*. For any language L, when L is given a particular definition D in $\mathbb{K}$ framework, then D assigns to every program in L a transition system $(Cfg, \rightarrow)$. Here Cfg denotes a set of program configurations and $\rightarrow$ is a binary relation over Cfg; the relation is called a „transition relation" and its elements are „transitions".

The configurations are not abstract, but they have an internal structure, which depends on the definition D. For imperative languages, the configurations may consists of a „program" part and a „data" part.

2.1 Terms

The program part can be represented as a term. $\mathbb{K}$ allows to define a multisorted algebraic signature (S, Σ); closed terms over this signature forms a (multisorted) term algebra. One may then choose a particular sort $s \in \Sigma$ and declare the set of all programs to be the set of all closed terms of the sort s.

Sorts are defined using `syntax` keyword. The definition

```
syntax H
syntax H ::= world()
syntax H ::= hello(H,H)
```

defines sort `H` and a nullary constructor `world` and a binary constructor `hello` of that sort. Uknown sorts on the left hand side of the operator `::=` are automatically defined, and when defining multiple constructors for one sort, the right hand sides can be chained with the operator `|`. The definition above is therefore equivalent to the following definition:

```
syntax H ::= world() | hello(H,H)
```

Furthermore, $\mathbb{K}$ allows the sort constructors to be given in an infix syntax and to contain „special" symbols. Therefore, instead of

```
syntax G ::= unit() | add(G, G) | inv(G)
```

it is possible to write

```
syntax G ::= ".G" | G "+" G | "-" G
```

to describe the signature of groups. 𝕂 also supports subsorting; the following definition makes the sort `Int` a subsort of the sort `Real`:

```
syntax Real ::= Int
```

```
syntax AExp  ::= Int | Id
               | AExp "+" AExp
syntax BExp  ::= Bool
               | AExp "<=" AExp
               | "!" BExp
               | BExp "&&" BExp
               | "(" BExp ")"
syntax Block ::= "{" "}"
               | "{" Stmt "}"
syntax Stmt  ::= Block
               | Id "=" AExp ";"
               | "while" "(" BExp ")" Block
               | Stmt Stmt
syntax Pgm   ::= Stmt
```

Figure 2.1: A definition of an algebraic signature of an imperative language.

𝕂 is distributed together with a basic library; together they provide a number of pre-defined sorts, including `Id` `Int`, `Bool`, `List`, and `Map`. The `Id` is a sort of C-style identifiers, the other names are rather self-explanatory. With use of these sorts, an algebraic signature of an imperative language can be defined as in Figure 2.1. From the definition, *arithmetic expressions* (represented by the sort `AExp`) are integers, identifiers, and sums of other arithmetic expressions; the arithmetic expressions can be compared to create *boolean expressions*, and so on.

2.2 Configurations

A part of a program configuration can be stored in a *cell*, which can be thought of as a labeled multiset [7]. Cells may contain terms of a given

```
configuration <T>
                <k> $PGM:Pgm </k>
                <state> .Map </state>
              </T>
```

Figure 2.2: A definition of the initial configuration of the language Imp.

sort or other cells. In a source code of a language definition in $\mathbb{K}$, cells are written in an xml-style notation.

Figure 2.2 contains a snippet of such source code. The keyword `configuration` here defines three cells (`T`, `k` and `state`), a single structure for all configurations, and an initial configuration. Every cell in the definition has some content: the `state` cell contains `.Map`, which has a constructor of sort `Map`, representing an empty map; the `k` cell contains a term of sort `Pgm` consisting of a variable with name `$PGM`, and the `T` cell contains the other two cells. The initial configuration for program `P` is just like that, except that the variable `$PGM` is replaced by a term, representing the program `P`.

We have said earlier that the programmer may choose a sort, whose terms will represent the „program" part of the configuration; that happens in the configuration definition. In this particular language definition, the „data" part of the configuration is represented by a term of sort `Map`.

2.3 Implementations of $\mathbb{K}$

Before going further, let us note that in the time of writing there exist two implementations of $\mathbb{K}$: UUIC-$\mathbb{K}$[1], developed by University of Illinois in Urbana-Campaign, and RV-$\mathbb{K}$[2], developed by RuntimeVerification, inc. The former is intended to be the reference implementation, the latter is optimized for RV's tools based on the c-semantics. RV-$\mathbb{K}$ differs from UIUC-$\mathbb{K}$ in a few aspects. One of the differences is that the RV-$\mathbb{K}$ requires the configuration to contain a cell with an `exit` attribute: the cell contains the value returned by `krun`. Because of that,

1. `https://github.com/kframework/k`
2. `https://github.com/runtimeverification/k`

```
module ARITH-SYNTAX
  syntax AExp  ::= Int
                 | AExp "/" AExp                [left]
                 > AExp "+" AExp                [left]
                 | "(" AExp ")"                 [bracket]
  syntax Pgm ::= AExp
endmodule
module ARITH
  imports ARITH-SYNTAX
  configuration <T>
                   <k> $PGM:Pgm </k>
                   <ret exit="">0</ret>
                </T>
endmodule
```

Figure 2.3: File `arith.k`. Note the `ret` cell with an `exit` attribute.

it does not supports the examples included in the official 𝕂 tutorial. However, all examples in this chapter from this point further work with both implementations.

2.4 Parsing

It can be seen from the Figure 2.1 that definitions of algebraic signatures in 𝕂 looks similar to definitions of context-free grammars using a BNF notation. This is not a coincidence; in fact, 𝕂 allows to specify the *concrete syntax* with use of special *attributes*. This subsection describes some of the parsing facilities 𝕂 provides; the Project does not use them, but the description enables the reader to easily experiment with the provided examples.

Figure 2.3 contains a definition of a simple language of arithmetic expressions. The definition contains a few things we have not described yet. First, it is split into two modules; the `ARITH-SYNTAX` module contains everything which is needed to generate a parser, while the `ARITH` module contains everything else. Second, one of the sort constructors is separated from the previous one by a > sign; that causes the preceding productions in the concrete syntax grammar to have a higher priority than the following ones and thus to bind tighter. Third, some constructors have *attributes* attached; the `left` attributes causes a binary constructor to be left-associative, and the `bracket` attribute means that the corresponding unary constructor should be parsed as a pair of brackets.

When stored in a file with name `arith.k`, the definition may be compiled using UIUC-𝕂 (the first command) or using RV-𝕂 (the second one):

```
$ kompile arith.k
$ kompile -O2 arith.k
```

The `krun` command will use the module `ARITH-SYNTAX` to generate the parser and the module `ARITH` to generate an interpreter. The names of the modules used can be specified by command-line arguments; if they are not specified (as above), 𝕂 infers them from the filename. Both parser and interpreter are stored in a directory `arith-kompiled`.

The compiled definition can be used to parse and execute a program written in the arithmetic language.

```
$ cat addition.arith
1 + 2 + 3
$ krun addition.arith
<T> <k> 1 + 2 + 3 </k> <ret> 0 </ret> </T>
```

The command parses the given source file, creates an initial configuration, walks in the generated transition system until it reaches a terminal configuration, and pretty-prints it. The generated transition system contains only one configuration (the initial one), as the current language definition of Arith does not have any semantic rules.

The pretty-printing implies that the abstract syntax is *unparsed* back to the concrete one; therefore, from the output above one can not conclude that the file was parsed correctly, as changing the left associativity to the right one (by changing `left` attributes to `right` ones) would not alter the output. However, `krun` can be instructed to output a textual version of its internal representation:

```
$ krun --output addition.kast
<T>`(`<k>`(`_+__ARITH-SYNTAX`(`_+__ARITH-SYNTAX`(
#token("1","Int"),#token("2","Int")),#token("3","Int"))),
`<ret>`(#token("0","Int")))
```

The output is easily readable for a machine; to parse it, the generated concrete syntax parser is not needed. It is less readable for humans, though. From the output one can easily get

```
$ cat simple.arith
5/2 + (1 + 3) / 2
$ krun simple.arith
<T> <k> 5 / 2 + ( 1 + 3 ) / 2 </k> <ret> 0 </ret> </T>
$ krun --output kast simple.arith
<T>`(`<k>`(`_+__ARITH-SYNTAX`(`_/__ARITH-SYNTAX`(
#token("5","Int"),#token("2","Int")),`_/__ARITH-SYNTAX`(`
_+__ARITH-SYNTAX`(#token("1","Int"),#token("3","Int")),
#token("2","Int")))),`<ret>`(#token("0","Int")))
```

Figure 2.4: Another example of a program in the language of arithmetic expressions. The last output corresponds to the expression `+( /( 5, 2 ), /( +( 1, 3 ), 2 ) )`

```
+( +( 1, 2 ), 3 )
```

simply be keeping only the content of `k` cell, removing superfluous characters and adding spaces. One can see that this expression, when interpreted as in prefix-notation, correspond to the content of `addition.c`. From a different example on Figure 2.4 one can see that the division has a priority over addition and the parenthesis bind the tightest.

2.5 Rules

So far, the definition of language Arith was able to generate only trivial transition systems with one configuration and no transitions. To generate configurations from the initial one, $\mathbb{K}$ provides a concept of *semantic rules*. In their simplest version, the rules have the form of $\varphi \Rightarrow \psi$, where φ, ψ are *patterns* - configurations with free variables, where every variable free in ψ have to be free in φ. We say that a pattern φ *matches* a concrete configuration *Cfg*, when φ may be turned into *Cfg* by substituting free variables of φ with concrete terms. In that case we say that the variables *bind* to the corresponding terms. When φ matches a configuration *Cfg*, a transition is generated into a new configuration Cfg', which is the result of substituting the free variables of ψ with the bound terms.

2.5.1 Basic rules

For example, the module ARITH of the language Arith may be extended with the following rule:

```
rule <T><k>I1:Int + I2:Int</k><ret>R:Int</ret></T>
  => <T><k>I1 +Int I2</k><ret>R</ret></T>
```

The left hand side of the rewriting operator (=>) contains three free variables (I1, I2, R), all of which required to have a sort Int. The variables I1, I2 are used as the two parameters of the constructor +, while the variable R represents the content of the ret cell. On the right hand side, I1 and I2 are given as parameters to built-in function +Int, which implements the addition of two integers; the variable R is used in the same place as on the left side, thus leaving the ret cell unchanged.

The compiled definition takes two numbers and adds them together:

```
$ kompile arith.k
$ cat simple.arith
1 + 2
$ krun simple.arith
<T> <k> 3 </k> <ret> 0 </ret> </T>
```

Here krun again printed the final configuration. It is possible to print an *i*-th configuration with use of the depth switch:

```
$ krun --depth 0 simple.arith
<T> <k> 1 + 2 </k> <ret> 0 </ret> </T>
$ krun --depth 1 simple.arith
<T> <k> 3 </k> <ret> 0 </ret> </T>
```

2.5.2 Local rewriting

The semantic rule above applies the rewriting operator (=>) to whole configuration, although it changes only the content of the k cell. 𝕂 implements a concept called *local rewriting*, which allows language definition developers to use the rewriting operator inside a cell or inside a term. With use of local rewriting, the above rule can be written as:

```
rule <T> <k> I1:Int + I2:Int => I1 +Int I2 </k>
         <ret> R:Int </ret> </T>
```

```
rule <k> I1:Int + I2:Int => I1 +Int I2 </k>
```

Figure 2.5: A rule for addition of two integers.

The rule can be even more simplified with use of an anonymous free variable, denoted by an underscore (_):

```
rule <T> <k> I1:Int + I2:Int => I1 +Int I2 </k>
         <ret> _ </ret> </T>
```

It is also possible to use the rewriting operator multiple times in one semantic rule:

```
rule <T> <k> (I1:Int => 0) + (I2:Int => I1 +Int I2) </k>
         <ret> _ </ret> </T> requires I1 =/=Int 0
rule <T> <k> (0 + I:Int) => I </k> <ret> _ </ret> </T>
```

A `requires` clause causes a rule to apply only if a certain condition holds; in this particular example, the rule should not apply if the variable `I1` is bound to zero. Without the clause, the rule would be able to apply indefinitely, without any effect.

2.5.3 Configuration abstraction

Semantic rules usually need to be aware only of a few configuration cells. In 𝕂, the semantics rules have to mention only the cells they really need to mention. The 𝕂 tool then, from the definition of configuration, infers the context in which such local rewriting takes place. The rule for addition can be written as in Figure 2.5. Such semantics rules are not only shorter and easier to write, but they are also independent on most of the configuration; when the structure of configuration changes, the rules may remain the same. This feature is called *configuration abstraction*.

2.6 Computations

When creating a language definition, it is often needed to compute something first, and then use the result of the computation to compute something else. In 𝕂, the notion of *first* and *then* is formalized in terms of *computations* and their *chaining*. Computations have the sort `K`; all user-defined sorts are automatically subsorted to `K`. Computations can be composed using the `~>` constructor, which is associative. The sort `K` has a nullary constructor `.K`, which represents an empty computation and acts as a unit with respect to `~>`. Thus `K` with `~>` and `.K` form a monoid.

In practice, the monoidal structure means that any term consisting of computations and the `~>` constructor behave as a *chain* of computations, with hidden empty computations everywhere inside. One can then insert a computation `c` to any position in the chain simply by rewriting an empty computation on that position `c`. It also allows replacing the rule 2.5 with

```
rule (<k> I1:Int + I2:Int => I1 +Int I2) ~> _</k>
```

without losing any existing behavior. If the old rule matches a configuration with a term `C` in the `k` cell, then the new rule matches the term `C ~> .K`, as the anonymous variable binds to the empty computation. But the new rule matches also the term `C`, because due the monoidal structure, configurations `C` and `C ~> .K` are equal. On the other hand, the new rule adds some behaviors, because it matches not only when the `k` cell contains exactly an addition of two integers, but also when it contains any sequence of computations, where the first computation is an addition of two integers. The first computation in the list is called *the top*. The parenthesis in the new rule are needed, as the constructor `~>` binds tighter then the operator `=>`.

The rules of the form

```
rule <k> SomeRewritingHere ~> _ </k>
```

can be also written as

```
rule SomeRewritingHere
```

For example, the rule

```
rule (<k> I1:Int + I2:Int => I1 +Int I2) ~> _</k>
```

can be also written as

```
rule I1:Int + I2:Int => I1 +Int I2
```

2.7 Strictness and evaluation strategies

How to compute the sum of three numbers, say $1 + 2 + 3$? In the language Arith, the + constructor is left-associative, so the natural approach is to compute $1 + 2$ first, which yields a result r, and then to compute $r + 3$. Although it is possible to write such rules manually, 𝕂 provides a number of tools, which enable the programmer to describe such computations on higher level of abstraction.

2.7.1 Sort predicates

For every sort `Srt` 𝕂 automatically generates a *sort predicate* `isSrt` of sort `Bool`. The predicate takes any term and returns `true` if the term is of sort `Srt`, otherwise returns `false`. One may override the default implementation of a sort predicate by writing a custom rule.

So far, the definition of Arith language can not evaluate nested expressions, as the built-in function `+Int` can be applied only on terms of sort Int. Moreover, the left side of the rule requires the involved terms to be of sort `Int` With use of sort predicate, a rule

```
rule <l> E1:AExp + E2:AExp => -1 </l>
requires notBool isInt(E1) orBool notBool isInt(E2)
```

can be added into the language definition; the rule rewrites a sum of two terms of sort `AExp` to `-1`, unless both of the terms have the sort `Int`. The `notBool` and `orBool` are built-in functions; `isInt` is a sort predicate for the sort `Int`.

2.7.2 Heating/cooling

Using the information above, a programmer may use the piece of 𝕂 code in Figure 2.6 to evaluate nested expressions. The idea here is that the left addend is evaluated first, then the right addend is evaluated and finally the two integers are added using the built-in function `+Int`. One way to interpret the rules is that the first two rules extract an unevaluated expression out of the addition, which creates a hole in the term, the extracted expression is then evaluated, and the third and forth rule plug the evaluated expression back into the hole. The constructors `holdAddR` and `holdAddL` are used to represent the original term without the extracted subterm; the sort `KItem` is a bit special, but

```
syntax KItem ::= holdAddR(AExp) | holdAddL(Int)
rule E1:AExp + E2:AExp => E1 ~> holdAddR(E2)
     requires notBool isInt(E1)
rule E1:Int + E2:AExp => E2 ~> holdAddL(E1)
     requires notBool isInt(E2)
rule E2:Int ~> holdAddL(E1:Int) => E1 + E2
rule E1:Int ~> holdAddR(E2) => E1 + E2
rule I1:Int + I2:Int => I1 +Int I2
```

Figure 2.6: Heating and cooling rules, written manually.

it is subsorted to the sort `K` as any user-defined sort. The process of extracting a subterm is known as *heating*, while the opposite process is *cooling*.

2.7.3 Evaluation contexts

Heating and cooling rules are very common; they are also tedious to write manually. In $\mathbb{K}$, the idea of evaluating a certain subterm first can be expressed in terms of *evaluation context*. With use of a keyword `context`, the $\mathbb{K}$ code

```
syntax KItem ::= holdAddR(AExp)
rule E1:AExp + E2:AExp => E1 ~> holdAddR(E2)
     requires notBool isInt(E1)
rule E2:Int ~> holdAddL(E1:Int) => E1 + E2
```

can be equivalently expressed as:

```
context HOLE:AExp + E:AExp [result(Int)]
```

The `context` declaration means exactly that: whenever the top of a `k` cell contains an addition of two `AExp`s and the first one is not of sort `Int`, extract the first one, push it on the top of the `k` cell and replace the addition with some placeholder; when the extracted subterm gets evaluated to `Int`, plug it back to the original context.

With use of context, the code on Figure 2.6 can be equivalently expressed as

```
context HOLE:AExp + E:AExp [result(Int)]
context I:Int + HOLE:AExp [result(Int)]
rule I1:Int + I2:Int => I1 +Int I2
```

```
3 + 1 + 7
( 3 + 1 ) ~> #freezer_+__ARITH-SYNTAX1_ ( 7 )
4 ~> #freezer_+__ARITH-SYNTAX1_ ( 7 )
4 + 7
11
```

Figure 2.7: The progress of evaluating the expression $3 + 1 + 7$. An ith line represents the configuration after i computational steps, starting from zero. For brevity, only the content of the `k` cell is shown.

which significantly reduces the amount of code. When compiled, the generated interpreter correctly evaluates the expression $3 + 1 + 7$; the evaluation progress is shown in Figure 2.7.

2.7.4 Strictness attributes

$\mathbb{K}$ defines a sort `KResult`, which represents results of computations. When no `result` attribute is given to a context declaration, the declaration behaves as with `[result(KResult)]`. When writing a larger language definition, it is convenient to identify the sorts of desired results and subsort them to `KResult`.

When defining a sort constructor, the constructor may be tagged with an attribute `strict`. In that case $\mathbb{K}$ generates a context declaration for every parameter of the constructor. With use of the `strict` attribute, a definition of the language Arith can be given using only a few lines (Figure 2.8).

2.8 Functions

In the preceding text, a number of built-in functions was used (e.g. `andBool`, `=/=Int` and `/Int`). Functions can appear in side conditions (the `requires` clause) or right hand sides of the rewriting operator. Their application does not create any transitions in the transition system and unlike constructors, functions cannot be pattern-matched. Custom functions can be defined with `syntax` keyword and a `function` attribute; their behavior is defined using the rewriting operator. See Figure 2.9 for an example.

```
module ARITH-SYNTAX
  syntax AExp  ::= Int
                 | AExp "/" AExp [left, strict]
                 > AExp "+" AExp [left, strict]
                 | "(" AExp ")"  [bracket]
  syntax Pgm ::= AExp
endmodule
module ARITH
  imports ARITH-SYNTAX
  configuration <T>
                  <k> $PGM:Pgm </k>
                  <ret exit=""> 0 </ret>
                </T>
  syntax KResult ::= Int
  rule I1:Int + I2:Int => I1 +Int I2
  rule I1:Int / I2:Int => I1 /Int I2 requires I2 =/=Int 0
endmodule
```

Figure 2.8: A definition of the language Arith.

```
syntax Int ::= eval(AExp) [function]
rule eval(I:Int) => I
rule eval(E1:AExp + E2:AExp) => eval(E1) +Int eval(E2)
rule eval(E1:AExp / E2:AExp) => eval(E1) /Int eval(E2)
```

Figure 2.9: An `eval` function.

3 C++

C++ is a „general purpose programming language" [12] originally created by Bjarne Stroustrup in the years between 1979 and 1983 [13]. After some time, many independent implementation emerged, and in 1998 the language was standardized as ISO/IEC 14882:1998 international standard ([14]); that version is now known as C++98. The language changed a lot since that time. New compilers emerged (e.g Clang), compilers implemented a lot of experimental features (e.g. Clang's Modules[1]), many core language and standard library defects were fixed and the memory model was standardized. The language is gradually evolving and once in a while, the C++ Standards Committee [15] emits an ISO standard. In the time of writing, the newest C++ standard is C++17 [11], which was technically completed in March 2017 and should be officially published in December 2017.

3.1 Standard documents

Although the most videly known documents created by the committee are the international standards, the committee is working all the time. The committe produces a large amount of documents, most of which are publicly available [16]; also, every document is assigned an unique identifier, which is then used to reference the document. Among the publicly available documents are also *workings drafts* of the ISO standard. The working draft is hosted on GitHub[2]; it therefore possible to build it directly from its LaTeX source or to download the emitted releases on the Release page.

In this thesis we mostly use the document `n4296` [10], which is the first draft released after the C++14 ISO standard ([17]); this document is also referenced in the source codes of the Project. We will also refer to the post-C++17 draft `n4700` [10], because its wording is much cleaner then C++14's.

1. `https://clang.llvm.org/docs/Modules.html`
2. `https://github.com/cplusplus/draft/`

3.2 Fundamentals

This section describes some of the fundamental concepts of the C++ language.

In C++, the memory consists of one or multiple sequences of contiguous bytes, where every *byte* has an unique *address* [18, §4.4/1]. Various constructs of the language can create and destroy *objects*; an object occupies a region of storage, not necessarily contiguous [18, §4.5/1]. Objects can contain *subobjects*; an objects not contained in any other objects is a *complete object*. Every complete object has to occupy at least one byte of the storage; therefore, we can think of objects as having an *identity* (we can identify them by the first byte of their storage).

FIXME: this is not true, see [10, 3.9/4]. Objects are used to hold *values*. Although the standard often refers to values, it does not specify what a *value* is (at least until [18]). One can think of value as of a *datum* together with its *interpretation*; an interpretation of a datum is some concrete, real-world entity[3]. More in-depth discussion about entities and their computer representations can be found in the first chapter of [19].

A `variable` can be either an object, or a *reference* [18, §6/6]. Reference is not an object; unlike objects, references does not need to occupy storage and do not have an identity.

An *expression* is a language construct, which can be evaluated to get a value and to cause a side-effect. Every expression has an associated *value category*: it can be either a *glvalue*, which means that an evaluation of the expression determines an identity of an object (or function or bitfield), or a *prvalue*, which means that the evaluation initializes an object or computes a value [18, 6.10/1]. The category of *glvalue*s is further divided into *lvalue*s and *xvalue*s.

3.3 Enumerations

Many languages implement enumerated types as a means for programmers to define a finite set of related values. Enumerations in the

3. This notion of *concrete entity* is something different than C++ *entity* as defined in the C++ language [10, §3/3].

```
enum E { A = 5, B = A + 3 };
enum E e = B;
```

Figure 3.1: A declaration of an *unscoped enumeration* `E` with *unscoped enumerators* `A` and `B`, followed by a declaration of a variable `e` of type `E`, initialized by the enumerator `B`. This code is valid in both C++14 and C99.

C++ language are based on enumerations from the C language. In this section we give a brief presentation of the C++ enumerations, as defined in [10]. Note that this text is concerned with what the language allows, rather than what is consider to be a good practice. For a discussion about the latter we point the reader to the relevant section of CppCoreGuidelines[4].

3.3.1 Unscoped enumerations

Figure 3.1 shows a basic C-style declaration of an enumeration `E`, with subsequent declaration of an object of the enumeration's type. Both declarations are valid in both C and C++. In C++14, the C-style enumerations are called *unscoped enumerations*. An enumeration may contain an unbounded number of named values, called *enumerators*. Here, `E` contains two enumerators: `A` and `B`. In C++14, enumerators of an unscoped enumeration are called *unscoped enumerators* and they are declared in the scope which contains the enumeration declaration (or, in the terms of the standard, the *enum-specifier* , . Therefore, it is possible to refer to them simply as in this example.

Because of this simplicity it is not possible to declare two unscoped enumerations with the same enumerator, in the same scope. It is still possible to create the second enumeration in a different scope, though; the enumeration may even have the same name, as in the Figure 3.2. If that happens, the old enumerator is *shadowed* with the new one, but it is still possible to access the old one using the scope resolution operator (i.e. `::B` in this example).

When a name of an enumeration (or a struct) is used in C, for example in a declaration of a variable of that enumeration type, it is

4. `http://isocpp.github.io/CppCoreGuidelines/CppCoreGuidelines`

```
enum E{B};
// enum F{B}; // an error
int main() {
    enum E{A,B};
    return (int)B;
}
```

Figure 3.2: A declaration of an enumeration with the same enumerator as in an already existing enumeration.

```
enum class E { A = 5, B = A + 3 };
E e = E::B;
```

Figure 3.3: A declaration of a *scoped enumeration* `E` with *scoped enumerators* `A` and `B`.

necessary to use the `enum` keyword (as in Figure 3.1). In C++, this is not necessary, as a name of an enumeration (`enum-name`) denotes a type (it is a `type-name`, see [10, Annex A.1],). It is still possible to do it the C way, though; the type names accompanied by the keyword `enum` or `struct` (or similar) are called *elaborated type specifiers*[10, §3.4.4], .

3.3.2 Scoped enumerations

The revision C++11 introduced `scoped enumerations`. They are declared using the phrase `enum class` or `enum struct`, and enumerators of a scoped enumeration are called *scoped enumerators*. Scoped enumerations differ from the plain enumerations in a number of aspects; the most significant difference is what they have their name for: scoped enumerators are declared in the scope of their enumeration[10, §7.2:11], . Because of that, the enumerators have to be referred to using the scope resolution operator (at least outside the enumeration declaration, see Figure 3.3). However, unscoped enumerators can be referred to this way, too ([10, §5.1.1/11],).

3.3.3 Underlying type

Every enumeration type has associated an `underlying type`. It is an integral type, which can represent all values of the enumeration; it also defines the size of the enumeration[5]. The standard library provides the programmer a way to find out the underlying type of an enumeration (Figure 3.4).

```
#include <type_traits>
enum E {A, B, C};
std::underlying_type_t<E> x = 5;
```

Figure 3.4: The type of the variable `x` is the underlying type of `E`.

An enumeration's underlying type can be explicitly specified [10, §7.2:5], . This is achieved by appending a colon followed by a name of an integral type right after the `enum-name` in the declaration, as shown in Figure 3.5. The most obvious reason why would programmer want to specify an enum's underlying type is in order to limit the enumeration's size, perhaps because of hardware constrains. Before C++11, this was often achieved through an extension of a particular compiler; for example, the GNU gcc compiler provides a special attribute `packed`, which causes an enumeration's size to be as small as possible (Figure 3.6).

If the underlying type of a scoped enumeration is not explicitly specified, it is automatically defined to `int`. All scoped enumerations, as well as unscoped enumerations with explicitly specified underlying type, are said to have a *fixed underlying type*. This stands in contrast to unscoped enumeration without explicitly specified underlying type. Their underlying type is not fixed, but it is an *implementation-defined*

5. Technically, it is more complicated: `https://stackoverflow.com/q/47444081/6209703`

```
enum E1 : char {A,B};
enum class E2 : short {C,D};
```

Figure 3.5: Explicitly specified underlying type.

```
enum E { A, B } __attribute__((packed));
```

Figure 3.6: A packed enum.

integral type, which can represent all values of the enumeration [10, §7.2:7], . That also suggests that the underlying type is not known prior the closing brace of the declaration.

The fact that the standard does not specify the exact underlying type for enumerations whose underlying type is not fixed is also the reason why in pre-C++11 (including C11), the enumerations whose size matters to the programmer have to be declared with use of compiler-specific language constructs (as in Figure 3.6). Without that, the compiler could simply choose the underlying type (and thus the size) of the enumeration (with respect to some standard-given restrictions); in practice, the chosen underlying type is often `int`.

3.3.4 Values of an enumeration

The standard also defines the set of values of an enumeration; those are the values, which can be written or read from an object of the enumeration's type. For an enumeration whose underlying type is not fixed, the set of values limits the choice of the underlying type; it is defined such that the values of all its enumerators and all values between the enumerators are also values of the enumeration. The definition is rather technical, though; an inquisitive reader shall find it in [10, §7.2:8], . For an enumeration with fixed underlying type, the values of the enumeration are exactly the values of its underlying type.

This behavior may be a source of confusion for programmers who do not know this; in past, it certainly was for the author of this text. The programmer may write a code similar to the one in Figure 3.7, which assumes that the value of `e` has to be either `E::A` or `E::B`. This assumption may be justified, if the programmer knows the set of values which may be returned from the function `giveMeAnEnum`. The programmer may try to go one step further and say: „If the function `giveMeAnEnum` tries to return something different than `E::A` or `E::B`, than the behavior is already undefined in *that* point; therefore, we do not need to handle that situation in `foo`." Such a reasoning is flawed, because the

```
enum class E { A, B };
E giveMeAnEnum();
int foo(E e) {
        E e = giveMeAnEnum();
        switch(e) {
                case E::A: return 3;
                case E::B: return 42;
        }
        /* Unreachable? Well... */
}
```

Figure 3.7: The reachability of the last line depends on whether `E` has a fixed underlying type or not.

function `giveMeAnEnum` may return e.g `static_cast<E>(-117)`, which is guaranteed to be well-defined, because `-117` is a valid value for `int`, which is the underlying type of `E` here. The programmer's reasoning would be valid if the enumeration `E` was a plain unscoped enumeration without explicitly specified underlying type (i.e if we removed the `class` keyword). We will not go into detail here, as the reasoning would depend on the technical parts of [10, §7.2:8], .

3.3.5 Enumerators

Declaration of an enumerator may contain an optional *initializer*, which have to be implicitly convertible to an integral type[6]. If the initializer is not specified, the value of the enumerator is zero if it is the first enumerator, otherwise it is the value of the previous enumerator plus one. When an enumeration is fully defined („following the closing brace of an enum-specifier"), the type of each enumerator is the type of the enumerator [10, §7.2:5], . For example, in the declaration of variable `e` in the Figure 3.3, the type of the expression `E::B` is `E`.

This is different from C, where enumerators have the type `int`. Still, unscoped enumerations can be implicitly converted to `int` [10, §7.2:10], , so they behave somewhat similarly to C. This conversion does not work for scoped enumerations. The other conversion (from

6. It is more complicated, but this is enough here.

```
enum E { A };
void foo(int){}
void goo() {
    foo(A);
    foo(E::A);
    foo(E::A+1};
}
```

Figure 3.8: Unscoped enumerations can be implicitly converted to `int`.

```
enum E { A };
void foo(int x) { E e = x; }
```

Figure 3.9: C++ does not allow an implicit conversion from int to enum. However, this is a valid C code.

integer types to enumerations) is not defined neither for scoped nor unscoped enumerations (Figure 3.9).

An initializer of an enumerator may refer to previously declared enumerators of the same enumeration (see 3.3). Inside the enumeration declaration (prior its closing brace), the type of its enumerators is not the type of the enumerator, but always an integral type[7]. Because of the stronger-than-C type system, in some cases it would be even inconvenient if the types of enumerators were types of their enumerations. For example, the declaration in Figure 3.3 declares an enumerator `B` with the initializer `A + 3`. If the type of `A` in this expression were `E`, then the expression would be ill-formed, because the scoped enumeration does not convert to int, so it cannot be added to an int.

3.3.6 Incomplete types

Some types cannot be used to define objects; those are called *incomplete* [10, §3.1:5], . A type may be incomplete at one point and complete at other point in a translation unit. Incomplete types are still useful; they may be used in function declarations to denote its parameters and return values, to define objects of pointer type or references. There are two categories of incomplete types: void types (e.g. `volatile void`)

7. We discuss this in more detail in Chapter 6

```
enum E1 : char;
enum class E2 : unsigned int;
enum class E3;
```

Figure 3.10: *Opaque-enum-declaration*s.

```
enum E;
```

Figure 3.11: Not a valid declaration, because „opaque-enum-declaration declaring an unscoped enumeration shall not omit the enum-base" [10, §7.2:2],

and incompletely-defined types. It is not allowed to apply the `sizeof` operator to an incomplete type. A programmer may choose to use incompletely-defined types to limit the number of dependencies between header files. An enumeration whose underlying type is not fixed is incomplete until the closing brace of its declaration.

3.3.7 Opaque enumerations

Enumerations with fixed underlying type can be declared *opaque*, that is, without the list of enumerators (see Figure 3.10). An enumeration declared using an *opaque-enum-declaration* is a complete type (it is not an incomplete type); therefore, it can be used almost as a fully declared enumeration, except that its enumerators are not available. Such an enumeration needs to have a fixed underlying type (see Figure 3.11). An opaque enumeration can be fully redeclared later in the program [10, §7.2:3], .

3.4 Initialization

As a part of this thesis we also implemented a language feature called *zero initialization* on class types. In this section we describe how variables are initialized in general, and how *zero initialization* fits into the overall scheme. The details of how we implemented *zero initialization* are in Section 5.3.

3.4.1 Initialization forms

When an identifier is being declared, an *initializer* can be used to specify an initial value of the identifier [10, §8.5/1], [18, §11.6/1]. Initializers have many syntactical forms, all of which are shown in Figure 3.12. The form of initialization `T x(a);` or `T x{a};` is known as *direct initialization* [10, §8.5/16]; this form is used in the declarations of variables `a` and `b` in the Figure 3.12.

An initialization of the form `T x{...};` or `T x = {...}` (where the elipsis are meta-characters indicating a comma-separated list) is known as *list-initialization*, where the former is *direct-list-initialization* and the latter *copy-list-initialization*. In the Figure 3.12, the variables `b` and `g` are *direct-list-initialized* and the variables `d`, `h` *copy-list-initialized*. The brace-enclosed comma-separated list is called *braced-init-list*; the initialization in declarations of the form `T a = E;` is known as *copy-initialization*.

```
int a(1);
int b{2};
int c = 3;
int d = {4};
int e{};
int f();
char g[]{'a', 'b'};
int h[3] = {1,2,3};
```

Figure 3.12: Declaration of variables `a`,`b`,`c`,`d`,`e`,`g`,`h`. The identifier `f` is not declared as a variable of type `int`, but as a function taking no parameters and returning an `int`.

It is important to note that all the previous categories consider only the form of the declaration; the semantics of such declarations heavily depends on the types involved. For example, the *copy-initialization* form is often used in a declaration of a reference; when used in a declaration of an object of a class type, a *move constructor* may be invoked instead of `copy constructor`.

3.4.2 Aggregates

The C++ language enables programmers to declare *classes*. From the programming methodology standpoint, not all classes are the same kind: for example, some classes are designed to hold values (e.g. `std::complex`); other ones encapsulate behavior and while the identities of their instances are really important to programmers, their values are not. Classes can also be categorized from another perspective: some of them maintain a non-trivial *invariant*, thus restricting the set of possible values their member variables may hold (e.g. `std::vector`); other classes consider their members to be independent of each other (e.g. `std::pair`), and the set of possible values of such class is then equal to the Cartesian product of the possible values of its members. Such types are commonly known as *product types*[8]. In C++, the notion of types composed freely from other types is present in the form of *aggregates*:

> An aggregate is an array or a class (Clause 9) with no user-provided constructors (12.1), no private or protected non-static data members (Clause 11), no base classes (Clause 10), and no virtual functions (10.3). [10, §8.5.1/1]

3.4.3 Aggregate initialization

In newer versions of the language, the restrictions on aggregate classes are more relaxed; for example, an aggregate class may have a base class, but under the condition that it is non-virtual and public [18, §11.6.1/1]. Nevertheless, the idea is still the same: an *aggregate* is „something *simply* composed of other things". Because of that, the language provides a special way to initialize an object of an aggregate type - an *aggregate initialization*.

The *aggregate initialization* happens, when an aggregate is initialized using *list-initialization*. If it happens, the *elements* of the aggregate, that is, the array elements or the non-static member variables, are initialized with the respective *initializer clauses* in the order of increasing indices, or in the order of declarations, respectively [10, §8.5.1/2]. If there are not enough initializer clauses in the initializer list, the remaining aggregate elements are initialized from an empty initializer list (see Figure 3.13).

8. More formal treatment of product types can be found in [20, Chapter 1.5].

```
struct S {
  char a;
  int b[2];
  double c;
};
S s = {'a', {2, 3}};
```

Figure 3.13: An exaple of an aggregate initialization. The member `c` is initialized with the value of the expression `double{}` , which is a floating-point zero.

3.4.4 Value initialization

When an object is created with no initializer, it is *default-initialized*, which in many situations means that no initialization is performed and the object has an undeterminate value. The language also defines a *value initialization*, whose purpose is to somehow „complement" the notion of „no initialization":

> [...] An object that is value-initialized is deemed to be constructed and thus subject to provisions of this International Standard applying to "constructed" objects, objects "for which the constructor has completed," etc., even if no constructor is invoked for the object's initialization. [10, §8.5/8]

A value-initialization is performed in various different cases. For example, the standard says

> An object whose initializer is an empty set of parentheses, i.e., (), shall be value-initialized.[...] [10, §8.5/11]

which is demostrated in Figure 3.14. (However, the declaration of `f` in Figure 3.12 does not declare an object, but a function; in that case, no value initialization is performed.) Value initialization is also performed for list-initialized objects of a class type with default constructor [10, §8.5.4/5.4]. Such class types are not aggregates, so no aggregate initialization can happen for them.

To value initialize an array means to value initialize all its elements; to value initialize an object of a non-class non-array type means to zero-initialize it. Value initialization of an object of a class type performs

```
struct A {
  T m;
  A():m() {}
};
```

Figure 3.14: A *value-initialization* of a member variable in a constructor.

```
struct A {
  int m;
};
struct B {
  virtual ~B(){}
  int m = 1;
  int n;
};
int bar() {
  A a{};
  B b1{};
  B b2;
  int c{};
}
```

Figure 3.15: Initialization from an empty initializer list.

default-initialization; the default-initialization is preceded by zero initialization if the class does not have a user-defined or deleted default constructor.

The source code in Figure 3.15 can serve as an example. The variables `a`, `b1` and `c` are list-initialized, while the variable `b2` is default-initialized; however, the meaning of the list-initialization among the variables differs. The variable `c` is value-initialized, and since it has a non-class non-array type, it is zero-initialed; after the initialization, the value of `x` is zero. The variable `a` is aggregate-initialized, because the class `A` is an aggregate; since the initializer-list is empty, its member variable `m` is initialized with the initializer `{}`, which performs zero-initialization as in the previous case. At the end, the integer `a.m` has the value of zero.

The variable `b1` is not aggregate-initialized, as the class `B` is not an aggregate, because it has a virtual member function (the destructor);

the variable `n2` is value-initialized instead. The value-initialization performs a zero-initialization, because the class `A` does not have neither user-provided nor deleted default constructor, but the default constructor is implicitly-generated. After the zero initialization, a default-initialization happens, which runs the implicitly-generated default constructor. At the end, the value of `b1.m` is one and the value of `b1.n` is zero. In contrast, the variable `b2` is only default-initialized; the value of `b2.m` is one as in the previous case, but the value of `b1.n` indeterminate.

3.4.5 Zero initialization

The exact meaning of *zero-initialization* is described in [10, §8.5/6]. References are not initialized at all, scalars are initialized to the value of zero converted to the particular scalar type, arrays are zero-initialized element-wise and objects of non-union class types are zero-initialized by zero-initializing all non-static data its members, base class subobjects and padding. We do not deal with union classes in this thesis.

In addition to the cases described in previous paragraphs, zero initialization happens in a two more situations.

- When initializing a character array with a string literal that is not long enough, the rest of the array is zero-initialized.

- Before other initializations of variables with non-static storage duration. Programmers may known this in a simplified form as the rule „in C/C++, global variables are initialized to zero".

This thesis is not concerned with these two situations.

3.5 Constant expressions

In C++, expressions are used almost everywhere. However, not every kind of expression can be used everywhere: one can hardly imagine the code in Figure 3.16 to be valid. The expressions used in an enum declaration or as a `case` label of a `switch` statement have to be computable in the compilation time. Those kind of expressions are generally known as *constant expressions*; they are defined in [10, §5.20].

```
enum class E { A = getchar() };
```

Figure 3.16: A „wild" expression in a context that requires a constant expression.

```
constexpr int const & id(int const &x) { return x; }
void test() {
        int const a = 7;
        //constexpr int const & b = id(a); // 1
        constexpr int c = id(a); // 2
        static int const d = 8;
        constexpr int const &e = id(d); // 3
}
```

Figure 3.17: ...

3.5.1 The intuition

The standard defines a couple terms with respect to constant expressions; this subsection attempts to give the reader an intuition which is behind them. The term *core constant expression* ([10, §5.20.2],) is used to represent an expression, which can be evaluated easily, using only the information the compiler has in the time of compilation. The term *constant expressions* ([10, §5.20:5],) denotes a core constant expression, whose value can exist and „makes sense" in the compile time.

The Figure 3.17 can serve as an example here. The function `id` takes one parameter, which is a reference to a constant integer, and returns it back. The expression `id(a)` is then a *core constant expression*, as it can be easily evaluated (its value is equivalent to the value of the expression `a`). It is not a *constant expression*, because its value refers to a non-static local variable, and the memory location of a (non-static) is unknowable in the compile time (such memory location does not even exist in the execution time, or the variable may be instantiated multiple times in recursive calls). Therefore, its value cannot be used as the initializer of a constexpr reference (1).

How is then possible that it may be used in the declaration (2)? Note that the variable is not initialized simply by the expression `id(a)`, but with the rvalue-to-lvalue conversion of the result of that `id(a)`;

such composed expression is a `constant expression`. In other words, the declaration of `c` does not concerned with the identity of the object it is initialized with, but only with its value - and that value is known in the compile time, as the variable `a` is initialized with the constant expression `7`. The declaration (3) works because the variable `d` was declared `static`.

```
int main() {
  const int x = 1;
  enum class E { A = x };
}
```

Figure 3.18: Sometimes, `const` is enough.

```
$ cat foo.cpp
int foo() { return 1; }
int main() {
    const int x = foo();
    enum class E { A = x };
}
$ clang++ -std=c++14 foo.cpp
error: enumerator value is not a constant expression
note: initializer of 'x' is not a constant expression
```

Figure 3.19: Sometimes, `const` is not enough.

4 Project overview

The project of C/C++ semantics in 𝕂 is hosted on GitHub[1]. It can be built easily by simply following the build instructions in the repository; however, the process deserves a few things to be mentioned here.

- The projects depends on RuntimeVerification's implementation of the 𝕂. The RV-𝕂 builds and runs without problems, with a minor exception: it requires the lexical parser flex[2] to be present in the system.
- The project uses clang as a library to parse C++ sources; the currently required version 3.9 is a bit outdated. It is not a problem, as clang can be built easily and the Project can be configured to use clang installed in any directory.
- The officially supported operating system is Ubuntu 16.4 LTS, which already contains prebuilt clang libraries in the required version. But the projects works on Fedora 26 without any problems.
- The building process can take up to thirty minutes on this text's author's machine.

4.1 Basic usage

Main user interface of the project consists of a script `kcc` [7], which implements a compiler based on the C/C++ semantics. The script mimics the interface of the GNU gcc compiler and supports many of gcc's command-line parameters. It is therefore possible to use it to build programs instead of gcc; however, the generated executables are many times slower than the ones built by gcc.

It has been said in Chapter 2 that in the 𝕂 framework, a semantics of a programming language L assigns to every program in L a set of program configuration with a transition system over them. The executable file `hello` generated by `kcc` (as shown in Figure 4.1) is

1. `https://github.com/kframework/c-semantics`
2. `https://github.com/westes/flex`

```
$ cat hello.C
extern "C" int puts(char const *s);
int main() {
        puts("Hello␣world");
}
$ kcc hello.C -o hello
$ ./hello
Hello world
```

Figure 4.1: A "hello world" program.

a perl script, which walks through the transition system in a step-by-step manner. The walk starts in an initial configuration and ends in a configuration for which no further transition is defined. The script then examines the final configuration and stops, possibly printing an error message whenever the walk ended abnormally.

It is possible to specify an exact number of computational steps to take by setting the variable `DEPTH` to the desired value. In this particular example, the executable is able to print only an incomplete portion of the text; then an error message is printed (Figure 4.2). The full list of accepted environment variables can be obtained by setting the environment variable `HELP`.

```
$ env DEPTH=675 ./hello
Hello woError: Execution failed.
```

Figure 4.2: Running a compiled program for an exact number of computational steps.

4.2 Under the hood

The Project internally consists of a clang-based tool `clang-kast` and many $\mathbb{K}$ modules. The modules are composed into three language definitions:

- the definition of C11 translation semantics, which we will refer to as *C translation semantics*;

- the definition of C++14 translation semantics, which we will refer to as *C++ translation semantics*, or simply *translation semantics*; and

- the definition of C11/C++14 execution semantics; which we will refer to as *execution semantics*.

Some modules are included in multiple language definitions. All the language definitions has to be compiled by `kompile`; the compilation of both `clang-kast` and the language definitions is driven by a Makefile.

All the 𝕂 source files are contained in the *semantics* directory in the project repository (Figure 4.3). The directory contains subdirectories `c11` and `cpp14`, which contain source files for the repsective languages; it also contains a subdirectory `common`, which contains (some) source files common to both languages. The directory of each language contains `language` and `library` subdirectories, where the former is split into subdirectories containing sources for translation semantics, execution semantics, and both. From this point further, the term *C++ semantics* will denote the content of `semantics/cpp14` subdirectory of the project.

```
$ tree -L 3 -d semantics
semantics
|-- common
|-- cpp14
|   |-- language
|   |   |-- common
|   |   |-- execution
|   |   |-- linking
|   |   \-- translation
|   \-- library
|-- c11
|   |-- language
|   |   |-- common
|   |   |-- execution
|   |   \-- translation
|   \-- library
\-- linking
```

Figure 4.3: The directory structure of the Project.

```
rule <k> sendString(FD::Int, S::String) => .K ...</k>
     <options> Opts::Set </options>
     requires lengthString(S) <=Int 0
     orBool (NoIO() in Opts)
```

When `kcc` is invoked on a C++ program, a clang-based tool `clang-kast` (from directory `cpp-parser`) is used to convert each source file into the 𝕂's internal representation (𝕂 AST). Every converted file is then individually used as a program, which is interpreted (using the 𝕂 tool `krun`) by the C++ *translation semantics*; the resulting terminal configuration can be thought of as an equivalent of an object file. The outputs are then joined together with runtime library and the result is wrapped in a generated Perl script. When the script is executed, it interprets (with `krun`) the linked program using the *C11/C++14 execution semantics*, possibly passing the script's command line arguments to the program.

4.2.1 Can we see it?

```
`<generatedTop>`(...
  `<thread>`(
    `<thread-id>`(#token("0","Int")),
    `<k>`(
      sendString(
        #token("1","Int"),
        #token("\"rld\\n\"","String")
      ) ~>
      sent(
        #token("1","Int"),
        #token("\"Hello world\\n\"","String")
      )...
    )...
  )...
)
```

Figure 4.4: An excerpt of a generated configuration. Large portions of the configuration were replaced by elipsis (...); the formatting (whitespaces) was added manually.

When an executable generated by `kcc` is run in an environment with the variable `VERBOSE` set, the executable prints the final configuration in a text form to the standard output. For the „hello world" program above (Figure 4.1), the konfiguration produced by the command

```
$ env VERBOSE=1 DEPTH=675 ./hello
```

has about 600 kilobytes. The excerpt in the Figure 4.4 contains a thread with two *computational items* on the top of its `K` cell. From that point, if the execution had not been stopped, the first computational item (`sendString`) would have sent the rest of the `Hello world` string to the standard output, then it would have been removed and the second computational item (`sent`) would have been processed.

The project contains some rules, which give semantics to the `sendString` term. The rule in the Figure 4.5 is one of them; it basically encodes the following piece of semantic information: „To send a nonempty string means to send its first character and then to send the rest, unless the IO is disabled". The Chapter 2 contains everything needed to understand this rule, except a small detail: the elipsis (...) in the `k` cell behaves exactly as `~> _` and expresses the notion of „we do not care what is in the rest of this cell". We provide an explanation of the rule only as a means for the reader to check his understanding. The rule says: „Every configuration, in which

1. there is an `options` cell containing a set not containing an `NoIO()` term, and in which
2. there is also a `k` cell having on its top a `sendString` item parametrized with an integer and a nonempty string `S`,

can be rewritten to another configuration by rewriting the `sendString` item to `#putc` of the first character, followed by (`~>`) the same `sendString` item, but without the first character of the string."

4.3 Sorts, syntax, semantics

The Project contains three types of modules: *sort* modules, which only define sorts; *syntax* modules, which define sort constructors and functions; and *semantic* modules, which contains mostly semantics rules

```
rule <k> sendString(FD::Int, S::String)
      => #putc(FD, ordChar(firstChar(S)))
      ~> sendString(FD, butFirstChar(S))
     ...</k>
     <options> Opts::Set </options>
     requires lengthString(S) >Int 0
     andBool notBool (NoIO() in Opts)
```

Figure 4.5: A semantics rule from file(...).

```
syntax PRVal  ::= prv(CPPValue, Trace, CPPTypeExpr)
syntax PRExpr ::= pre(Expr, Trace, CPPTypeExpr)
syntax PRVal  ::= PRExpr
```

Figure 4.6: Syntactic constructs for representation of *prvalue*s.

usually a few definitions of sorts and syntactic construct used only by that module. This structure reduces intermodule dependencies, because semantic modules need to depend only on syntactic modules, and syntactic modules only on sort modules.

4.4 Value categories

In order to represent expressions of different value categories, the C++ semantics define sorts `LVal`, `XVal` and `PRVal` for lvalues, xvalues and prvalues, respectively. Each one of those sorts have a *value* constructor, which represents an evaluated expression, and a subsort with an *expression* constructor, which represents an unevaluated expression. For example, *prvalue*s are represented by the sort `PRVal` as shown in the Figure 4.6. All value sorts are subsorted to the sort `Val`. The C++ semantics also defines a function `isEvalVal`, which rewrites to `true` for terms representing an evaluated expression.

4.5 KResults

The C++ semantics relies on strictness attributes and evaluation context, as described in Section 2.7. To be able to do that, the sort predicate `isKResult` needs to be able to distinguish the terms that should be the result of computations. This can be achieved by making the appropriate sorts to be subsorts of `KResult` and by tweaking `isKResult` with additional rewrite rules. For example, the semantics contains a rule as the one in Figure 4.7; the rule says that an evaluated expression (e.g. represented by a `prv` term) is a `KResult` if and only if it has a non-reference type.

```
rule isKResult(
        Lbl:KLabel(_::CPPValue, _::Trace, T::CPPType)
     )
     => notBool isCPPRefType(T)
     requires isValKLabel(#klabel(Lbl))
```

Figure 4.7: An evaluated expression may be a `KResult`.

The set of the computational results of the translation semantics has to be different then the set of the results of the execution semantics. For example, some expressions occurring in a body of a function cannot be fully evaluated when the function declaration is being processed, but they can be evaluated when the function is called. Therefore, some terms (e.g. a `pre` term) are results in the translation semantics, but they have to be evaluated further in the execution semantics. Because of that, the translation semantics contains a similar rule non-evaluated expressions (like `pre`). This is important for the implementation of generalized constant expressions (Section 5.4).

5 Implementation

This chapter focuses on those parts of the C/C++ semantics project, which are related to the goal and contribution of this thesis. The chapter describes the implementation of the main features, shows relations between the implementation, standard and general architecture of the semantics, and highlights some aspects of the C++ language one may perhaps oversee when using the language as a programmer. The last section of this chapter then gives a short evaluation of the implementation.

5.1 Our approach

When we began the work on this thesis, we had almost no knowledge of the 𝕂 framework and we knew the C++ language almost only from the position of a practicioner. Before we started extending the Project with any language features, we went through the official 𝕂 tutorial[1] to familiarize ourselves with the framework. We then implemented the language features approximately in the order in which they are described in this chapter. The implementation work on each of the features can be divided into four phases, which are described in the following paragraphs.

Understand the feature In the first phase we focused on understanding the selected language feature. During this phase we wrote many snippets of C++ code related to the feature; the snippets were based on our reading of the standard as well as on our own programming experience and on C++ related answers on StackOverflow. One useful tool in this phase was GodBold's Compiler Explorer[2]; which allowed us to easily compile a piece of code with different versions of various compilers.

Understand the relevant parts of the Project We then explored the Project to see which parts of the feature are already implemented

1. `http://www.kframework.org/index.php/K_Tutorial`
2. `https://gcc.godbolt.org/`

(if any), how the C++ semantics implements similar features and in which places the feature might be implemented. A simple `grep` tool was helpful in this place. We also the debugging switch of `kcc`, as well as its ability to run the compilation up to a specified depth.

Implement it The third phase consisted of extending the implementation with our code and test-cases, which was usually done in a *test-first* manner. Sometimes we reused some snippets from the first phase, but usually we created new tests. During this phase we often found some bugs in the Project; if that happened, we sometimes needed to fix it immediately, but usually we just created a different test-cases which would not trigger the bug and postponed the fix for later.

Do the rest When the feature was implemented, we fixed the bugs we found or removed the technical debt we encountered or created in the implementation phase. A few times we needed to reiterate the whole process, because we discovered a new aspect of the feature we did not fully understand yet.

5.2 Enumerations

When we were trying to understand enumerations, we created approximately 40 snippets of C++ source code. It then went out that in order to implement enumerations, several parts of the semantics had to be modified. The translation tool `clang-kast` was slightly modified to produce AST nodes for enumeration declarations; to process the declarations in the semantics, a new file was added to the C++ translation semantics and a new set of cells was added to common part of configuration. It was also needed to implement enumerator lookup, which required addition of a few cells to configuration and a slight modification of some of the existing name lookup rules. The semantics was to some extent already prepared to work with enumerations, and some of the relevant rules (e.g. for conversions) needed no change. To ensure correctness of implementation, several test cases were added to the test suite. Overall, most of the modifications were additive, and only little of the existing code needed to be changed. During the implementation process, a few minor bugs were discovered and fixed.

5.2.1 Declaration

Enumeration declaration, including the *opaque declaration*, is implemented in module `CPP-DECL-ENUM` of the semantics. Every enumeration declaration is processed as follows:

1. A new `cppenum` cell is created in the current translation unit. An error is reported if there already exist an enumeration with the same name, unless the declaratation is opaque.
2. The enumeration being declared is added to environment, so that it could be later looked up.
3. For full declarations, the enumerators are processed in the order of their declarations. Processed enumerators are stored in sub-cells of the `cppenum` cell; for unscoped enumerations, the enumerators are also added to the scope surrounding the enumeration declaration.
4. For unscoped enumerations without a fixed underlying type, the set of enumeration values and the underlying type is computed and stored in the `cppenum` cell.

For declarations of enumerations with no fixed underlying type, the standard keeps one aspect of the declaration unspecified. Prior the closing bracket of the enumerator declaration [3], the type of an enumerator with initializer is the type of the initializer, and type of an enumerator without initializer is the type of previous enumerator, whenever possible. If there is no initializer specified for the first enumerator, the type of the enumerator is unspecified; it is also unspecified for enumerators (without initializers) whose value does not fit into the type of previous enumerator. For example, in the declaration on figure 5.1, the type of enumerator `A` in the declaration of `B` is not specified, and therefore the value of `B` is not specified, too. Similarly, the value of enumerator `D` on most platforms does not fit to unsigned

3. And inside the enumeration declaration in particular.

char, which is the type of enumerator `C`, and its type is thus unspecified. Note that the types of enumerators are unspecified only inside of the declaration of the enumeration, i.e. prior the closing bracket of the declaration. Type of every enumerator after the complete declaration is always the type of the enumeration, so this unspecified behaviour is usually not a problem in practice, unless one writes code similar to the one in image 5.1.

```
enum E {
        A, B=sizeof(A), C=(unsigned char)255, D
};
```

Figure 5.1: Declaration of an enumeration with unspecified values of enumerators.

However, the semantics should be aware of this behaviour. Many real-world programs use enumerations whose enumerators does not have initializers, since the value of the enumerators is by default numbered from zero. Earlier versions of the semantics caused the semantic-based compiler kcc to stop the compilation whenever an undefined or unspecified behaviour was encoutered, which would be an unforunate thing to do for such programs. For this reason, the project maintainer added an error-reporting and recovery support to the semantics [4]. The current version of the semantics issues a warning, whenever this unspecified behaviour occur. Ideally, the warning would be suppressed if the unspecified type is never used, but this enhancement was not implemented.

5.2.2 Enumerator lookup

The name of an enumerator can be referred to using the scope resolution operator applied to a name of the enumeration. This was implemented easily using only a few rules in the (static) semantics. Furthemore, the enumerators of an unscoped enumeration are declared in the scope immediately containing the declaration of the enumeration. To implement this, we have decided to add a few new

4. `https://github.com/kframework/c-semantics/commit/584fa6ff4a90aca45de99d6b210177258ebd96d4`

cells, which map names of enumerators of enums defined in the surrounding scope to their corresponding type. The lookup then reuses the rules from the previous case.

The rules for enumerator lookup also have to consider the context in which the lookup is performed. As noted earlier, it is mandated by the standard that the types of declared enumerators are different inside the declaration then after it. One may find that surprising; however, this is needed in order to easily create enumerator initializers, which depends on values of previous enumerators of the same enumeration, as it is ilustrated in figure 5.2. If the type of the enumerator `A` in the initializer of the enumerator `B` was the type of the enumeration (as it is after the declaration), the initializer expression would be ill-formed, as in C++ enumerations are not implicitly convertible to arithmetic types. Thus, an enumerator prior the closing bracket has always an integral type.

```
enum F {
        A=1, B=A+2
};
```

Figure 5.2: An enumerator depends on value of previous enumerator.

5.2.3 Underlying type

The C++ *language* does not have a direct support for determining the underlying type of an enumeration. Instead, such facility is provided by the standard library, as shown in Figure 3.4; the standard library then uses compiler's intrinsic functions to get the underlying type. For example, the GNU gcc compiler provides a function[5] `__underlying_type`, which translates to the underlying type of its argument (Figure 5.3).

The clang fronted supports this intrinsic, too. Our tool `clang-kast` translates this intrinsic to the constructor `GnuEnumUnderlyingType` of sort `AType`; the translation semantics then contain a few straightforward rules which find the declared enumeration and its underlying type.

5. It is not a function in the C/C++ sense, as it operates on types.

```
enum class E : short {};
__underlying_type(E) x = 5;
```

Figure 5.3: Use of gcc's intrinsic function for determining the underlying type of an enumeration.

5.2.4 State of the implementation

We successfully implemented enumeration declaration and enumerator lookup for both scoped and unscoped enumerations. Our implementation allows enumerations to be declared in all scopes where enumerations can be declared according to the standard, that is, in a namespace scope, in a class scope or in a block scope (inside a function). We also implemented opaque enumeration declaration and redeclaration and a support for the standard library `std::underlying_type_t` template. During the implementation we fixed a number of bugs in the Project, including a bug which caused the lookup for a double nested name to fail. We also found a few defects in other projects (see Section 6.1).

We have not implemented the non-standard `__attribute__((packed))` extension; however, it is likely that it will be implemented in the future, because the C semantics supports it. We also do not have precise semantics for the enumerator initializers, which should be different kinds of constant expression depending on the fixedness of the enumeration's underlying type (see also Section 5.4). Most of the implementation was merged into upstream in April 2017, but a few features are still implemented only in our (publicly-available) fork of the project.

5.3 Zero initialization of class types

When we were choosing the features to implement as a part of this thesis, we did not intend to implement anything class-related at all. Later, when most of the enumerations-related language features was implemented, the maintainer of the Project suggested[6] that we could

6. https://github.com/kframework/c-semantics/pull/275#issuecomment-294241691

```
rule #figureInit(/*...*/, ExpressionList(.List),/*...*/)
     => #valueInit(/*...*/)

rule #figureInit(/*...*/, ExpressionList(L::List),/*...*/)
     => constructorInit(/*...*/)
```

Figure 5.4: The rules causing the nondeterminism

implement zero-initialization on class types (see Section 3.4). Since we wanted to gain deeper understanding of both the C++ language and the Project before attempting to tackle the problem of generalized constant expressions, and the implementation part of the zero-initialization of class-types seemed to be relatively straightforward, we agreed. However, the implementation of this feature turned out to be far more challenging than we thought.

5.3.1 A bug: nondeterminism

The first problem we encountered was quite surprising. It took us many hours, perhaps a few tens of hours, to identify it; however, the bug was quite easy to fix. The Project contained two rules as in Figure 5.4, which had to differentiate between two cases in list-initialization: when the initializer list is empty, the object should be value-initialized (as in [10, §8.5.4/3.4], the first rule), and when it is non-empty, the object should be constructed with a constructor (as in [10, §8.5.4/3.6], the second rule). However, the second rule could apply even when the initializer list was empty, which was the bug.

The bug was hard to identify, because if both rules could apply, the 𝕂 framework usually applied the first. However, when we began the implementation of the zero initialization on class types, the nondeterminism manifested. When the newly-written zero-initialization code contained a bug which should cause the compilation to get stuck, then the second rule applied, thus bypassing our newly-written code.

If we had had a deeper understanding of how non-determinism works in 𝕂, and had we considered the possibility of existence of nondeterministic behavior in the translation semantics, we would probably found the bug sooner. Once we knew about this non-determinism, it was easy to fix it[7].

5.3.2 Another bug: no default initialization

The second problem caused some classes to not be default-initialized. The semantics contained a rule responsible for zero-initializing an object of a class type with subsequent default-initialization; the rule aimed to skip the default-initialization phase for classes with trivial default-constructor, but the phase was skipped in the other case instead. This caused an assertion failure for code in Figure 5.5. This bug was also easy to fix[8].

```
struct Def {
  int x;
  int y = 7;
};

int main() {
  Def def{};
  assert(def.y == 7);
}
```

Figure 5.5: Non-trivial implicitly-defaulted default constructor.

5.3.3 Implementation

To zero-initialize an object of a class type means to zero-initialize all its non-static members, its base classes and also all the padding among the subobjects. The initialization of non-static members was straightforward, because we could reuse an already existing syntax for member access and create computational items like

`zeroInit(Base . no-template Name(C, X), T)`

7. https://github.com/kframework/c-semantics/pull/287
8. Fixed in PR https://github.com/kframework/c-semantics/pull/291. Note that the accompanying test-case tests default-initialization without preceding zero-initialization, which is not exactly what the patch fixes. This is probably due to the fact that in the time of fixing this, zero initialization of classes was not implemented yet, and therefore the correct test-case would not execute properly.

```
struct Base1 { int x; };
struct Base2 { int y; };
struct Derived : Base1, Base2 {};
int foo(Derived const &d) {
  return d.x + d.y;
}
```

Figure 5.6: Accessing a non-static member variable of a base class.

which zero-initialize a (non-templated) member variable `X` (of type T) of an object `Base` of a class-type `C`. This syntax with dot member access is similar to the syntax of C++. In order to implement the zero-initialization of base-classes, we needed to create a means of obtaining a base-class subobject; in order to test the implementation, we needed to fix the semantics of accessing a non-static member variable of a base class (see Figure 5.6). The fix [9] was reviewed by the project maintainer and is likely to be merged soon.

In order to implement zero-initialization of class padding, we needed to extend the code for class declaration. For every non-static data member it was necessary to store not only its offset in the class, but also an *unaligned offset* - the offset the member would have if its type had no requirements on alignment. Zero-initialization of the padding is then implemented as zero-initialization of an array of characters.

5.3.4 State of the implementation

We successfully implemented zero initialization of classes, and during the implementation work we fixed a few bugs in the Project. The implementation is ready to be merged into upstream and as far as we know, it is complete. The only limitation known to us is that zero-initialization of bit-fields and virtual base classes is not tested, because those language features are still not supported in the C++ semantics.

9. `https://github.com/kframework/c-semantics/pull/292`

5.4 Generalized constant expressions (constexpr)

In order to support generalized constant expressions in the Project, three things needs to be done. First, the translation semantics has to be modified to allow evaluation of complex expressions, and function calls in particular. Second, the translation semantics needs to capture the precise meaning of constant expressions as defined in the standard. Third, the semantics for constant expressions has to be used in appropriate places, e.g. to evaluate enumerator initializers and case labels. In this thesis, we focus on the first thing.

Before we started the work, the translation semantics was already able to evaluate simple arithmetic or boolean C++ expressions (like `1 + 1 == 2`. However, the translation semantics was not able to evaluate function calls. To evaluate them, the configurations of execution semantics contains a set of cells which represents *call stack*, but the configuration of translation semantics does not contain anything like that.

When designing the support for generalized constant expressions, several design alternatives seemed to be viable. One of them was to extend the translation semantics with new rules and configuration cell which would allow to evaluate the not-supported-yet expressions; another alternative was to reuse some of the rules from the execution semantics and modify the configuration accordingly. We decided to first try to reuse as much of existing code as possible, with the perspective that if that fails, we will fall back to other design alternative or try to explore the design space more carefully. It turned out that we could reuse almost everything; although the changes were quite extensive, there were only a few newly added rules and cells, and most of them had the character of „move this there", „use this file from the execution semantics in the translation semantics" and „add a `require` clause here".

5.4.1 Configuration

Although the source code files for C and C++ execution semantics are located in different directories, they are compiled together as a part of one language definition. The main configuration structure for the execution semantics is defined in the module `C-CONFIGURATION`[10]; this module imports another modules, some of which (for example, the module `COMMON-CONFIGURATION`[11]) contains definitions of other parts of the configuration.

Before we started the work on generalized constant expressions, the module `C-CONFIGURATION` contained among other things a definition of configuration cell `thread-local`, which held most of the resources needed to run a single thread of execution - circa 100 lines of code. We decided to extract the definition into a newly created module `COMMON-THREAD-LOCAL`[12] a reuse it in the C++ translation semantics.

The problem with the reuse was that the cell `thread-local` contained some cells, whose names collided with names of cells already existing in the translation semantics. The collision was easily resolved by renaming the cells from the translation semantics (renaming the cells included in the `thread-local` cell would affect large parts of the C semantics).

5.4.2 Collision of semantic rules

We also wanted to reuse as many semantics rules as possible from the execution semantics. A problem with this was that some syntactic constructs had a different meaning in the translation semantics than in the execution semantics. The constructor `IfStmt` (Figure 5.7) can serve as an example here. In the translation semantics, an if-statement could be processed only as a part of a function declaration; the translation semantics had to perform a contextual conversion on the expression and keep the rest of the work for the execution semantics. In the execution semantics, the statement could be processed only when executing a function, so it was needed to evaluate the condition first and then to rewrite the term to either the first statement (if-block) or the second (else-block).

However, if we tried to execute a function with an `IfStmt` during the translation semantics, the translation rules could still apply, which would be a problem. We solved this issue by two means. First, we intro-

10. in file `semantics/c11/language/execution/configuration.k`
11. in file `semantics/common/configuration.k`
12. in file `semantics/common/thread_local.k`

```
syntax Stmt ::= IfStmt(Expr, Stmt, Stmt)
```

Figure 5.7: Syntactic construct for the C++ `if` statement.

duced a nullary[13] boolean function `Execution`, which returns `true` in the execution semantics or when executing an expression during the translation semantics, and `false` when doing „real" translation. Second, for many syntactic constructs (including `IfStmt` and `WhileStmt`) we introduced a new version of the construct only for the purpose of the translation semantics (e.g. `IfTStmt`, `WhileTStmt`) and changed the translation rules to work with these.

5.4.3 KResults

Another problem was that the meaning of the `strict` attribute in the translation semantics was different than in the execution semantics, because the `KResults` were different (see Section 4.5). When we tried to evaluate an expression during a function call in the translation semantics, some cooling rules applied too early and the computation got stuck. We fixed this by modifying one rule for the sort predicate `isKResult` to return `false` for terms representing not fully evaluated expressions (e.g. `pre`, `le`), when the function `Execution` evaluates to `true`.

5.4.4 State of the implementation

In the time of writing, the implementation of generalized constant expression is only experimental and not ready to be merged into the upstream. We still do not support some language constructs, the implementation contains a few bugs and does not fully adhere to the standard, and the semantics for undefined behavior developed in [7] is not supported during the compile-time execution. That being said, we also have to mention that the whole project of C++ semantics in $\mathbb{K}$ is still very experimental, contains bugs and do not support many language construct (including, for example, the `switch` statement).

13. It may still examine the current configuration using reflection.

```
constexpr int fact(int n) {
        int f = 1;
        while (n > 0) {
                f = f * n;
                n = n - 1;
        }
        return f;
}
```

Figure 5.8: A `constexpr` factorial.

```
constexpr int fact(int n) {
        return n <= 1 ? 1 : n * fact(n - 1);
}
int main() {
        constexpr int f_4 = fact(4);
        enum class Facts {
                F_4 = f_4,
                F_5 = fact(5)
        };
        assert(int(Facts::F_4) == 24);
        assert(int(Facts::F_5) == 120);
}
```

Figure 5.9: A *recursive* `constexpr` factorial with tests.

Despite the fact that the implementation of generalized constant expressions is still incomplete, we do consider the current state of the implementation to be a success, for at least two reasons. First, the `kcc` is now able to compute values of functions in the compilation time (i.e. in the translation semantics), including functions containing control-flow statements (Figure 5.8) or recursive calls (Figure 5.9). Second, in doing so, we were able to avoid code duplication and reuse large parts of the existing execution semantics. Therefore we have a reason to believe that the approach we chose was good and that it will be possible to complete the implementation in near future.

5.5 Summary

Enumerations are fully implemented and most of the implementation is already merged in the upstream repository. Zero initialization of class types is also fully implemented, modulo some features which are not yet supported in the project at all; it is ready to be merged into the upstream. We have an elegant implementation of generalized constant expressions which is able to evaluate functions with recursion or non-trivial control-flow. The implementation is incomplete, not adhering to the standard and not production ready; however, it is based on good principles and can serve as a basis for future development.

6 Relation to other projects

6.1 Defects found in other projects

During the work, we discovered a few defects in major C++ compilers and also in the C++ standard. While most of them was already known among the experts, some of them are new; such defects were reported. In this section, we give a brief overview of the defects we found, including the ones which were already known but not fixed yet.

6.1.1 A gcc redeclaration bug

We discovered a new bug in the C++ parser of the gcc compiler. When an opaque enumeration is redeclared using fully qualified name (as in Figure 6.1), the compiler rejects it, although such code is valid (and the clang compiler compiles it without problems). We reported the bug to the official bugzilla[1]; the bug was then confirmed by a gcc c++ developer, who had subsequently shown that the bug works not only for enumerations, but also for classes. The bug affects many versions of the gcc compiler, including the version 7.2, which was the latest in the time of writing.

```
enum E : int;
enum ::E : int{};
```

Figure 6.1: A minimal example of a redeclaration error

6.1.2 A Clang inconsistency bug

The Figure 6.2 shows a forward-declaration of an enumeration `E` in the namespace `N`, followed by a redeclaration of the enumeration; the redeclaration itself is contained in the scope of the global namespace. In which scope are the enumerators `A`, `B` declared? The standard says:

1. `https://gcc.gnu.org/bugzilla/show_bug.cgi?id=83132`

```
namespace N { enum E : int; }
enum N::E : int {A,B};
namespace N {
    int x = int(::N::B); // 1
}
int y = int(::A); // 2
int z = int(A); // 3
```

Figure 6.2: Clang inconsistency bug.

> Each enum-name and each unscoped enumerator is declared in the scope that immediately contains the enum-specifier. [10, §7.2:11],

Therefore, from the standard's point of view, the declaration (1) is ill-formed, while the declarations (2) and (3) are valid. However, the compilers we have tried (the GNU gcc compiler in version 7.2, the MSVC compiler in version 19, and Intel's ICC in version 18) all declare the enumerators in the scope, in which the enumeration is declared, and they therefore accept the declaration (1) and reject (2) and (3).

Clang takes the same approach as other compilers and accepts (1) and rejects (2), but due a bug, it also accepts (3). We reported this behavior to the Clang's Bugzilla as a bug[2]. That was confirmed by the owner of the Clang C++ frontend, Richard Smith, who also found a different manifestation of the bug.

The problems with opaque declarations of unscoped enumerations are known at least from 2012, when (the same) Richard Smith submitted a related issue[3] to the list of core language issues. The proposed resolution is to disallow the opaque declarations of unscoped enumerations; see also the discussion in the ISO C++ Google Group[4].

6.1.3 Type of an enumeration

We found a contradiction in the C++14 standard; the contradiction is still present in C++17 as well as in an early C++20 draft (N4700). The

2. https://bugs.llvm.org/show_bug.cgi?id=35401
3. http://www.open-std.org/jtc1/sc22/wg21/docs/cwg_active.html#1485
4. https://groups.google.com/a/isocpp.org/d/embed/msg/std-discussion/_W1pUFzl3og/0Kpb0li2YMwJ

contradiction is also not mentioned in the current (97th) revision of "C++ Standard Core Language Active Issues", so we think that it is a new defect.

```
enum class E {
    A,
    B = E::A // 1
};
```

Figure 6.3: What is the type of `E::A` in (1)?

The contradiction can be illustrated on the code in Figure 6.3. The problem is that the standard gives two different answers to the question: „What is the type of the expression `E::A` in the declaration of `B`?". An answer can be based on [10, §7.2/5], , which says:

> [...] Following the closing brace of an enum-specifier, each enumerator has the type of its enumeration. If the underlying type is fixed, **the type of each enumerator prior to the closing brace is the underlying type** and the constant-expression in the enumerator-definition shall be a converted constant expression of the underlying type [...]

Therefore, the type of `E::A` should be `int`, which is the default underlying type of scoped enumerations. However, the standard also says ([10, §5.1.1/11],):

> A nested-name-specifier that denotes an enumeration (7.2), followed by the name of an enumerator of that enumeration, is a qualified-id that refers to the enumerator. The result is the enumerator. **The type of the result is the type of the enumeration.** The result is a prvalue.

So the type of `E::A` should be `E`. This is clearly a defect in the standard. The only way to satisfy both requirements is to make the types `E` and `int` the same, which would directly contradict the standard, because „Each enumeration defines a type that is different from all other types."[10, §7.2/5], .

One possible solution to this conflict is to change [10, §5.1.1/11], to say that the type of the result is the type of the **enumerator**, or to remove the relevant sentence completely. That is also how compilers seem to implement it: if `E::A` had the type of an enumeration, such enumerator declaration would be ill-formed ([10, §5.1.1/5],). Our language definition also takes this approach.

This issue was first discussed on StackOverflow[5], than it was raised in the ISO C++ Google Group[6]. Later we also reported this (incorrectly) as an editorial issue of the standard[7] and it seems that someone is already working on that.

5. https://stackoverflow.com/q/42369314/6209703
6. https://isocpp.org/forums/iso-c-standard-discussion?place=msg%2Fstd-discussion%2Fj63Em3eUa5c%2FDwdIEjBJAgAJ
7. https://github.com/cplusplus/draft/issues/1845

7 Conclusion

7.1 Future work

- Merge the remaining code into upstream.
- Finish the implementation of constant expressions. Working with literal types.
- Finish the rest of the C++ semantics (switch, templates, sfinae, lambda functions, automatic type deduction, type traits, virtual inheritance, multiple inheritance, variable templates, variadic templates ...), so that `kcc` can compile all the programs which can be compiled with gcc/clang.
- Somehow deal with C++ versions and defects. C++17? C++20?
- What about the concurrency?
- Design and implement some ACSL-like assertion language.
- Function calls in contracts? Can we check them?
- High-level reasoning about templates? And what about Concepts?

Bibliography

[1] G. Rosu, "K - a semantic framework for programming languages and formal analysis tools," in *Dependable Software Systems Engineering*, ser. NATO Science for Peace and Security, D. Peled and A. Pretschner, Eds. IOS Press, 2017.

[2] G. Roşu, "From rewriting logic, to programming language semantics, to program verification," in *Logic, Rewriting, and Concurrency: Essays Dedicated to José Meseguer*, ser. LNCS, vol. 9200. Springer, September 2015, pp. 598–616.

[3] D. Bogdănaş and G. Roşu, "K-Java: A Complete Semantics of Java," in *Proceedings of the 42nd Symposium on Principles of Programming Languages (POPL'15)*. ACM, January 2015, pp. 445–456.

[4] D. Park, A. Ştefănescu, and G. Roşu, "KJS: A complete formal semantics of JavaScript," in *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'15)*. ACM, June 2015, pp. 346–356.

[5] A. Tokarčík, "An executable formal semantics of agda," Masters Thesis, Masaryk University, Faculty of Informatics, Brno, 2015. [Online]. Available: http://theses.cz/id/y9dksc/

[6] C. Ellison, "A formal semantics of C with applications," Ph.D. dissertation, University of Illinois, July 2012.

[7] C. Hathhorn, C. Ellison, and G. Roşu, "Defining the undefinedness of c," in *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'15)*. ACM, June 2015, pp. 336–345.

[8] D. Guth, C. Hathhorn, M. Saxena, and G. Rosu, "Rv-match: Practical semantics-based program analysis," in *Computer Aided Verification - 28th International Conference, CAV 2016, Toronto, ON, Canada, July 17-23, 2016, Proceedings, Part I*, ser. LNCS, vol. 9779. Springer, July 2016, pp. 447–453.

[9] "ISO International Standard ISO/IEC 14882:2011 - Programming Language C++." International Organization for Standardization, Geneva, CH, Standard (withdrawn), Sep. 2011.

[10] "Standard for Programming Language C++ - Working Draft, N4296," November 2014. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2014/n4296.pdf

[11] "ISO International Standard ISO/IEC 14882:2017 - Programming Language C++." International Organization for Standardization, Geneva, CH, Standard (upcoming), Dec. 2017.

[12] Bjarne Stroustrup. The c++ programming language. [Online]. Available: http://www.stroustrup.com/C++.html

[13] Bjarne Stroustrup. Bjarne stroustrup's faq. [Online]. Available: http://www.stroustrup.com/bs_faq.html

[14] "ISO International Standard ISO/IEC 14882:1998 - Programming Language C++." International Organization for Standardization, Geneva, CH, Standard (withdrawn), Sep. 1998.

[15] Jtc1/sc22/wg21 - the c++ standards committee homepage. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg21/

[16] C++ standards committee papers. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg21/docs/papers/

[17] "ISO International Standard ISO/IEC 14882:2014 - Programming Language C++." International Organization for Standardization, Geneva, CH, Standard, Dec. 2014.

[18] "Standard for Programming Language C++ - Working Draft, N4700," October 2017. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2014/n4700.pdf

[19] A. A. Stepanov and P. McJones, *Elements of Programming*. Addison-Wesley Professional, 2009.

[20] T. Univalent Foundations Program, *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study: https://homotopytypetheory.org/book, 2013.