BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

JUST-IN-TIME COMPILATION OF THE DEPENDENTLY-TYPED LAMBDA CALCULUS

JUST-IN-TIME PŘEKLAD ZÁVISLE TYPOVANÉHO LAMBDA KALKULU

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. JAKUB ZÁRYBNICKÝ**
AUTOR PRÁCE

SUPERVISOR **Ing. ONDŘEJ LENGÁL, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2021

Brno University of Technology
Faculty of Information Technology

Department of Intelligent Systems (DITS) Academic year 2020/2021

Master's Thesis Specification

Student: **Zárybnický Jakub, Bc.**
Programme: Information Technology
Field of study: Intelligent Systems
Title: **Just-in-Time Compilation of Dependently-Typed Lambda Calculus**
Category: Compiler Construction
Assignment:

1. Investigate dependent types, simply-typed and dependently-typed lambda calculus, and their evaluation models (push/enter, eval/apply).
2. Get familiar with the Graal virtual machine and the Truffle language implementation framework.
3. Create a parser and an interpreter for a selected language based on dependently-typed lambda calculus.
4. Propose a method of normalization-by-evaluation for dependent types and implement it for the selected language.
5. Create a just-in-time (JIT) compiler for the language using the Truffle API.
6. Compare the runtime characteristics of the interpreter and the JIT compiler, evaluate the results.

Recommended literature:

- https://www.graalvm.org/
- Löh, Andres, Conor McBride, and Wouter Swierstra. "A tutorial implementation of a dependently typed lambda calculus." Fundamenta Informaticae 21 (2001): 1001-1031.
- Marlow, Simon, and Simon Peyton Jones. "Making a fast curry: push/enter vs. eval/apply for higher-order languages." Journal of Functional Programming 16.4-5 (2006): 415-449.

Requirements for the semestral defence:

- Items 1 to 3.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor: **Lengál Ondřej, Ing., Ph.D.**
Head of Department: Hanáček Petr, doc. Dr. Ing.
Beginning of work: November 1, 2020
Submission deadline: May 19, 2021
Approval date: November 11, 2020

Abstract

A number of programming languages have managed to greatly improve their performance by replacing their custom runtime system with Just-in-time (JIT) optimizing compilers like GraalVM or RPython. This thesis evaluates whether such a transition would also benefit dependently-typed programming languages by implementing a minimal language based on the $\lambda\star$-calculus using the Truffle language implementation framework on the GraalVM platform, a partial evaluation-based JIT compiler based on the Java Virtual Machine.

This thesis introduces the type theoretic notion of dependent types, specifies a minimal dependently-typed language, and implements two interpreters for this language: a simple AST-based interpreter, and a Truffle-based interpreter. A number of optimization techniques that use the capabilities of a JIT compiler are then applied to the Truffle-based interpreter. The performance of these interpreters is then evaluated on a number of normalization and elaboration tasks designed to be comparable with other system, and the performnce is then compared with a number of state-of-the-art dependent languages and proof assistants.

[...specific numbers]

Abstrakt

[...česky]

Keywords

Truffle, Java Virtual Machine, just-in-time compilation, compiler construction, dependent types, lambda kalkul

Klíčová slova

Truffle, Virtuální stroj JVM, just-in-time kompilace, tvorba překladačů, závislé typy, lambda kalkulus

Reference

ZÁRYBNICKÝ, Jakub. *Just-in-Time Compilation of the Dependently-Typed Lambda Calculus*. Brno, 2021. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Ondřej Lengál, Ph.D.

Rozšířený abstrakt

[...česky]

Just-in-Time Compilation of the Dependently-Typed Lambda Calculus

Declaration

I hereby declare that this Master's thesis was created as an original work by the author under the supervision of Ing. Ondřej Lengál Ph.D.

I have listed all the literary sources, publications, and other sources that were used during the preparation of this thesis.

.....
Jakub Zárybnický
May 10, 2021

Acknowledgements

I would like to thank my supervisor Ing. Ondřej Lengál Ph.D. for entertaining a second one of my crazy thesis proposals.

I would like to thank Edward Kmett for suggesting this idea: it has been a challenge.

I would also like to thank my family and friends who supported me when I needed it the most throughout my studies.

Contents

Chapter 1

Introduction

Proof assistants like Coq, F*, Agda or Idris, or other languages with dependent types like Cayenne or Epigram, allow programmers to write provably correct-by-construction code in a manner similar to a dialog with the compiler [**?**]. They also face serious performance issues when applied to problems or systems on a large-enough scale [**?**] [**?**]. Their performance grows exponentially with the number of lines of code in the worst case [**?**], which is a significant barrier to their use. While many of the performance issues are fundamentally algorithmic, a better runtime system would improve the rest. However, custom runtime systems or more capable optimizing compilers are time-consuming to build and maintain. This thesis seeks to answer the question of whether just-in-time compilation can help to improve the performance of such systems.

Moving from custom runtime systems to general language platforms like e.g., the Java Virtual Machine (JVM) or RPython [**?**], has improved the performance of several dynamic languages: project like TruffleRuby, FastR, or PyPy. It has allowed these languages to reuse the optimization machinery provided by these platforms, improve their performance, and simplify their runtime systems.

As there are no standard benchmarks for dependently typed languages, we design a small, dependently-typed core language to see if using specific just-in-time (JIT) compilation techniques produces asymptotic runtime improvements in the performance of β-normalization and $\beta\eta$-conversion checking, which are among the main computational tasks in the elaboration process and is also the part that can most likely benefit from JIT compilation. The explicit non-goals of this thesis are language completeness and interoperability, as neither are required to evaluate runtime performance.

State-of-the-art proof assistants like Coq, Agda, Idris, or others is what we can compare our results with. There are also numerous actively developed research projects in this area; Lean is a notable one that I found too late in my thesis to incorporate its ideas. However, the primary evaluation will be against the most well established proof assistants. Reformulat

As for the languages that use Truffle, the language implementation framework that allows interpreters to use the JIT optimization capabilities of GraalVM, an alternative implementation of the Java Virtual Machine: there are numerous general-purpose functional languages, the most prominent of which are TruffleRuby and FastR. Both were reimplemented on the

Truffle platform, resulting in significant performance improvements[1, 2]. We will investigate the optimization techniques they used, and reuse those that are applicable to our language.

There is also a number of functional languages on the Java Virtual Platform that do not use the Truffle platform, like Clojure, Scala or Kotlin, as well as purely functional languages like Eta or Frege. All of these languages compile directly to JVM byte code: we may compare our performance against their implementation, but we would not be able to use their optimization techniques. To the best of my knowledge, neither meta-tracing nor partial evaluation have been applied to the dependently-typed lambda calculus.

The closest project to this one is Cadenza [**?**], which served as the main inspiration for this thesis. Cadenza is an implementation of the simply-typed lambda calculus on the Truffle framework. While it is unfinished and did not show as promising performance compared to other simply-typed lambda calculus implementations as its author hoped, this project applies similar ideas to the dependently-typed lambda calculus, where the presence of type-level computation should lead to larger gains. Rephrase

In this thesis, I will use the Truffle framework to evaluate how well are the optimizations provided by the just-in-time compiler GraalVM suitable to the domain of dependently-typed languages. GraalVM helps to turn slow interpreter code into efficient machine code by means of *partial evaluation* [**?**]. During partial evaluation, specifically the second Futamura projection [**?**], an interpreter is specialized together with the source code of a program, yielding executable code. Parts of the interpreter could be specialized, some optimized, and some could be left off entirely. Depending on the quality of the specializer, this may result in performance gains of several orders of magnitude.

Truffle makes this available to language creators, they only need to create an interpreter for their language. It also allows such interpreters to take advantage of GraalVM's *polyglot* capabilities, and directly interoperate with other JVM-based languages, their code and values [**?**]. Development tooling can also be derived for Truffle languages quite easily [**?**]. Regardless of whether Truffle can improve their performance, both of these features would benefit dependently-typed or experimental languages.

While this project was originally intended just as a λΠ calculus compiler and an efficient runtime, it has ended up much larger due to a badly specified assignment. I also needed to study type theory and type checking and elaboration algorithms that I have used in this thesis, and which form a large part of chapters 2 and 3.

Starting from basic λ-calculus theory and building up to the systems of the lambda cube, we specify the syntax and semantics of a small language that I refer to as Montuno (Chapter~2). We go through the principles of λ-calculus evaluation, type checking and elaboration, implement an interpreter for Montuno in a functional style (Chapter 3). In the second part of the thesis, we evaluate the capabilities offered by Truffle and the peculiarities of Truffle languages (Chapter 4), implement an interpreter for Montuno on the Truffle framework (Chapter 5), and apply various JIT optimizations to it (Chapter 6). After designing and using a set of benchmarks to evaluate the language's performance, we close with a large list of possible follow-up work (Chapter 7).

[1] Unfortunately, there are no officially published benchmarks, but a number of articles claim that TruffleRuby is 10-30x faster than the official C implementation. [**?**]

[2] FastR is between 50 to 85x faster than GNU R, depending on the source. [**?**]

Chapter 2

Language specification: λ⋆-calculus with extensions

2.1 Introduction

Proof assistants like Agda or Idris are built around a fundamental principle called the Curry-Howard correspondence that connects type theory and mathematical logic, demonstrated in Figure 2.1. In simplified terms it says that given a language with a self-consistent type system, writing a well-typed program is equivalent to proving its correctness [**?**]. It is often shown on the correspondence between natural deduction and the simply-typed λ-calculus, as in Figure 2.2. Proof assistants often have a small core language around which they are built: e.g. Coq is built around the Calculus of Inductive Constructions, which is a higher-order typed λ-calculus.

Mathematical logic	Type theory
$\top$ true	$()$ unit type
$\bot$ false	$\emptyset$ empty type
$p \wedge q$ conjunction	$a \times b$ sum type
$p \vee q$ disjunction	$a + b$ product type
$p \Rightarrow q$ implication	$a \to b$ exponential (function) type
$\forall x \in A, p$ universal quantification	$\Pi_{x:A} B(x)$ dependent product type
$\exists x \in A, p$ existential quantification	$\Sigma_{x:A} B(x)$ dependent sum type

Figure 2.1: Curry-Howard correspondence between mathematical logic and type theory

Compared to the type systems in languages like Java, dependent type systems can encode much more information in types. We can see the usual example of a list with a known

Natural deduction	$\lambda\rightarrow$ calculus
$\dfrac{}{\Gamma_1, \alpha, \Gamma_2 \vdash \alpha}$ axiom	$\dfrac{}{\Gamma_1, x : \alpha, \Gamma_2 \vdash x : \alpha}$ variable
$\dfrac{\Gamma, \alpha \vdash \beta}{\Gamma \vdash \alpha \rightarrow \beta}$ implication introduction	$\dfrac{\Gamma, x : \alpha \vdash t : \beta}{\Gamma \vdash \lambda x.t : \alpha \rightarrow \beta}$ abstraction
$\dfrac{\Gamma \vdash \alpha \rightarrow \beta \qquad \Gamma \vdash \alpha}{\Gamma \vdash \beta}$ modus ponens	$\dfrac{\Gamma \vdash t : \alpha \rightarrow \beta \qquad \Gamma \vdash u : \alpha}{\Gamma \vdash tu : \beta}$ application

Figure 2.2: Curry-Howard correspondence between natural deduction and $\lambda\rightarrow$-calculus

length in Listing 2.1: the type `Vect` has two parameters, one is the length of the list (a Peano number), the other is the type of its elements. Using such a type we can define safe indexing operators like `head`, which is only applicable to non-empty lists, or `index`, where the index must be given as a finite number between zero and the length of the list (`Fin len`). List concatenation uses arithmetic on the type level, and it is possible to explicitly prove that concatenation preserves list length.

```
data Vect : (len : Nat) -> (elem : Type) -> Type where
  Nil  : Vect Z elem
  (::) : (x : elem) -> (xs : Vect len elem) -> Vect (S len) elem

-- Definitions elided
head : Vect (S len) elem -> elem
index : Fin len -> Vect len elem -> elem
(++) : (xs : Vect m elem) -> (ys : Vect n elem) -> Vect (m + n) elem

proofConcatLength
  : {m, n : Nat} -> {A : Type} -> (xs : Vect n A) -> (ys : Vect m A)
    -> length (xs ++ ys) = length xs + length ys
```

Listing 2.1: Vectors with explicit length in the type, source: the Idris base library

On the other hand, these languages are often restricted in some ways. General Turing-complete languages allow non-terminating programs: non-termination leads to a inconsistent type system, so proof assistants use various ways of keeping the logic sound and consistent. Idris, for example, requires that functions are total and finite. It uses a termination checker, checking that recursive functions use only structural or primitive recursion, in order to ensure that type-checking stays decidable.

This chapter aims to introduce the concepts required to specify the syntax and semantics of a small dependently-typed language and use these to produce such a specification, a necessary prerequisite so that we can create interpreters for this language in later chapters. This chapter, however, does not attempt to be a complete reference in the large field of type theory.

2.2 Languages

2.2.1 λ-calculus

We will start from the untyped lambda calculus, as it is the language that all following ones will build upon. Introduced in the 1930s by Alonzo Church as a model of computation, it is a very simple language that consists of only three constructions: abstraction, application, and variables, written as in Figure 2.3.

e	::=	v	variable
	|	$M\ N$	application
	|	$\lambda v.\ M$	abstraction

(a) Standard (Church) notation

e	::=	v
	|	$(N)\ M$
	|	$[v]\ M$

(b) De Bruijn notation

Figure 2.3: λ-calculus written in Church and de Bruijn notation

β-reduction The λ-abstraction $\lambda x.\ t$ represents a program that, when applied to the expression x, returns the term t. For example, the expression $(\lambda x.x\ x)\ t$ produces the expression $t\ t$. This step, applying a λ-abstraction to a term, is called *β-reduction*, and it is the basic *rewrite rule* of λ-calculus. Another way of saying that is that the x is assigned/replaced with the expression T, and it is written as the substitution $M[x := T]$

$$(\lambda x.\ t)\ u \longrightarrow_\beta t[x := u]$$

α-conversion We however need to ensure that the variables in the substituted terms do not overlap and if they do, we need to rename them. This is called *α-conversion* or α renaming. In general, the variables that are not bound in λ-abstractions, *free variables*. may need to be replaced before every β-reduction so that they do not become *bound* after substitution.

$$(\lambda x.\ t) \longrightarrow_\alpha (\lambda y.\ t[x := y])$$

η-conversion Reducing a λ-abstraction that directly applies its argument to a term or equivalently, rewriting a term in the form of $\lambda x.f\ x$ to f is called *η-reduction*. The opposite rewrite rule, from f to $\lambda x.f\ x$ is $\bar{\eta}$-expansion, and because the rewriting works in both ways, it is also called the *η-conversion*.

$$\lambda x.f\ x \longrightarrow_\eta f$$
$$f \longrightarrow_{\bar{\eta}} \lambda x.f\ x$$

δ-reduction β-reduction together with α-renaming are sufficient to specify λ-calculus, but there are three other rewriting rules that we will need later: *δ-reduction* is the replacement of a constant with its definition.

Reduce args	Reduce under abstraction: **Yes**	**No**
Yes	$E := \lambda x.E \mid x\ E_1 ... E_n$ Normal form	$E := \lambda x.e \mid x\ E_1 ... E_n$ Weak normal form
No	$E := \lambda x.E \mid x\ e_1 ... e_n$ Head normal form	$E := \lambda x.e \mid x\ e_1 ... e_n$ Weak head normal form

Figure 2.4: Normal forms in λ-calculus

$$id\ t \longrightarrow_\delta (\lambda x.x)\ t$$

ζ-reduction For local variables, equivalent process is called the *ζ-reduction*.

$$let\ id = \lambda x.x\ in\ id\ t \longrightarrow_\zeta (\lambda x.x)\ t$$

ι-reduction We will also use other types of objects than just functions. Applying a function that extracts a value from an object is called the *ι-reduction*. In this example, the object is a pair of values, and the function π_1 is a projection that extracts the first value of the pair.

$$\pi_1(a, b) \longrightarrow_\iota a$$

Normal form By repeatedly βδιζ-reducing an expression–applying functions to their arguments, replacing constants and local variables with their definitions, evaluating objects, and α-renaming variables if necessary, we get a β-normal form, or just *normal form* for short. This normal form is unique up to α-conversion, according to the Church-Rossier theorem.

$$\begin{aligned}
& \text{let } pair = \lambda m.(m, m) \text{ in } \pi_1\ (pair\ (id\ 5)) \\
\longrightarrow_\zeta\ & \pi_1\ ((\lambda m.(m, m))\ (id\ 5)) \\
\longrightarrow_\beta\ & \pi_1\ (id\ 5, id\ 5) \\
\longrightarrow_\iota\ & id\ 5 \\
\longrightarrow_\delta\ & (\lambda x.x)\ 5 \\
\longrightarrow_\beta\ & 5
\end{aligned}$$

Other normal forms There are also other normal forms, they all have something to do with unapplied functions. If we have an expression and repeatedly use only the β-reduction, we end up with a function, or a variable applied to some free variables. These other normal forms specify what happens in such a "stuck" case. In Figure 2.4, e is an arbitrary λ-term and E is a term in the relevant normal form [**?**]. Closely related to the concept of a normal form are *normalization strategies* that specify the order in which sub-expressions are reduced.

Strong normalization An important property of a model of computation is termination, the question of whether there are expressions for which computation does not stop. In the context of the λ-calculus it means whether there are terms, where repeatedly applying rewriting rules does not produce a unique normal form in a finite sequence steps. While for some expressions this may depend on the selected rewriting strategy, the general property is as follows: If for all well-formed terms a there does not exist any infinite sequence of reductions $a \longrightarrow_\beta a' \longrightarrow_\beta a'' \longrightarrow_\beta \cdots$, then such a system is called *strongly normalizing*.

The untyped λ-calculus is not a strongly normalizing system, though, and there are expressions that do not have a normal form. When such expressions are reduced, they do not get smaller, but they *diverge*. The ω combinator:

$$\omega = \lambda x.x\ x$$

is one such example that produces an infinite term. Applying ω to itself produces a divergent term whose reduction cannot terminate:

$$\omega\ \omega \longrightarrow_\delta (\lambda x.x\ x)\omega \longrightarrow_\beta \omega\ \omega$$

The fixed-point function, the Y combinator, is also notable:

$$Y = \lambda f.(\lambda x.f\,(x\ x))\ (\lambda x.f\,(x\ x))$$

This is one possible way of encoding general recursion in λ-calculus, as it reduces by applying f to itself:

$$Y\,f \longrightarrow_{\delta\beta} f\,(Y\,f) \longrightarrow_{\delta\beta} f\,(f\,(Y\,f)) \longrightarrow_{\delta\beta} \ldots$$

This, as we will see in the following chapter, is impossible to encode in the typed λ-calculus without additional extensions.

As simple as λ-calculus may seem, it is a Turing-complete system that can encode logic, arithmetic, or data structures. Some examples include *Church encoding* of booleans, pairs, or natural numbers (Figure 2.5).

2.2.2 λ→-calculus

It is often useful, though, to describe the kinds of objects we work with. Already, in Figure 2.5 we could see that reading such expressions can get confusing: a boolean is a function of two parameters, whereas a pair is a function of three arguments, of which the first one needs to be a boolean and the other two contents of the pair.

The untyped λ-calculus defines a general model of computation based on functions and function application. Now we will restrict this model using types that describe the values that can be computed with.

The simply typed λ-calculus, also written λ→ as "→" is the connector used in types, introduces the concept of types. We have a set of basic types that are connected into terms

$$\begin{aligned} 0 &= \lambda f.\lambda x.\, x \\ 1 &= \lambda f.\lambda x.\, f\ x \end{aligned}$$

(a) Natural numbers

$$\begin{aligned} succ &= \lambda n.\lambda f.\lambda x.f\ (n\ f\ x) \\ plus &= \lambda m.\lambda n.m\ succ\ n \end{aligned}$$

(b) Simple arithmetic

$$\begin{aligned} true &= \lambda x.\lambda y.x \\ false &= \lambda x.\lambda y.y \\ not &= \lambda p.p\ false\ true \\ and &= \lambda p.\lambda q.p\ q\ p \\ ifElse &= \lambda p.\lambda a.\lambda b.p\ a\ b \end{aligned}$$

(c) Logic

$$\begin{aligned} cons &= \lambda f.\lambda x.\lambda y.f\ x\ y \\ fst &= \lambda p.p\ true \\ snd &= \lambda p.p\ false \end{aligned}$$

(d) Pairs

Figure 2.5: Church encoding of various concepts

using the arrow $\rightarrow$, and type annotation or assignment $x : A$. We now have two languages: the language of terms, and the language of types. These languages are connected by a *type judgment*, or *type assignment* $x : T$ that asserts that the term x has the type T [**?**].

Church- and Curry-style There are two ways of formalizing the simply-typed λ-calculus: $\lambda\rightarrow$-Church, and $\lambda\rightarrow$-Curry. Church-style is also called system of typed terms, or the explicitly typed λ-calculus as we have terms that include type information, and we say:

$$\lambda x : A.x : A \rightarrow A,$$

or using parentheses to clarify the precedence

$$\lambda(x : A).x : (A \rightarrow A).$$

Curry-style is also called the system of typed assignment, or the implicitly type λ-calculus as we assign types to untyped λ-terms that do not carry type information by themselves, and we say $\lambda x.x : A \rightarrow A$. [**?**].

There are systems that are not expressible in Curry-style, and vice versa. Curry-style is interesting for programming, we want to omit type information; and we will see how to manipulate programs specified in this way in Chapter 3. We will use Church-style in this chapter, but our language will be Curry-style, so that we incorporate elaboration into the interpreter.

Well-typed terms Before we only needed evaluation rules to fully specify the system, but specifying a system with types also requires typing rules that describe what types are allowed. We will also need to distinguish *well-formed terms* from *well-typed terms*: well-formed terms are syntactically valid, whereas well-typed terms also obey the typing rules. Terms that are well-formed but not yet known to be well typed are called *pre-terms*, or terms of *pre-syntax*.

There are some basis algorithms of type theory, in brief:

- given a pre-term and a type, *type checking* verifies if the term can be assigned the type.
- given just a pre-term and no type, *type inference* computes the type of an expression
- and finally *type elaboration* is the process of converting a partially specified pre-term into a complete, well-typed term [**?**].

Types and context The complete syntax of the λ→-calculus is in Figure 2.6. Reduction operations are the same as in the untyped lambda calculus, but we will need to add the language of types to the previously specified language of terms. This language consists of a set of *base types* which can consist of e.g. natural numbers or booleans, and *composite types*, which describe functions between them. We also need a way to store the types of terms that are known, a typing *context*, which consists of a list of *type judgments* in the form $x : T$, which associate variables to their types.

e			(*terms*)
	:=	v	variable
	|	$M\ N$	application
	|	$\lambda x.\ t$	abstraction
	|	$x : \tau$	annotation
τ			(*types*)
	:=	β	base types
	|	$\tau \to \tau'$	composite type
Γ			(*typingcontext*)
	:=	$\emptyset$	empty context
	|	$\Gamma, x : \tau$	type judgement
v			(*values*)
	:=	$\lambda x.\ t$	closure

Figure 2.6: λ→-calculus syntax

Typing rules The simply-typed λ-calculus can be completely specified by the typing rules in Figure 2.7 [**?**]. These rules are read similarly to logic proof trees: as an example, the rule **App** can be read as "if we can infer f with the type $A \to B$ and a with the type A from the context Γ, then we can also infer that function application $f\ a$ has the type B". Given these rules and the formula

$$\lambda a : A.\lambda b : B.a : A \to B \to A$$

we can also produce a derivation tree that looks similar to logic proofs and, as mentioned before, its semantics corresponding to the logic formula "if A and B, then A" as per the Curry-Howard equivalence.

$$\dfrac{\dfrac{\dfrac{}{a : A, b : B \vdash a : A}}{a : A \vdash \lambda b : B.a : B \to A}}{\vdash \lambda a : A.\lambda b : B.a : A \to B \to A}$$

We briefly mentioned the problem of termination in the previous section; the simply-typed λ-calculus is strongly normalizing, meaning that all well-typed terms have a unique normal form. In other words, there is no way of writing a well-typed divergent term; the Y combinator is impossible to type in $\lambda\!\to$ and any of the systems in the next chapter [**?**].

$$\frac{x : A \in \Gamma}{\Gamma \vdash x : A}\ (\textsc{Var})$$

$$\frac{\Gamma \vdash f : A \to B \qquad \Gamma \vdash a : A}{\Gamma \vdash fa : B}\ (\textsc{App})$$

$$\frac{\Gamma, x : A \vdash b : B}{\Gamma \vdash \lambda x : A.b : A \to B}\ (\textsc{Abs})$$

Figure 2.7: $\lambda\!\to$-calculus typing rules

2.2.3 λ-cube

The $\lambda\!\to$-calculus restricts the types of arguments to functions; types are static and descriptive. When evaluating a well-typed term, the types can be erased altogether without any effect on the computation. In other words, terms can only depend on other terms.

Generalizations of the $\lambda\!\to$-calculus can be organized into a cube called the Barendregt cube, or the λ-cube [**?**] (Figure 2.8). In $\lambda\!\to$ only terms depend on terms, but there are also three other combinations represented by the three dimensions of the cube: types depending on types $(\Box, \Box)$, or also called type operators; terms depending on types $(\Box, \star)$, called *polymorphism*; and terms depending on types $(\star, \Box)$, representing *dependent types*.

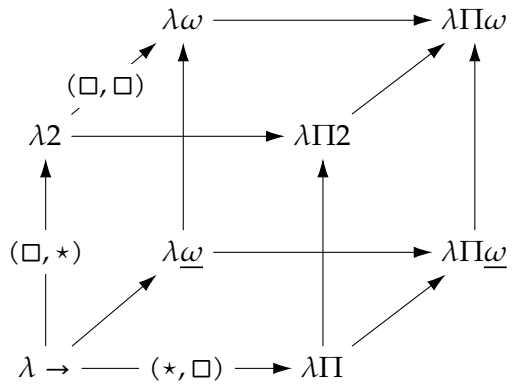

Figure 2.8: Barendregt cube (also λ-cube)

Sorts To formally describe the cube, we will need to introduce the notion of sorts. In brief,

$$t : T : \star : \Box.$$

The meaning of the symbol : is same as before, "x has type y". The type of a term t is a type T, the type of a type T is a kind $*$, and the type of kinds is the sort $\Box$. The symbols $\star$ and $\Box$ are called *sorts*. As with types, sorts can be connected using arrows, e.g. $(\star \to \star) \to \star$. To contrast the syntaxes of the following languages, the syntax of $\lambda\!\to$ is here:

$$\begin{array}{lcl} \mathit{types} & := & T \mid A \to B \\ \mathit{terms} & := & v \mid \lambda x : A.t \mid a\ b \\ \mathit{values} & := & \lambda x : A.t \end{array}$$

$\lambda\underline{\omega}$-calculus Higher-order types or type operators generalizes the concepts of functions to the type level, adding λ-abstractions and applications to the language of types.

$$\begin{array}{lcl} \mathit{types} & := & T \mid A \to B \mid A\ B \mid \Lambda A.B(a) \\ \mathit{terms} & := & v \mid \lambda x : A.t \mid a\ b \\ \mathit{values} & := & \lambda x : A.t \end{array}$$

λ2-calculus The dependency of terms on types adds polymorphic types to the language of types: $\forall X : k.A(X)$, and type abstractions (Λ-abstractions) and applications to the language of terms. This system is also called System F, and it is equivalent to propositional logic [**?**].

$$\begin{array}{lcl} \mathit{types} & := & T \mid A \to B \mid \forall A.B \\ \mathit{terms} & := & v \mid \lambda x : A.t \mid a\ b \mid \Lambda A.t \\ \mathit{values} & := & \lambda x : A.t \mid \Lambda A.t \end{array}$$

$\lambda\Pi$-calculus Allowing types to depend on terms means that type of a function can depend on its term-level arguments, hence dependent types, represented by the type $\Pi a : A.B(a)$. This dependency is the reason for the name of dependently-typed languages. This system is well-studied as the Logical Framework (LF) [**?**].

$$\begin{array}{lcl} \mathit{types} & := & T \mid A \to B \mid \Pi a : A.B \\ \mathit{terms} & := & v \mid \lambda x : A.b \mid a\ b \mid \Pi a : A.b \\ \mathit{values} & := & \lambda x : A.b \mid \Pi x : A.b \end{array}$$

Pure type system These systems can all be described by one set of typing rules instantiated with a triple (S, A, R). Given the set of sorts $S = \{\star, \Box\}$ we can define relations A and R where, for example, $A = \{(\star, \Box)\}$ is translated to the axiom $\vdash \star : \Box$ by the rule **Start**, and $R = \{(\star, \Box)\}$[1] means that a kind can depend on a type using the rule **Product**.

[1]The elements of R are written as (s_1, s_2), which is equivalent to (s_1, s_2, s_2).

$$
\begin{array}{lcll}
S & := & \{\star, \Box\} & \text{set of sorts} \\
A & \subseteq & S \times S & \text{set of axioms} \\
R & \subseteq & S \times S \times S & \text{set of rules}
\end{array}
$$

The typing rules in Figure 2.9 apply to all the above-mentioned type systems. The set R exactly corresponds to the dimensions of the λ-cube, so instantiating this type system with $R = \{(\star, \star)\}$ would produce the $\lambda{\to}$-calculus, whereas including all the dependencies $R = \{(\star, \star), (\Box, \star), (\star, \Box), (\Box, \Box)\}$ produces the $\lambda\Pi\omega$-calculus. If we also consider that the function arrow $A \to B$ is exactly equivalent to the type $\Pi a : A.B(a)$ if the variable a is not used in the expression $B(a)$, the similarity to Figure 2.7 should be easy to see.

$$
\frac{}{\vdash s_1 : s_2}\ (s_1, s_2) \in A \qquad \text{(Start)}
$$

$$
\frac{\Gamma \vdash A : s}{\Gamma, x : A \vdash x : A}\ s \in S \qquad \text{(Var)}
$$

$$
\frac{\Gamma \vdash x : A \qquad \Gamma \vdash B : s}{\Gamma, y : B \vdash x : A}\ s \in S \qquad \text{(Weaken)}
$$

$$
\frac{\Gamma \vdash f : \Pi_{x:A} B(x) \qquad \Gamma \vdash a : A}{\Gamma \vdash fa : B[x := a]} \qquad \text{(App)}
$$

$$
\frac{\Gamma, x : A \vdash b : B \qquad \Gamma \vdash \Pi_{x:A} B(x) : s}{\Gamma \vdash (\lambda x : A.b) : \Pi_{x:A} B(x)}\ s \in S \qquad \text{(Abs)}
$$

$$
\frac{\Gamma \vdash A : s_1 \qquad \Gamma, x : A \vdash B : s_2}{\Gamma \vdash \Pi_{x:A} B(x) : s_3}\ (s_1, s_2, s_3) \in R \qquad \text{(Product)}
$$

$$
\frac{\Gamma \vdash a : A \qquad \Gamma \vdash A' : s \qquad A \longrightarrow_\beta A'}{\Gamma \vdash a : A'}\ s \in S \qquad \text{(Conv)}
$$

Figure 2.9: Typing rules of a pure type system

Universes This can be generalized even more. Instantiating this system with an infinite set of sorts $S = \{Type_0, Type_1, ...\}$ instead of the set $\{\star, \Box\}$ and setting A to $\{(Type_0, Type_1), (Type_1, Type_2), ...\}$ leads to an infinite hierarchy of *type universes*, and is in fact an interesting topic in the field of type theory. Proof assistants commonly use such a hierarchy [**?**].

Type in Type Going the other way around, simplifying S to $S = \{\star\}$ and setting A to $\{(\star, \star)\}$, leads to an inconsistent logic system called $\lambda\star$, also called a system with a *Type in Type* rule. This leads to paradoxes similar to the Russel's paradox in set theory.

Maybe show Girard's paradox?

In many pedagogic implementations of dependently-typed λ-calculi I saw, though, this was simply acknowledged: separating universes introduces complexity but the distinction is not as important for many purposes.

For the goal of this thesis–testing the characteristics of a runtime system–the distinction is unimportant. In the rest of the text we will use the inconsistent $\lambda\star$-calculus, but with all the constructs mentioned in the preceding type systems. We will now formally define

these constructs, together with several extensions to this system that will be useful in the context of just-in-time compilation using Truffle, e.g., (co)product types, booleans, natural numbers.

Proof assistants and other dependently-typed programming languages use systems based on $\lambda\Pi\omega$-calculus, which is called the Calculus of Constructions. They add more extensions: induction and subtyping are common ones. We will discuss only a subset of them in the following section, as many of these are irrelevant to the goals of this thesis.

2.3 Types

While it is possible to derive any types using only three constructs: Π-types (dependent product), Σ-types (dependent sum), and W-types (inductive types), that we haven't seen so far; we will define specific *"wired-in"* types instead, as they are more straightforward to both use and implement.

We will specify the syntax and semantics of each type at the same time. For syntax, we will define the terms and values, for semantics we will use four parts: type formation, a way to construct new types; term introduction (constructors), ways to construct terms of these types; term elimination (destructors), ways to use them to construct other terms; and computation rules that describe what happens when an introduced term is eliminated. The algorithms to normalize and type-check these terms will be mentioned in the following chapter. In this section we will solely focus on the syntax and semantics.

2.3.1 Π-types

As mentioned above, the type $\Pi a : A.B$, also called the *dependent product type* or the *dependent function type*, is a generalization of the function type $A \to B$. Where the function type simply asserts that its corresponding function will receive a value of a certain type as its argument, the Π-type makes the value available in the rest of the type. Figure 2.10 introduces its semantics; they are similar to the typing rules of $\lambda\to$-calculus function application, except for the substitution in the type of B in rule **Elim-Pi**.

$$\frac{\Gamma \vdash A : \star \qquad \Gamma, x : A \vdash B : \star}{\Gamma \vdash \Pi x : A.B}\ \textbf{(Type-Pi)}$$

$$\frac{\Gamma, a : A \vdash b : B}{\Gamma \vdash \lambda x.b\ :\ \Pi x : A.B}\ \textbf{(Intro-Pi)} \qquad \frac{\Gamma \vdash f\ :\ \Pi x : A.B \qquad \Gamma \vdash a : A}{\Gamma \vdash f\ a : B[x := a]}\ \textbf{(Elim-Pi)}$$

$$\frac{\Gamma, a : A \vdash b : B \qquad \Gamma \vdash a : A}{\Gamma \vdash (\lambda x : A.b)a \longrightarrow_\beta b[x := a]}\ \textbf{(Eval-Pi)}$$

Figure 2.10: Π-type semantics

While a very common example of a Π-type is the length-indexed vector $\Pi(n : \mathbb{N}).Vec(\mathbb{R}, n)$, it is also possible to define a function with a *"dynamic"* number of arguments like in the following listing. It is a powerful language feature also for its programming uses, as it makes it possible to e.g. implement a well-typed function `printf` that, e.g., produces the function $Nat \to Nat \to String$ when called as `printf "%d%d"`.

$$
\begin{array}{rcl}
\mathit{succOrZero} & : & \Pi(b : \mathit{Bool}).\ \mathit{if}\ b\ \mathit{then}\ (\mathit{Nat} \to \mathit{Nat})\ \mathit{else}\ \mathit{Nat} \\
\mathit{succOrZero} & = & \Pi(b : \mathit{Bool}).\ \mathit{if}\ b\ \mathit{then}\ (\lambda x.\ x + 1)\ \mathit{else}\ 0 \\
\mathit{succOrZero}\ \mathit{true}\ 0 & \longrightarrow_{\beta\delta} & 1 \\
\mathit{succOrZero}\ \mathit{false} & \longrightarrow_{\beta\delta} & 0
\end{array}
$$

Implicit arguments The type-checker can infer many type arguments. Agda adds the concept of implicit function arguments [**?**] to ease the programmer's work and mark inferrable type arguments in a function's type signature. Such arguments can be specified when calling a function using a special syntax, but they are not required [**?**]. We will do the same, and as such we will split the syntax of a Π-type back into three separate constructs, which can be seen in Figure 2.11.

$$
\begin{array}{rcllll}
\mathit{term} & := & a \to b & \mid\ (a : A) \to b & \mid\ \{a : A\} \to b & \text{(abstraction)} \\
 & \mid & f\ a & \mid & \mid\ f\ \{a\} & \text{(application)} \\
\mathit{value} & := & \Pi a : A.b & & &
\end{array}
$$

Figure 2.11: Π-type syntax

The plain *function type* $A \to B$ is simple to type but does not bind the value provided as the argument A. The *explicit Π-type* $(a : A) \to B$ binds the value a and makes it available to use inside B, and the *implicit Π-type* $\{a : A\} \to B$ marks the argument as one that type elaboration should be able to infer from the surrounding context. The following is an example of the implicit argument syntax, a polymorphic function *id*.

$$
\begin{array}{rcllcl}
\mathit{id} & : & \{A : \star\} \to A \to A & := & \Pi(x : A).x \\
\mathit{id}\ \{\mathit{Nat}\} & : & \mathit{Nat} \to \mathit{Nat} & \longrightarrow_{\beta\delta} & \lambda(x : \mathit{Nat}).x \\
\mathit{id}\ 1 & : & \mathit{Nat} & \longrightarrow_{\beta\delta} & 1
\end{array}
$$

2.3.2 Σ-types

The Σ-type is also called the *dependent pair type*, or alternatively the dependent tuple, dependent sum, or even the dependent product type. Like the Π-type was a generalization of the function type, the Σ-type is a generalization of a product type, or simply a *pair*. Semantically, the Σ-type is similar to the tagged union in C-like languages: the type $\Sigma(a : A).B(a)$ corresponds to a value (a, b), only the type $B(a)$ can depend on the first member of the pair. This is illustrated in Figure 2.12, where the dependency can be seen in rule **Intro-Sigma**, in the substitution $B[x := a]$.

Above, we had a function that could accept different arguments based on the value of the first argument. Below we have a type that simply uses Σ in place of Π in the type: based on the value of the first member, the second member can be either a function or a value, and still be a well-typed term.

$$
\begin{array}{rcl}
\mathit{FuncOrVal} & : & \Sigma(b : \mathit{Bool}).\ \mathit{if}\ b\ \mathit{then}\ (\mathit{Nat} \to \mathit{Nat})\ \mathit{else}\ \mathit{Nat} \\
(\mathit{true}, \lambda x.\ x + 1) & : & \mathit{FuncOrVal} \\
(\mathit{false}, 0) & : & \mathit{FuncOrVal}
\end{array}
$$

$$\frac{\Gamma \vdash A : \star \qquad \Gamma, x : A \vdash B : \star}{\Gamma \vdash \Sigma_{x:A} B : \star} \textbf{(Type-Sigma)}$$

$$\frac{\Gamma \vdash a : A \qquad \Gamma, x : A \vdash B : \star \qquad \Gamma \vdash b : B[x := a]}{\Gamma \vdash (a, b) : \Sigma_{x:A} B} \textbf{(Intro-Sigma)}$$

$$\frac{\Gamma \vdash p : \Sigma_{x:A} B}{\Gamma \vdash \pi_1\ p : A} \textbf{(Elim-Sigma1)} \qquad \frac{\Gamma \vdash p : \Sigma_{x:A} B}{\Gamma \vdash \pi_2\ p : B[x := fst\ p]} \textbf{(Elim-Sigma2)}$$

$$\frac{\Gamma \vdash a : A \qquad \Gamma, x : A \vdash B : \star \qquad \Gamma \vdash b : B[x := a]}{\Gamma \vdash \pi_1\ (a, b) \longrightarrow_\iota a : A} \textbf{(Eval-Sigma1)}$$

$$\frac{\Gamma \vdash a : A \qquad \Gamma, x : A \vdash B : \star \qquad \Gamma \vdash b : B[x := a]}{\Gamma \vdash \pi_2\ (a, b) \longrightarrow_\iota b : B} \textbf{(Eval-Sigma2)}$$

Figure 2.12: Σ-type semantics

Pair Similar to the function type, given the expression $\Sigma(a : A).B(a)$, if a does not occur in the expression $B(a)$, then it is the non-dependent pair type. The pair type is useful to express an isomorphism also used in general programming practice: a conversion between a function of two arguments, and a function of one argument that returns a function of one argument:

$$\begin{array}{rcll} & A \times B \to C & \Leftrightarrow & A \to B \to C \\ curry & := & \lambda(f : A \times B \to C). & \lambda(x : A).\lambda(y : B).\ f\ (x, y) \\ uncurry & := & \lambda(f : A \to B \to C). & \lambda(x : A \times B).\ f\ (\pi_1\ x)\ (\pi_2\ y) \end{array}$$

Tuple The n-tuple is a generalization of the pair, a non-dependent set of an arbitrary number of values, otherwise expressible as a set of nested pairs: commonly written as $(a_1, ..., a_n)$.

Record A record type is similar to a tuple, only its members have unique labels. In Figure 2.13 we see the semantics of a general record type, using the notation $\{l_i = t_i\} : \{l_i : T_i\}$ and a projection *record.member*.

$$\frac{\forall i \in \{1..n\}\ \Gamma \vdash T_i : \star}{\Gamma \vdash \{l_i : T_i^{i \in \{1..n\}}\} : \star} \textbf{(Type-Rec)}$$

$$\frac{\forall i \in \{1..n\} \quad \Gamma \vdash t_i : T_i}{\Gamma \vdash \{l_i = t_i^{i \in \{1..n\}}\} : \{l_i : T_i^{i \in \{1..n\}}\}} \textbf{(Intro-Rec)}$$

$$\frac{\Gamma \vdash t : \{l_i : T_i^{i \in \{1..n\}}\}}{\Gamma \vdash t.l_i : T_i} \textbf{(Elim-Rec)}$$

$$\frac{\forall i \in \{1..n\}\ \Gamma \vdash t_i : T_i \qquad \Gamma \vdash t : \{l_i : T_i^{i \in \{1..n\}}\}}{\Gamma \vdash \{l_i = t_i^{i \in \{1..n\}}\}.l_i \longrightarrow_\iota t_i : B} \textbf{(Eval-Rec)}$$

Figure 2.13: Record semantics

In Figure 2.14 we have a syntax that unifies all of these concepts: a Σ-type, a pair, an n-tuple, a named record. A non-dependent n-tuple type is written as $A \times B \times C$ with values (a, b, c).

Projections of non-dependent tuples use numbers, e.g., $p.1$, $p.2$, ... A dependent sum type is written in the same way as a named record: $(a : A) \times B$ binds the value $a : A$ in the rest of the type B, and on the value-level enables the projection $obj.a$.

$$\begin{array}{rcll} term & := & T_1 \times \cdots \times T_n \mid (l_1 : T_1) \times \cdots \times (l_n : T_n) \times T_{n+1} & \text{(types)} \\ & \mid & t.i \mid t.l_n & \text{(destructors)} \\ & \mid & (t_1, \cdots, t_n) & \text{(constructor)} \\ value & := & (t_1, \cdots, t_n) & \end{array}$$

Figure 2.14: Σ-type syntax

Coproduct The sum type or the coproduct $A + B$ can have values from both types A and B, often written as $a : A \vdash inl\ A : A + B$, where *inl* means "on the left-hand side of the sum $A + B$". This can be generalized to the concept of *variant types*, with an arbitrary number of named members; shown below, using Haskell syntax:

$$data\ Maybe\ a = Nothing \mid Just\ a$$

For the purposes of our language, a binary sum type is useful, but inductive variant types would require more involved constraint checking, so we will ignore those, only using simple sum types in the form of $A + B$. This type can be derived using a dependent pair where the first member is a boolean.

$$Char + Int \quad \simeq \quad \Sigma(x : Bool).\ if\ x\ Char\ Int$$

2.3.3 Value types

Finite sets Pure type systems mentioned in the previous chapter often use types like **0**, **1**, and **2** with a finite number of inhabitants, where the type **0** (with zero inhabitants of the type) is the empty or void type. Type **1** with a single inhabitant is the unit type, and the type **2** is the boolean type. Also, the infinite set of natural numbers can be defined using induction over **2**. For our purposes it is enough to define a fixed number of types, though.

Unit The unit type **1**, or commonly written as the 0-tuple "`()`", is sometimes used as a universal return value. As it has no evaluation rules, though, we can simply add a new type *Unit* and a new value and term *unit*, with the rule $unit : Unit$.

Booleans The above-mentioned type **2** has two inhabitants and can be semantically mapped to the boolean type. In Figure 2.15 we introduce the values (constructors) *true* and $false$, and a simple eliminator *if* that returns one of two values based on the truth value of its argument.

$$\frac{}{\vdash Bool : \star}\ \textbf{(Type-Nat)}$$

$$\frac{}{\vdash true : Bool}\ \textbf{(Intro-True)} \qquad \frac{}{\vdash false : False}\ \textbf{(Intro-False)}$$

$$\frac{\Gamma \vdash a_1 : A \qquad \Gamma \vdash a_2 : A}{\Gamma, x : Bool \vdash if\ x\ a_1\ a_2 : A}\ \textbf{(Elim-Bool)}$$

$$\frac{\Gamma \vdash a_1 : A \qquad \Gamma \vdash a_2 : A}{\Gamma \vdash if\ true\ a_1\ a_2 \longrightarrow_\iota a_1 : A}\ \textbf{(Eval-True)} \qquad \frac{\Gamma \vdash a_1 : A \qquad \Gamma \vdash a_2 : A}{if\ false\ a_1\ a_2 \longrightarrow_\iota a_2 : A}\ \textbf{(Eval-False)}$$

Figure 2.15: `Bool` semantics

Natural numbers The natural numbers form an infinite set, unlike the above value types. On their own, adding natural numbers to a type system does not produce non-termination, as the recursion involved in their manipulation can be limited to primitive recursion as e.g., used in Gödel's System T [**?**]. The constructions introduced in Figure 2.16 are simply the constructors *zero* and *succ*, and the destructor *natElim* unwraps at most one layer of *succ*.

Alternatively, a dependent evaluator into ΠIn:Nat.T(n) although unused, e.g. ncatlab

$$\frac{}{\vdash Nat : \star}\ \textbf{(Type-Nat)}$$

$$\frac{}{\vdash zero : Nat}\ \textbf{(Intro-Zero)} \qquad \frac{\Gamma \vdash n : Nat}{\Gamma \vdash succ\ n : Nat}\ \textbf{(Intro-Succ)}$$

$$\frac{\Gamma \vdash a_1 : A \qquad \Gamma, n : Nat \vdash a_2 : A}{\Gamma, x : Nat \vdash natElim\ x\ a_1\ (\lambda x.a_2)}\ \textbf{(Elim-Nat)}$$

$$\frac{\Gamma \vdash a_1 : A \qquad \Gamma, n : Nat \vdash a_2 : A}{\Gamma \vdash natElim\ zero\ a_1\ (\lambda x.a_2) \longrightarrow_\iota a_1 : A}\ \textbf{(Eval-Zero)}$$

$$\frac{\Gamma \vdash a_1 : A \qquad \Gamma, n : Nat \vdash a_2 : A \qquad \Gamma \vdash n : Nat}{natElim\ (succ\ n)\ a_1\ (\lambda x.a_2) \longrightarrow_\iota a_2[x := n] : A}\ \textbf{(Eval-Succ)}$$

Figure 2.16: `Nat` semantics

2.3.4 μ-types

There are multiple ways of encoding recursion in λ-calculi with types, based on whether a recursive expression is delimited using types, or whether it is also reflected in the type of a recursive expression. Recursion must be defined carefully if the type system needs to be consistent, as non-restricted general recursion leads to non-termination and inconsistency. *Iso-recursive types* use explicit folding and unfolding operations, that convert between the recursive type $\mu a.T$ and $T[a := \mu a.T]$, whereas in *equi-recursive types* these operations are implicit and inserted by the type-checker.

As both complicate the type-checker, we will use a simpler value-level recursive combinator fix. While this does compromise the consistency of the type system, it is sufficient for the purposes of runtime system characterization.

$$\frac{\Gamma \vdash f : A \to A}{\Gamma \vdash fix\, f : A}\ \textbf{(Type-Fix)}$$

$$\frac{\Gamma, x : A \vdash t : A}{\Gamma \vdash fix\ (\lambda x.t) \longrightarrow_\beta t[x := (\lambda x.t)] : A}\ \textbf{(Eval-Fix)}$$

Figure 2.17: `fix` semantics

The semantics of the function fix are described in Figure 2.17. This definition is sufficient to define e.g., the recursive computation of a Fibonacci number or a local recursive binding as below.

```
fib = fix (λf. λn. if (isLess n 2) n (add (f (n - 1)) (f (n-2))))

evenOdd
  : (isEven : Nat → Bool) × (isOdd : Nat → Bool) × Top
  = fix (λf. ( if isZero x then true else f.isOdd (pred x)
             , if isZero x then false else f.isEven (pred x)
             , Top
             ))
```

2.4 Remaining constructs

These constructs together form a complete core language capable of forming and evaluating expressions. Already, this would be a usable programming language. However, the *surface language* is still missing: the syntax for defining constants and variables, and interacting with the compiler.

Local definitions The λ-calculus is, to use programming language terminology, a purely functional programming language: without specific extensions, any language construct is an expression. We will use the syntax of Agda, and keep local variable definition as an expression as well, using a *let* − *in* construct, with the semantics given in Figure 2.18.

$$\frac{\Gamma \vdash a : A \qquad \Gamma, x : A \vdash b : B}{\Gamma \vdash \text{let } x = a \text{ in } b : B}\ \textbf{(Type-Let)}$$

$$\frac{\Gamma \vdash v : A \qquad \Gamma, x : A \vdash e : B}{\text{let } x = v \text{ in } e \longrightarrow_\zeta e[x := v]}\ \textbf{(Eval-Let)}$$

Figure 2.18: *let* − *in* semantics

Global definitions Global definitions are not strictly necessary, as with local definitions and the fixed-point combinator we could emulate them. However, global definitions will be useful later in the process of elaborations, when global top-level definitions will separate blocks that we can type-check separately. We will add three top-level expressions: a declaration that only assigns a name to a type, and a definition with and without type. Definitions without types will have them inferred.

$$
\begin{array}{rcl}
top & := & id : term \\
& | & id : term = term \\
& | & id = term
\end{array}
$$

Holes A construct that serves solely as information to the compiler and will not be used at runtime is a *hole*. It can take the place of a term in an expression and marks the missing term as one to be inferred ("filled in") during elaboration[2]. In fact, the syntax for a global definition without a type will use a hole in place of its type. The semantics of a hole are omitted on purpose as they would also require specifying the type inference algorithm.

$$
\begin{array}{rcl}
term & := & \_
\end{array}
$$

Interpreter directives Another type of top-level expressions is a pragma, a direct command to the compiler. We will use these when evaluating the time it takes to normalize or elaborate an expression, or when enabling or disabling the use of "wired-in" types, e.g. to compare the performance impact of using a Church encoding of numbers versus a natural type that uses hardware integers. We will once again use the syntax of Agda:

$$
\begin{array}{rcl}
top & := & \{-\#\ BUILTIN\ id\ \#-\} \\
& | & \{-\#\ ELABORATE\ term\ \#-\} \\
& | & \{-\#\ NORMALIZE\ term\ \#-\}
\end{array}
$$

Polyglot Lastly, one language feature that will only be described and implemented in Chapter 5: a *"polyglot"* construct that offers a way to execute code in a different language, which is a feature of the Truffle framework. The selected syntax is a three-part expression that contains the name of language to be used, the foreign code, and the type of the result of evaluating this foreign code:

$$
\begin{array}{rcl}
term & := & [|\ id\ |\ foreign\ |\ term\ |]
\end{array}
$$

The syntax and semantics presented here altogether comprise a working programming language. A complete listing of the semantics is included in Appendix B. The syntax, written using the notation of the ANTLR parser generator is in Listing 2.2. The syntax does not mention constants like *true* or *Nat*, as they will be implemented as global definitions bound in the initial type-checking context and do not need to be recognized during parsing.

With this, the language specification is complete, and we can move on to the next part, implementing a type-checker and an interpreter for this language.

[2]Proof assistants also use the concept of a metavariable, often with the syntax $?\alpha$.

```
FILE : STMT (STMTEND STMT)* ;
STMT : "{-#" PRAGMA "#-}"
     | ID ":" EXPR
     | ID (":" EXPR)? "=" EXPR
     ;
EXPR : "let" ID ":" EXPR "=" EXPR "in" EXPR
     | "λ" LAM_BINDER "." EXPR
     | PI_BINDER+ "→" EXPR
     | ATOM ARG*
     ;
LAM_BINDER : ID | "_" | "{" (ID | "_") "}" ;
PI_BINDER : ATOM ARG* | "(" ID+ ":" EXPR ")" | "{" ID+ ":" EXPR "}" ;
ARG : ATOM | "{" ID ("=" TERM)? "}" ;
ATOM : "[" ID "|" FOREIGN "|" TERM "]"
     | EXPR "×" EXPR
     | "(" EXPR ("," EXPR)+ ")"
     | "(" EXPR ")"
     | ID "." ID
     | ID
     | NAT
     | "*"
     | "_"
     ;
 STMTEND : ("\n" | ";")+ ;
 ID : [a-zA-Z] [a-zA-Z0-9] ;
 SKIP : [ \t] | "--" [^\r\n]* | "{-" [^#] .* "-}" ;
// pragma discussed in text
```

Listing 2.2: The complete grammar, written using simplified ANTLR

Chapter 3

Language implementation: Montuno

3.1 Introduction

Now with a complete language specification, we can move onto the next step: writing an interpreter. The algorithms involved can be translated from specification to code quite naturally. at least in the style of interpreter we will create at first. The second interpreter in Truffle will require a quite different programming paradigm and deciding on many low-level implementation details, e.g., how to implement actual function calls.

In this chapter we will introduce the algorithms at the core of an interpreter and build a tree-based implementation for the language, elaborating on key implementation decisions. This interpreter will be referred to using the working name Montuno[1].

This interpreter can be called an AST (abstract syntax tree) interpreter, as the principal parts all consist of tree traversals, due to the fact that all the main data structures involved are trees: pre-terms, terms, and values are recursive data structures. The main algorithms to be discussed are: evaluation, normalization, and elaboration, all of them can be translated to tree traversals in a straightforward way.

Language The choice of a programming language is mostly decided by the eventual target platform Truffle, as we will be able to share parts of the implementation between the two interpreters. The language of GraalVM and Truffle is Java, although other languages that run on the Java Virtual Machine can be used[2]. My personal preference lies with more functional languages like Scala or Kotlin, as the code often is cleaner and more concise[3], so in the end, after comparing the languages, I have selected Kotlin due to its multi-paradigm nature: Truffle requires the use of annotated classes, but this first interpreter can be written in a more natural functional style.

[1] Montuno, as opposed to the project Cadenza, to which this project is a follow-up. Both are music terms, *cadenza* being a "long virtuosic solo section", whereas *montuno* is a "faster, semi-improvised instrumental part".

[2] Even though Kotlin seems not to be recommended by Truffle authors, there are several languages implemented in it, which suggests there are no severe problems. "[...] and Kotlin might use abstractions that don't properly partially evaluate." (from https://github.com/oracle/graal/issues/1228)

[3] Kotlin authors claim 40% reduction in the number of lines of code, (from https://kotlinlang.org/docs/faq.html)

Libraries Truffle authors recommend against using many external libraries in the internals of the interpreter, as the techniques the libraries use may not work well with Truffle . Source Therefore, we will need to design our own supporting data structures based on the fundamental data structures provided directly by Kotlin. Only two external libraries would be too complicated to reimplement, and both of these were chosen because they are among the most widely used in their field:

- a parser generator, ANTLR, to process input into an abstract syntax tree,
- a terminal interface library, JLine, to implement the interactive interface.

For the build and test system, the recommended choices of Gradle and JUnit were used.

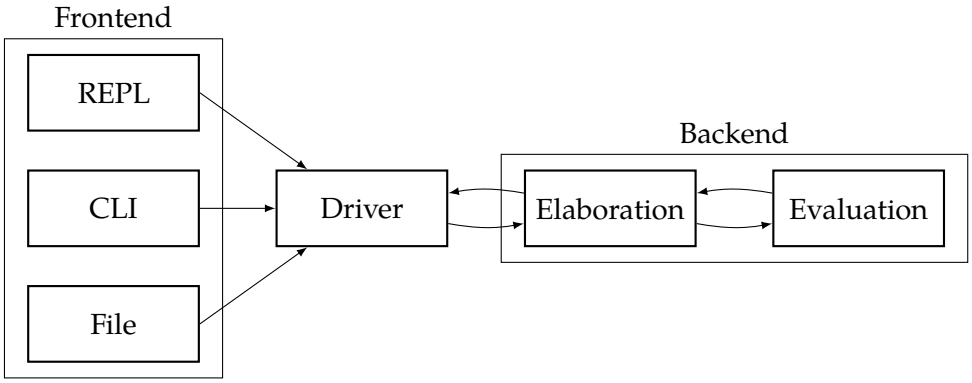

Figure 3.1: Overview of interpreter components

3.1.1 Program flow

A typical interpreter takes in the user's input, processes it, and outputs a result. In this way, we can divide the interpreter into a frontend, a driver, and a backend, to reuse compiler terminology. A frontend handles user interaction, be it from a file or from an interactive environment, a backend implements the language semantics, and a driver connects them, illustrated in Figure 3.1.

Frontend The frontend is intended to be a simple way to execute the interpreter, offering two modes: a batch processing mode that reads from a file, and an interactive terminal environment that receives user input and prints out the result of the command. Proof assistants like Agda offer deeper integration with editors like tactics-based programming or others, similar to the refactoring tools offered in development environments for object-oriented languages, but that is unnecessary for the purposes of this thesis.

Backend The components of the backend, here represented as *elaboration* and *evaluation*, implement the data transformation algorithms that are further illustrated in Figure 3.2. In brief, the *elaboration* process turns user input in the form of partially-typed, well-formed *pre-terms* into fully-annotated well-typed *terms*. *Evaluation* converts between a *term* and a *value*: a term can be compared to program data, it can only be evaluated, whereas a value is the result of such evaluation and can be e.g., compared for equality.

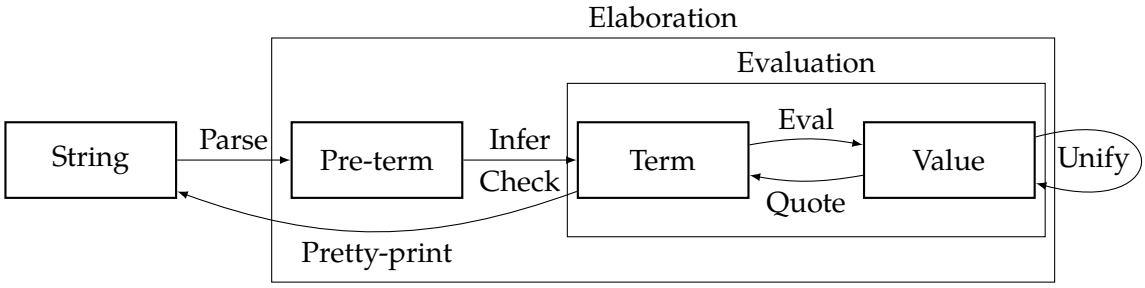

Figure 3.2: Data flow overview

Data flow In Figure 3.2, *Infer* and *Check* correspond to type checking and type inference, two parts of the *bidirectional typing* algorithm that we will use. *Unification* (*Unify*) forms a major part of the process, as that is how we check whether two values are equal. *Eval* corresponds to the previously described $\beta\delta\zeta\iota$-reduction implemented using the *normalization-by-evaluation* style, whereas *Quote* builds a term back up from an evaluated value. To complete the description, *Parse* and *Pretty-print* convert between the user-readable, string representation of terms and the data structures of their internal representation. For the sake of clarity, the processes are illustrated using their simplified function signatures in Listing 3.1.

```
fun parse(input: String): PreTerm;
fun pprint(term: Term): String;
fun infer(pre: PreTerm): Pair<Term, Val>;
fun check(pre: PreTerm, wanted: Val): Term;
fun eval(term: Term): Val;
fun quote(value: Val): Term;
fun unify(left: Val, right: Val): Unit;
```

Listing 3.1: Simplified signatures of the principal functions

In this chapter, we will first define the data types, especially focusing on closure representation. Then, we will specify and implement two algorithms: *normalization-by-evaluation*, and *bidirectional type elaboration*, and lastly, we finish the interpreter by creating its driver and frontend.

3.2 Data structures

We have specified the syntax of the language in the previous chapter, which we first need to translate to concrete data structures before trying to implement the semantics. Sometimes, the semantics impose additional constraints on the design of the data structures, but in this case, the translation is quite straight-forward.

Properties Terms and values form recursive data structures. We will also need a separate data structure for pre-terms as the result of parsing user input. All of these structures represent only well-formed terms and in addition, terms and value represent the well-typed subset of well-formed terms. Well-formedness should be ensured by the parsing process, whereas type-checking will take care of the second property.

Pre-terms As pre-terms are mostly just an encoding of the parse tree without much further processing, the complete data type is only included in Appendix C. The `PreTerm` class hierarchy mostly reflects the `Term` classes with a few key differences, like the addition of compiler directives or variable representation, so in the rest of this section, we will discuss terms and values only.

Location A key feature that we will also disregard in this chapter is term location that maps the position of a term in the original source expression, mostly for the purpose of error reporting. As location is tracked in a field that occurs in all pre-terms, terms, and values, it will only be included in the final listing of classes in Appendix C.

$$
\begin{array}{rcllll}
\mathit{term} & := & v & \mid \mathit{constant} & & \\
 & \mid & a\ b & \mid a\ \{b\} & & \\
 & \mid & a \to b & \mid (a : A) \to b & \mid \{a : A\} \to b & \\
 & \mid & a \times b & \mid (l : A) \times b & \mid a.l & \\
 & \mid & \text{let } x = v \text{ in } e & \mid [\mid id \mid foreign \mid type \mid] & & \\
 & \mid & \_ & & & \\
\mathit{value} & := & \mathit{constant} & & & \\
 & \mid & \lambda x : A.b & \mid \Pi x : A.b & & \\
 & \mid & (a_1, \cdots, a_n) & & & \\
 & \mid & \_ & & &
\end{array}
$$

Figure 3.3: Terms and values in Montuno (revisited)

The terms and values that were specified in Chapter 2 are revisited in Figure 3.3, there are a two main classes of terms: those that represent computation (functions and function application), and those that represent data (pairs, records, constants).

Data classes Most *data* terms can be represented in a straight-forward way, as they map directly to features of the host language, Kotlin in our case. Kotlin has a standard way of representing primarily data-oriented structures using `data classes`. These are classes whose primary purpose is to hold data, so-called Data Transfer Objects (DTOs), and are the recommended approach in Kotlin[4]. In Listing 3.2 we have the base classes for terms and values, and a few examples of structures that map directly from the syntax to a data object.

```
sealed class Term
sealed class Value

data class TLet(val id: String, val bind: Term, val body: Term) : Term()
data class TSigma(val id: String, val type: Term, val body: Term) : Term()
data class TPair(val left: Term, val right: Term) : Term()

data class VPair(val left: Value, val right: Value) : Value()
```

Listing 3.2: Pair and *let* − *in* representations

[4] https://kotlinlang.org/docs/idioms.html

Terms that encode computation, whether delayed (λ-abstraction) or not (application) will be more involved. Variables *can* be represented in a straight-forward way, but a string-based representations is not the most optimal way. We will look at these three constructs in turn.

3.2.1 Functions

Closure Languages, in which functions are first-class values, all use the concept of a closure. A closure is, in brief, a function in combination with the environment in it was created. The body of the function can refer to variables other than its immediate arguments, which means the surrounding environment needs to be stored as well. The simplest example is the *const* function $λx.λy.x, which, when partially applied to a single argument, e.g., let $five = const5$, needs to store the value 5 until it is eventually applied to the remaining second argument: $five15 \longrightarrow 5$.

HOAS As Kotlin supports closures on its own, it would be possible to encode λ-terms directly as functions in the host language. This is possible, and it is one of the ways of encoding functions in interpreters. This encoding is called the higher-order abstract syntax (HOAS), which means that functions[5] in the language are equal to functions in the host language. Representing functions using HOAS produces very readable code, and in some cases, e.g., on GHC produces code an order of magnitude faster than using other representations [**?**]. An example of what it looks like is in Listing 3.3.

```
data class Closure<T>(val fun: (T) -> T)

val constFive = Closure<Int> { (n) -> 5 }
```

Listing 3.3: Higher-order abstract syntax encoding of a closure

Explicit closures However, we will need to perform some operations on the AST that need explicit access to environments and the arguments of a function. The alternative to reusing functions of the host language is a *defunctionalized* representation, also called *explicit closure* representation. We will need to use this representation later, when creating the Truffle version: function calls will need to be objects, nodes in the program graph, as we will see in Chapter 4. In this encoding, demonstrated in Listing 3.4, we store the term of the function body together with the state of the environment when the closure was created.

```
data class Closure<T>(val fun: Term, val environment: Map<Name,Term>)

val constFive = Closure<Int>(TLocal("x"), mapOf("x" to 5))
```

Listing 3.4: Defunctionalized function representation

[5] In descriptions of the higher-order abstract syntax, the term *binders* is commonly used instead of function or λ-abstractions, as these constructs *bind* a value to a name.

3.2.2 Variables

Representing variables can be as straight-forward as in Listing 3.4: a variable can be a simple string containing the name of the variable. This is also what our parser produces in the pre-term representation. Also, when describing reduction rules and substitution, we have also referred to variables by their names. That is not the best way of representing variables.

Named Often, when specifying a λ-calculus, the process of substitution $t[x := e]$ is kept vague, as a concern of the meta-theory in which the λ-calculus is encoded. When using variable names (strings), the terms themselves and the code that manipulates them are easily understandable. Function application, however, requires variable renaming (α-conversion), which involves traversing the entire argument term and replacing each variable occurrence with a fresh name that does not yet occur in the function body. However, this is a very slow process, and it is not used in any real implementation of dependent types or λ-calculus.

Nameless An alternative to string-based variable representation is a *nameless* representation, which uses numbers in place of variable names [**?**]. These numbers are indices that point to the current variable environment, offsets from the top or the top of the environment stack. The numbers are assigned, informally, by *counting the lambdas*, as each λ-abstraction corresponds to one entry in the environment. The environment can be represented as a stack to which a variable is pushed with every function application, and popped when leaving a function. The numbers then point to these entries. These two approaches can be seen side-by-side in Figure 3.4.

	fix	$succ$
Named	$(\lambda f.(\lambda x.f\ (x\ x))\ (\lambda x.f\ (x\ x)))\ g$	$\lambda x.x\ (\lambda y.x\ y)$
Indices	$(\lambda(\lambda 1\ (0\ 0))\ (\lambda 1\ (0\ 0)))\ g$	$\lambda 0\ (\lambda 1\ 0)$
Levels	$(\lambda(\lambda 0\ (1\ 1))\ (\lambda 0\ (1\ 1)))\ g$	$\lambda 0\ (\lambda 0\ 1)$

Figure 3.4: Named and nameless variable representations

de Bruijn indices The first way of addressing, de Bruijn indexing, is rather well-known. It is a way of counting from the top of the stack, meaning that the argument of the innermost lambda has the lowest number. It is a "relative" way of counting, relative to the top of the stack, which is beneficial during e.g. δ-reduction in which a reference to a function is replaced by its definition: using indices, the variable references in the function body do not need to be adjusted after such substitution.

de Bruijn levels The second way is also called the "reversed de Bruijn indexing" [**?**], as it counts from the start of the stack. This means that the argument of the innermost lambda has the highest number. In the entire term, one variable is only ever addressed by one number, meaning that this is an "absolute" way of addressing, as opposed to the "relative" indices.

Locally nameless There is a third alternative that combines both named and nameless representations, and it has been used in e.g., the Lean proof assistant [**?**]. De Bruijn indices are used for bound variables and string-based names for free variables. This also avoids any need for bound variable substitution, but free variables still need to be resolved later during the evaluation of a term.

Our choice We will use a representation that has been used in recent type theory implementations [**?**] [**?**]: de Bruijn indices in terms, and de Bruijn levels in values. Such a representation avoids any need for substitution as terms are that substituted *into* an existing value do not need to have the "relative" indices adjusted based on the size of the current environment, whereas the "absolute" addressing of levels in values means that values can be directly compared. This combination of representations means that we can doing avoid any substitution at all, as any adjustment of variables is performed during the evaluation from term to value and back.

Implementation Kotlin makes it possible to construct type-safe wrappers over basic data types that are erased at runtime but that support custom operations. Representing indices and levels as `inline classes` means that we can perform add and subtract them using the natural syntax e.g. `ix + 1`, which we will use when manipulating the environment in the next section. The final representation of variables in our interpreter is in Listing 3.5.

```
inline class Ix(val it: Int) {
    operator fun plus(i: Int) = Ix(it + i)
    operator fun minus(i: Int) = Ix(it - i)
    fun toLvl(depth: Lvl) = Lvl(depth.it - it - 1)
}

inline class Lvl(val it: Int) {
    operator fun plus(i: Int) = Lvl(it + i)
    operator fun minus(i: Int) = Lvl(it - i)
    fun toIx(depth: Lvl) = Ix(depth.it - it - 1)
}

data class VLocal(val it: Lvl) : Val()
data class TLocal(val it: Ix) : Val()
```

Listing 3.5: Variable representation

3.2.3 Class structure

Variables and λ-abstractions were the two non-trivial parts of the mapping between our syntax and Kotlin values. With these two pieces, we can fill out the remaining parts of the class hierarchy. The full class listing is in Appendix C, here only a direct comparison of the data structures is shown on the *const* function in Figure 3.5, and the most important differences between them are in Figure 3.6.

```
PLam("x", Expl,    TLam("x", Expl,    VLam("x", Expl,
  PLam("y", Expl,    TLam("y", Expl,    VCl([valX], VLam("y", Expl,
    PVar("x")))        TLocal(1)))        VCl([valX, valY], VLocal(0)))))
```

Figure 3.5: Direct comparison of `PreTerm`, `Term`, and `Value` objects

	Variables	Functions	Properties
`PreTerm`	String names	`PreTerm` AST	well-formed
`Term`	de Bruijn index	`Term` AST	well-typed
`Value`	de Bruijn level	Closure (`Term` AST + `Values` in context)	normal form

Figure 3.6: Important distinctions between `PreTerm`, `Term`, and `Value` objects

3.3 Normalization

3.3.1 Approach

Normalization-by-evaluation Normalization is a series of βδζι-reductions, as defined in Chapter 2. While there are systems that implement normalization as an exact series of reduction rules, it is an inefficient approach that is not common in the internals of state-of-the-art proof assistants. An alternative way of bringing terms to normal form is the so-called *normalization-by-evaluation* (NbE) [**?**]. The main principle of this technique is interpretation from the syntactic domain of terms into a computational, semantic domain of values and back. In brief, we look at terms as an executable program that can be *evaluated*, the result of such evaluation is then a normal form of the original term. NbE is total and provably confluent [**?**] for any abstract machine or computational domain.

Neutral values If we consider only closed terms that reduce to a single constant, we could simply define an evaluation algorithm over the terms defined in the previous chapter. However, normalization-by-evaluation is an algorithm to bring any term into a full normal form, which means evaluating terms inside function bodies and constructors. NbE introduces the concept of "stuck" values that cannot be reduced further. In particular, free variables in a term cannot be reduced, and any terms applied to a stuck variable cannot be further reduced and are "stuck" as well. These stuck values are called *neutral values*, as they are inert with regards to the evaluation algorithm.

Semantic domain Proof assistants use abstract machines like Zinc or STG; any way to evaluate a term into a final value is viable. This is also the reason to use Truffle, as we can translate a term into an executable program graph, which Truffle will later optimize as necessary. In this first interpreter, however, the computational domain will be a simple tree-traversal algorithm.

The set of neutral values in Montuno is rather small (Figure 3.7): an unknown variable, function application with a neutral *head* and arbitrary terms in the *spine*, and a projection eliminator.

$$neutral \quad := \quad var \quad | \quad neutral\, a_1 \,...a_n \quad | \quad neutral.l_n$$

Figure 3.7: Neutral values

Specification The NbE algorithm is fully formally specifiable using four operations: the above-mentioned evaluation and quoting, reflection of a neutral value (*NeVal*) into a value, and reification of a value into a normal value (*NfVal*) that includes its type, schematically shown in Figure 3.8. In this thesis, though, will only describe the relevant parts of the specification in words, and say that NbE (as we will implement it) is a pair of functions $nf = quote(eval(term))$,

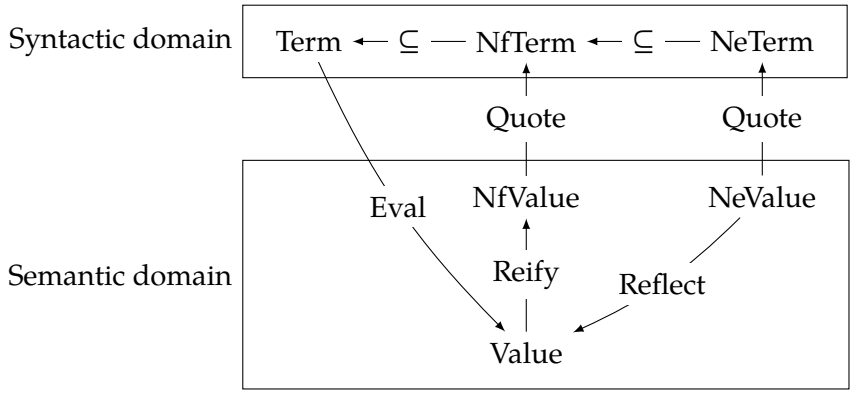

Figure 3.8: Syntactic and semantic domains in NbE [**?**]

3.3.2 Normalization strategies

Normalization-by-evaluation is, however, at its core inefficient for our purposes [**?**]. The primary reason to normalize terms in the interpreter is for type-checking and inference and that, in particular, needs normalized terms to check whether two terms are equivalent. NbE is an algorithm to get a full normal form of a term, whereas to compare values for equality, we only need the weak head-normal form. To illustrate: to compare whether a λ-term and a pair are equal, we do not need to compare two fully-evaluated values, but only to find out whether the term is a pair of a λ-term, which is given by the outermost constructor, the *head*.

In Chapter 2 we saw an overview of normal forms of λ-calculus. To briefly recapitulate, a normal form is a fully evaluated term with all sub-terms also fully evaluated. A weak head-normal form is a form where only the outermost construction is fully evaluated, be it a λ-abstraction or application of a variable to a spine of arguments.

Reduction strategy Normal forms are associated with a reduction strategy, a set of small-step reduction rules that specify the order in which sub-expressions are reduced. Each strategy brings an expression to their corresponding normal form. Common ones are *applicative order* in which we first reduce sub-expressions left-to-right, and then apply functions to them; and *normal order* in which we first apply the leftmost function, and only then reduce its arguments. In Figure 3.9 there are two reduction strategies that we will emulate.

$$x \xrightarrow{name} x$$

$$\frac{}{(\lambda x.e) \xrightarrow{name} (\lambda x.e)}$$

$$\frac{e_1 \xrightarrow{name} (\lambda x.e) \qquad e[x := e_2] \xrightarrow{name} e'}{(e_1\ e_2) \xrightarrow{name} e'}$$

$$\frac{e_1 \xrightarrow{name} e_1' \not\equiv \lambda x.e}{(e_1\ e_2) \xrightarrow{name} (e_1'\ e_2)}$$

(a) Call-by-name to weak head normal form

$$x \xrightarrow{norm} x$$

$$\frac{e \xrightarrow{norm} e'}{(\lambda x.e) \xrightarrow{norm} (\lambda x.e')}$$

$$\frac{e_1 \xrightarrow{name} (\lambda x.e) \qquad e[x := e_2] \xrightarrow{norm} e'}{(e_1\ e_2) \xrightarrow{norm} e'}$$

$$\frac{e_1 \xrightarrow{name} e_1' \not\equiv \lambda x.e \qquad e_1' \xrightarrow{norm} e_1'' \qquad e_2 \xrightarrow{norm} e_2'}{(e_1\ e_2) \xrightarrow{norm} (e_1''\ e_2)}$$

(b) Normal order to normal form

Figure 3.9: Reduction strategies for λ-calculus [**?**]

In general programming language theory, a concept closely related to reduction strategies is an evaluation strategy. These also specify when an expression is evaluated into a value, but in our case, they apply to our host language Kotlin.

Call-by-value Call-by-value, otherwise called eager evaluation, corresponds to applicative order reduction strategy [**?**]. Specifically, when executing a statement, its sub-expressions are evaluated inside-out and immediately reduced to a value. This leads to predictable program performance (the program will execute in the order that the programmer wrote it, evaluating all expressions in order), but this may lead to unnecessary computations performed: given an expression `const 5 (ackermann 4 2)`, the value of `ackermann 4 2` will be computed but immediately discarded, in effect wasting processor time.

Call-by-need Call-by-need, also lazy evaluation, is the opposite paradigm. An expression will be evaluated only when its result is first accessed, not when it is created or defined. Using call-by-need, the previous example will terminate immediately as the calculation `ackermann 4 2` will be deferred and then discarded. However, it also has some drawbacks, as the performance characteristics of programs may be less predictable or harder to debug.

Call-by-value is the prevailing paradigm, used in all commonly used languages with the exception of Haskell. It is sometimes necessary to defer the evaluation of an expression, however, and in such cases lazy evaluation is emulated using closures or zero-argument functions: e.g., in Kotlin a variable can be initialized using the syntax `val x by lazy { ackermann(4, 2) }`, and the value will only be evaluated if it is ever needed.

Call-by-push-value There is also an alternative paradigm, called call-by-push-value, which subsumes both call-by-need and call-by-value as they can be directly translated to CBPV–in the context of λ-calculus specifically. It defines additional operators *delay* and *force* to accomplish this, one to create a *thunk* that contains a deferred computation, one to evaluate

the thunk. Also notable is that it distinguishes between values and computations: values can be passed around, but computations can only be executed, or deferred.

Emulation We can emulate normalization strategies by implementing the full normalization-by-evaluation algorithm, and varying the evaluation strategy. Kotlin is by default a call-by-value language, though, and evaluation strategy is an intrinsic property of a language so, in our case, this means that we need to insert `lazy` annotations in the correct places, so that no values are evaluated other than those that are actually used. In the case of the later Truffle implementation, we will need to implement explicit *delay* and *force* operations of call-by-push-value, which is why we introduced all three paradigms in one place.

3.3.3 Implementation

The basic outline of the implementation is based on Christiansen's [**?**]. In essence, it implements the obvious evaluation algorithm: evaluating a function captures the current environment in a closure, evaluating a variable looks up its value in the environment, and function application inserts the argument into the environment and evaluates the body of the function.

Environments The brief algorithm description used a concept we have not yet translated into Kotlin: the environment, or evaluation context. When presenting the $\lambda\!\rightarrow$-calculus, we have seen the typing context Γ, to which we add a value context.

$$\Gamma \quad := \quad \bullet \quad | \quad \Gamma, x : t$$

The environment, following the above definition, is a stack: defining a variable pushes a pair of a name and a type to the top, which is then popped off when the variable goes out of scope. An entry is pushed and popped whenever we enter and leave a function context, and the entire environment needs to be captured in its current state whenever we create a closure. When implementing closures in Truffle, we will also need to take care about which variables are actually used in a function. That way, we can capture only those that need to be captured and not the entire environment.

Linked list The natural translation of the environment definition is a linked list. It would also be the most efficient implementation in a functional language like Haskell, as appending to an immutable list is very cheap there. In Kotlin, however, we need to take care about not allocating too many objects and will need to consider mutable implementations as well.

Mutable/immutable In Kotlin and other JVM-based languages, an `ArrayDeque` is a fast data structure, a mutable implementation of the stack data structure. In general, array-backed data structures are faster than recursive ones on the JVM, which we will use in the Truffle implementation. In this first interpreter, however, we can use the easier-to-use immutable linked list implementation. It is shown in Listing 3.6, a linked list specialized for values; an equivalent structure is also implemented for types.

```
data class VEnv(val value: Val, val next: VEnv?)

fun VEnv?.len(): Int = if (this == null) 0 else 1 + next.len()
operator fun VEnv?.plus(v: Val): VEnv = VEnv(v, this)
operator fun VEnv?.get(n: Ix): Val
   = if (n.it == 0) this!!.value else this!!.next[n - 1]
```

Listing 3.6: Environment data structure as an immutable linked list

Environment operations We need three operations from an environment data structure: insert (bind) a value, look up a bound value by its level or index, and unbind a variable that leaves the scope. In Listing 3.6, we see two of them: the operator `plus`, used as `env + value`, binds a value, and operator `get`, used as `env[ix]`, looks a value up. Unbinding a value is implicit, because this is an immutable linked list: the reference to the list used in the outer scope is not changed by any operations in the inner scope. These operations are demonstrated in Listing 3.7, on the `eval` operations of a variable and a *let* − *in* binding.

There we also see the basic structure of the evaluation algorithm. Careful placement of `lazy` has been omitted, as it splits the algorithm into two: parts that need to be evaluated lazily and those that do not, but the basic structure should be apparent. The snippet uses the Kotlin `when-is` construct, which checks the class of the argument, in this case we check if `this` is a `TLocal`, `TLet`, etc.

```
fun eval(ctx: Context, term: Term, env: VEnv): Val = when (term) {
  is TLocal ->
    env[term.ix] ?: VNeutral(HLocal(Lvl(ctx.lvl - term.ix - 1), spineNil))
  is TLet -> eval(ctx, term.body, env + eval(ctx, term.defn, env))
  is TLam -> VLam(term.name, VCl(env, term.body))
  is TApp -> when (fn := eval(ctx, term.lhs, env)) {
    is VLam -> eval(ctx, fn.cl.term, fn.cl.env + eval(ctx, term.rhs, env))
    is VNeutral -> VNeutral(fn.head, fn.spine + term.right)
  }
  // ...
}
```

Listing 3.7: Demonstration of the `eval` algorithm

Eval In Listing 3.7, a variable is looked up in the environment, and considered a neutral value if the index is bigger than the size of the current environment. In `TLet` we see how an environment is extended with a local value. A λ-abstraction is converted into a closure. Function application, if the left-hand side is a `VLam`, evaluates the body of this closure, and if the left-hand side is a neutral expression, then the result is also neutral value and its spine is extended with another argument. Other language constructs are handled in a similar way,

Quote In Listing 3.8, we see the second part of the algorithm. In the domain of values, we do not have plain variable terms, or *let* − *in* bindings, but unevaluated functions and "stuck" neutral terms. A λ-abstraction, in order to be in normal form, needs to have its body also in normal form, therefore we insert a neutral variable into the environment in place of the argument, and eval/quote the body. A neutral term, on the other hand, has

at its head a neutral variable. This variable is converted into a term-level variable, and the spine reconstructed as a tree of nested `TApp` applications.

```
fun quote(ctx: Context, v: Val): Term = when (v) {
  is VNeutral -> {
    x = TLocal(Ix(ctx.depth - v.head - 1))
    for (vSpine in v.spine.reversed()) {
        x = TApp(x, quote(ctx, vSpine))
    }
    x
  }
  is VLam -> TLam(v.name,
      quote(ctx, eval(ctx, v.cl.body, v.cl.env + VNeutral(HLocal(ctx.lvl)))))
  // ...
}
```

Listing 3.8: Demonstration of the `quote` algorithm

These two operations work together, to fully quote a value, we need to also lazily `eval` its sub-terms. The main innovation of the normalization-by-evaluation approach is the introduction of neutral terms, which have the role of a placeholder value in place of a value that has not yet been supplied. As a result, the expression $quote(eval(term, emptyEnv))$ produces a lazily evaluated normal form of a term in a weak head-normal form, with its sub-terms being evaluated whenever accessed. Printing out such a term would print out the fully normalized normal form.

Primitive operations Built-in language constructs like *Nat* or *false* that have not been shown in the snippet are mostly inserted into the initial context as values that can be looked up by their name. In general, though, constructs with separate syntax, e.g. Σ-types, consist of three parts:

- their type is bound in the initial context;
- the term constructor is added to the set of terms and values, and added in `eval()`;
- the eliminator is added as a term and as a spine constructor, i.e., an operation to be applied whenever the neutral value is provided.

The full listing is provided in the supplementary source code, as it is too long to be included in text.

3.4 Elaboration

3.4.1 Approach

The second part of the internals of the compiler is type elaboration. Elaboration is the transformation of a partially-specified, well-formed program submitted by a user into a fully-specified, well-typed internal representation [**?**]. In particular, we will use elaboration to infer types of untyped Curry-style λ-terms, and to infer implicit function arguments that were not provided by the user, demonstrated in Figure 3.10.

function signature:	$id : \{A\} \to A \to A$
provided expression:	$id\ id\ 5$
elaborated expression:	$(id\ \{Nat \to Nat\}\ id)\ \{Nat\}\ 5$

Figure 3.10: Demonstration of type elaboration

Bidirectional typing Programmers familiar with statically-typed languages like Java are familiar with type checking, in which all types are provided by the user, and therefore are inputs to the type judgment $\Gamma \vdash e : t$. Omitting parts of the type specification means that the type system not only needs to check the types for correctness, but also infer (synthesize) types: the type t in $\Gamma \vdash e : t$ is produced as an output. In some systems, it is possible to omit all type annotations and rely only on the type constraints of built-in functions and literals. Bidirectional systems that combine both input and output modes of type judgment are now a standard approach [**?**], often used in combination with constraint solving.

Judgments The type system is composed of two additional type judgments we haven't seen yet, that describe the two directions of computation in the type system:

- $\Gamma \vdash e \Rightarrow t$ is "given the context Γ and term e, infer (synthesize) its type t", and
- $\Gamma \vdash e \Leftarrow t$ is "given the context Γ, term e and type t, check that t is a valid type for t".

The entire typing system described in Chapter 2 can be rewritten using these type judgments. The main principle is that language syntax is divided into two sets of constructs: those that constrain the type of a term and can be checked against an inferred term, and those that do not constrain the type and need to infer it entirely.

$$\frac{a : t \in \Gamma}{\Gamma \vdash a \Rightarrow t}\ (\textbf{Var}) \qquad \frac{c \text{ is a constant of type } t}{\Gamma \vdash c \Rightarrow t}\ (\textbf{Const})$$

$$\frac{\Gamma, x : t \vdash e \Leftarrow u}{\Gamma \vdash \lambda x.e \Leftarrow t \to u}\ (\textbf{Abs}) \qquad \frac{\Gamma \vdash f \Rightarrow t \to u \qquad \Gamma \vdash a \Rightarrow t}{\Gamma \vdash f\ a \Rightarrow u}\ (\textbf{App})$$

$$\frac{\Gamma \vdash a \Rightarrow t \qquad \Gamma \vdash a = b}{\Gamma \vdash a \Leftarrow b}\ (\textbf{ChangeDir}) \qquad \frac{\Gamma \vdash a \Leftarrow t}{\Gamma \vdash (a : t) \Rightarrow t}\ (\textbf{Ann})$$

Figure 3.11: Bidirectional typing rules for the $\lambda\!\to$-calculus

Bidirectional $\lambda\!\to$ typing In Figure 3.11, this principle is demonstrated on the simply-typed λ-calculus with only variables, λ-abstractions and function application. The first four rules correspond to rules that we have introduced in Chapter 2, with the exception of the constant rule that we have not used there. The two new rules are (**ChangeDir**) and (**Ann**): (**ChangeDir**) says that if we know that a term has an already inferred type, then we can satisfy any rule that requires that the term checks against a type equivalent to this one. (**Ann**) says that to synthesize the type of an annotated term $a : t$, the term first needs to check against that type.

Rules (**Var**) and (**Const**) produce an assumption, if a term is already in the context or a constant, then we can synthesize its type. In rule (**App**), if we have a function with an

inferred type then we check the type of its argument, and if it holds then we can synthesize the type of the application $f\ a$. To check the type of a function in rule (**Abs**), we first need to check whether the body of a function checks against the type on the right-hand side of the arrow.

While slightly complicated to explain, this description produces a provably sound and complete type-checking system [**?**] that, as a side effect, synthesizes any types that have not been supplied by the user. Extending this system with other language constructs is not complex: the rules used in Montuno for local and global definitions are in Figure 3.12.

$$\frac{\Gamma \vdash t \Leftarrow \star \qquad \Gamma \vdash a \Leftarrow t \qquad \Gamma, x : t \vdash b \Rightarrow u}{\Gamma \vdash \text{let } x : t = a \text{ in } b \Rightarrow u}\ (\textbf{Let-In})$$

$$\frac{\Gamma \vdash t \Leftarrow \star \qquad \Gamma \vdash a \Leftarrow t}{\Gamma \vdash x : t = a \Rightarrow t}\ (\textbf{Defn})$$

Figure 3.12: Bidirectional typing rules for *let* − *in* and top-level definitions

Meta-context One concern was not mentioned in the previous description: when inferring a type, we may not know all its component types: in rule (**Abs**), the type of the function we check may only be constrained by the way it is called. Implicit function arguments $\{A\ B\} \to A \to B \to A$ also only become specific when the function is actually called. The solution to this problem is a *meta-context* that contains *meta-variables*.

These stand for yet undetermined terms [**?**], either as placeholders to be filled in by the user in interactive proof assistants (written with a question mark, e.g. as $?\alpha$), or terms that can be inferred from other typing constraints using unification. These meta-variables can be either inserted directly by the user in the form of a hole "_", or implicitly, when inferring the type of a λ-abstraction or an implicit function argument [**?**].

There are several ways of implementing this context depending on the scope of meta-variables, or whether it should be ordered or the order of meta-variables does not matter. A simple-to-implement but sufficiently useful for our purposes is a globally-scoped meta-context divided into blocks placed between top-level definitions.

```
id : {A} → A → A = λx.x
?α = Nat
?β = ?α → ?α
five = (id ?β id) ?α 5
```

Listing 3.9: Meta-context for the expression `id id 5`

The meta-context implemented in Montuno is demonstrated in Listing 3.9. When processing a file, we process top-level expressions sequentially. The definition of the *id* function is processed, and in the course of processing *five*, we encounter two implicit arguments, which are inserted on the top-level as the meta-variables $?\alpha$ and $?\beta$.

3.4.2 Unification

Returning to the rule (**ChangeDir**) in Figure 3.12, a critical piece of the algorithm is the equality of two types that this rule uses. To check a term against a type $\Gamma \vdash a \Leftarrow t$, we first infer a type for the term $\Gamma \vdash a \Rightarrow u$, and then test its equivalence to the wanted type $t = u$.

Conversion checking The usual notion of equivalence in λ-calculus is *α-equivalence of β-normal forms*, that we discussed in Chapter 2, which corresponds to structural equality of the two terms. *Conversion checking* is the algorithm that determines if two terms are convertible using a set of conversion rules.

Unification As we also use meta-variables in the type elaboration process, these variables need to be solved in some way. This process of conversion checking together with solving meta-variables is called *unification* [**?**], and is a well-studied problem in the field of type theory.

Pattern unification In general, solving meta-variables is undecidable [**?**]. Given the constraint $?\alpha\ 5 = 5$, we can produce two solutions: $?\alpha = \lambda x.x$ and $?\alpha = \lambda x.5$. There are several possible approaches and heuristics: first-order unification solves for base types and cannot produce functions as a result; higher-order unification can produce functions but is undecidable; *pattern unification* is a middle ground and can produce functions as solutions, with some restrictions.

Renaming In this thesis, I have chosen to reuse the algorithm from [**?**] which, in brief, assumes that a meta-variable is a function whose arguments are all local variables in scope at the moment of its creation. Then, when unifying the meta-variable with another (non-variable) term, it builds up a list of variables the term uses, and stores such a solution as a *renaming* that maps the arguments to a meta-variable to the variables with which it was unified. As the algorithm is rather involved but tangential to the goals of this thesis, I will omit a detailed description and instead point an interested reader at the original source [**?**].

3.4.3 Implementation

As with the implementation of normalization-by-evaluation, we will look at the most illustrative parts of the implementation. This time, the comparison can be made directly side-by-side, between the bidirectional typing algorithm and its implementation.

What was not mentioned explicitly is that the type elaboration algorithm has `PreTerm`s as its input, and produces `Term`s in the case of type checking, and pairs of `Term`s and `Value`s (the corresponding types) in the case of type inference. Unification, not demonstrated here, is implemented as parallel structural recursion over two `Value` objects.

In Figure 3.13, we see the previously described rule that connects the checking and synthesis parts of the algorithm and uses unification. Unification solves meta-variables as a side-effect, here it is only in the role of a guard as it does not produce a value. The code exactly

follows the typing rule, the pre-term is inferred, resulting in a pair of a well-typed term and its type, the type is unified with the wanted type and, if successful, the produced term is the return value.

$$\frac{\Gamma \vdash a \Rightarrow t \qquad \Gamma \vdash a = b}{\Gamma \vdash a \Leftarrow b}\ \textbf{(ChangeDir)}$$

```
fun LocalContext.check(pre: PreTerm, wanted: Value): Term = when (pre) {
  // ...
  else -> {
    val (t, actual) = infer(pre.term)
    unify(actual, wanted)
    t
  }
}
```

Figure 3.13: Side-by-side comparison of the **ChangeDir** rule

Figure 3.14 shows the exact correspondence between the rule and its implementation, one read left-to-right, the other top-to-bottom. Checking of the type and value are straightforward, translation of $\Gamma, x : t \vdash b \Rightarrow u$ binds a local variable in the environment, so that the body of the *let* − *in* expression can be inferred, and the result is a term containing the inferred body and type, wrapped in a `TLet`.

$$\frac{\Gamma \vdash t \Leftarrow \star \qquad \Gamma \vdash a \Leftarrow t \qquad \Gamma, x : t \vdash b \Rightarrow u}{\Gamma \vdash \text{let } x : t = a \text{ in } b \Rightarrow u}\ \textbf{(Let-In)}$$

```
fun LocalContext.infer(pre: PreTerm): Pair<Term, Value> = when (pre)
  is RLet -> {
    val t = check(pre.type, VStar)
    val a = check(pre.defn, t)
    val (b, u) = localDefine(pre.name, a, t).infer(pre.body)
    Pair(TLet(pre.name, t, a, b), u)
  }
  // ...
}
```

Figure 3.14: Side-by-side comparison of the **Let-in** rule

Lastly, the rule for a term-level λ-abstraction is demonstrated in Figure 3.15. The type produced on the last line of the snippet is a `VPi` unlike the rule, as the rule was written for the λ→-calculus; it is semantically equivalent, however. This rule demonstrates the creation of a new meta-variable as without a placeholder, we are not able to infer the type of the body of the function. This meta-variable might or might not be solved in the course of inferring the body: either way, both the term and the type only contain a reference to a globally-scoped meta-variable and not the solution.

3.5 Driver

This concludes the complex part of the interpreter, what follows are rather routine concerns. Next part of the implementation is the driver that wraps the backend, and handles

$$\frac{\Gamma, x : t \vdash e \Leftarrow u}{\Gamma \vdash \lambda x.e \Leftarrow t \to u}\ (\textbf{Abs})$$

```
fun LocalContext.infer(pre: PreTerm): Pair<Term, Value> = when (pre)
  is RLam -> {
    val a = newMeta()
    val (b, t) = localBind(pre.name, a).infer(pre.body)
    Pair(TLam(pre.name, b), VPi(pre.name, a, VCl(env, t.quote())))
  }
  // ...
}
```

Figure 3.15: Side-by-side comparison of the **Abs** rule

its interaction with the surrounding world. In particular, the parser, pretty-printer, and state management.

Parser Lexical and syntactic analysis is not the focus of this work, so simply I chose the most prevalent parsing library in Java-based languages, which seems to be ANTLR[6]. It comes with a large library of languages and protocols from which to take inspiration[7], so creating the parser was a rather simple matter. ANTLR provides two recommended ways of consuming the result of parsing using classical object-oriented design patterns: a listener and a visitor. I used neither as they were needlessly verbose or limiting[8].

Instead of these, a custom recursive-descent AST transformation was used that is demonstrated in Listing 3.10. This directly transforms the `ParseContext` objects created by ANTLR into our `PreTerm` data type.

```
fun TermContext.toAst(): PreTerm = when (this) {
  is LetContext -> RLet(id.toAst(), type.toAst(), defn.toAst(), body.toAst())
  is LamContext -> rands.foldRight(body.toAst()) { l, r -> RLam(l.toAst(), r) }
  is PiContext -> spine.foldRight(body.toAst()) { l, r -> l.toAst()(r) }
  is AppContext -> operands.fold(oprator.toAst()) { l, r -> r.toAst()(l) }
  else -> throw UnsupportedOperationException(javaClass.canonicalName)
}
```

Listing 3.10: Parser to `PreTerm` transformation as a depth-first traversal

The data type itself is shown in Listing 3.11. As with terms and values, it is a recursive data structure, presented here in a slightly simplified manner compared to the actual implementation, as it omits the part that tracks the position of a term in the original source. The grammar that is used as the source for the parser generator ANTLR was already presented once in the conclusion of Chapter 2, so the full listing is only included in Appendix B.

Pretty-printer A so-called pretty-printer is a transformation from an internal representation of a data structure to a user-readable string representation. The implementation of

[6] https://www.antlr.org/

[7] https://github.com/antlr/grammars-v4/

[8] In particular, ANTLR-provided visitors require that all return values share a common super-class. Listeners don't allow return values and would require explicit parse tree manipulation.

```
sealed class PreTerm
typealias Pre = PreTerm
typealias N = String

sealed class TopLevel
data class RDecl(val n: N, val type: Pre) : TopLevel()
data class RDefn(val n: N, val type: Pre?, val term: Pre) : TopLevel()
data class RTerm(val cmd: Command, val term: Pre) : TopLevel()

object RU : Pre()
object RHole : Pre()
data class RVar(val n: N) : Pre()
data class RNat(val n: Int) : Pre()
data class RApp(val lhs: Pre, val rhs: Pre) : Pre()
data class RLam(val n: N, val body: Pre) : Pre()
data class RPi(val n: N, val type: Pre, val body: Pre) : Pre()
data class RLet(val n: N, val type: Pre, val defn: Pre, val body: Pre) : Pre()
data class RForeign(val lang: N, val eval: N, val type: Pre) : Pre()
```

Listing 3.11: Data type `PreTerm`

such a transformation is mostly straight-forward, complicated only by the need to correctly handle operator precedence and therefore parentheses.

This part is implemented using the Kotlin library `kotlin-pretty`, which is itself inspired by the Haskell library `prettyprinter` which, among other things, handles correct block indentation and ANSI text coloring: that functionality is also used in error reporting in the terminal interface.

An excerpt from this part of the implementation is included in Listing 3.12, which demonstrates the precedence enumeration `Prec`, the optionally parenthesizing operation `par`, and other constructions of the `kotlin-pretty` library.

```
enum class Prec { Atom, App, Pi, Let }
fun Term.pretty(ns: NameEnv?, p: Prec = Prec.Atom): Doc<Nothing> = when (this) {
  is TVar -> ns[ix].text()
  is TApp -> par(p, Prec.App,
    arg.pretty(ns, Prec.App) spaced body.pretty(ns, Prec.Atom))
  is TLet -> {
    val d = listOf(
      ":".text() spaced ty.pretty(ns, Prec.Let),
      "=".text() spaced bind.pretty(ns, Prec.Let),
    ).vCat().align()
    val r = listOf(
      "let $n".text() spaced d,
      "in".text() spaced body.pretty(ns + n, Prec.Let)
    ).vCat().align()
    par(p, Prec.Let, r)
  } // ...
}
```

Listing 3.12: Pretty-printer written using `kotlin-pretty`

State management Last component of the driver code is global interpreter state, which consists mainly of a table of global names, which is required for handling incremental interpretation or suggestions (tab-completion) in the interactive environment. It also tracks the position of the currently evaluated term in the original source file for error reporting.

Overall, the driver receives user input in the form of a string, parses it, expression by expression supplies it to the backend, receiving back a global name, or an evaluated value, which it pretty-prints and returns back to the user-facing frontend code.

3.6 Frontend

We will consider only two forms of user interaction: batch processing of a file via a command-line interface, and a terminal environment for interactive use. Later, with the Truffle interpreter, we can also add an option to compile a source file into an executable using Truffle's capability to produce *Native Images*.

```
> cat demo.mt
id : {A} -> A -> A = \x. x
{-# TYPE id #-}
{-# ELABORATE id 5 #-}
{-# NORMALIZE id 5 #-}

> montuno demo.mt
{A} -> A -> A
id {Nat} 5
5
> montuno --type id
{A} -> A -> A
```

Listing 3.13: Example usage of the CLI interface

CLI We will reuse the entry point of Truffle languages, a `Launcher` class, so that integration of the Truffle interpreter is easier later, and so that we are able to create single executable that is able to use both interpreters.

`Launcher` handles pre-processing command-line arguments for us, a feature for which we would otherwise use an external library like `JCommander`. In the Truffle interpreter, we will also use the *execution context* it prepares using various JVM options but for now, we will only use `Launcher` for argument processing.

Two modes of execution are implemented, one mode that processes a single expression provided on the command line and `--normalize`s it, `--elaborate`s it, or find its `--type`. The second mode is sequential batch processing mode that reads source code either from a file or from standard input, and processes all statements and commands in it sequentially.

As we need to interact with the user we encounter another problem, that of error reporting. It has been mentioned in passing several times, and in this implementation of the interpreter, it is handled only partially. To report an error well, we need its cause and location. Did the user forget to close a parenthesis, or is there a type error and what can they do

to fix it? Syntactic errors are reported well in this interpreter, but elaboration errors only sometimes.

Error tracking pervades the entire interpreter, position records are stored in all data structures, location of the current expression is tracked in all evaluation and elaboration contexts, and requires careful placement of update commands and throwing and catching of exceptions. As error handling is implemented only passably and is not the focus of this thesis, it is only mentioned briefly here.

In Listing 3.13, a demonstration of the command-line interface is provided: normalization of an expression, batch processing of a file, and finally, starting up of the REPL.

REPL Read-Eval-Print Loop is the standard way of implementing interactive terminal interfaces to programming languages. The interpreter receives a string input, processes it, and writes out the result. There are other concerns, e.g., implementing name completion, different REPL-specific commands or, in our case, switching the backend of the REPL at runtime.

From my research, JLine is the library of choice for interactive command-line applications in Java, so that is what I used. Its usage is simple, and implementing a basic interface takes only 10s of lines. The commands reflect the capabilities of the command-line interface: (re)loading a file, printing out an expression in normalized or fully elaborated forms, and printing out the type of an expression. These are demonstrated in a simple way in Listing 3.14.

```
> montuno
Mt> :load demo.mt
Mt> <TAB><TAB>
Nat Bool zero succ true false if natElim id const
Mt> :normalize id 5
5
Mt> :elaborate id 5
id {Nat} 5
Mt> :type id
{A} -> A -> A
Mt> :quit
```

Listing 3.14: REPL session example