Chapter 1

Simplifying Parsers and Expressions

Writing domain-specific lint rules unlocks the potential for more powerful and interesting transformations utilising specialised domain knowledge. Desirable: * inspectability for analysis (that's what we're here for!) and optimisation The purpose of this chapter is to describe the intermediate representations of parsers (§1.1) and functions (§1.2). Show that terms must be simplified to a normal form Demonstrate equivalence to dsl optimisations in staged metaprogramming

1.1   Representing and Simplifying Parsers

This is an INTERMEDIATE SYMBOLIC REPRESENTATION (?) more specialised than general-purpose scala ast This section is about simplifying in our semantic domain (parsers)

Scalafix runs at the meta-level, outside of the phase distinction of compile- and run-time. Staged metaprogramming applies optimisations at compile-time, whereas these "optimisations" at applied post-compilation

For example, given two AST nodes `Term.Name("p")` and `Term.Name("q")` corresponding to named parsers p and q, suppose a transformation involves combining them with the *ap* combinator `<*>`. One may consider using quasiquotes to achieve this: `q"p <*> q"` would automatically expand to `Term.ApplyInfix(Term.Name("p")`, `Term.Name("<*>"), Type.ArgClause(Nil), Term.ArgClause(List(Term.Name("q")), None))`. However, this loses the static inspectability of the individual parsers p and q – although quasiquotes can be used as extractor patterns to recover the original AST nodes, their usage as such is discouraged as they can easily result in unintended match errors. The recommended approach is to pattern match on the AST nodes directly, which is obviously unergonomic even for this small example: to extract the RHS term q, one would have to perform a nested pattern match on the `Term.ApplyInfix` term and its `Term.ArgClause` node representing the arguments of the infix function application.

It is hopefully obvious that this would a very painful process for the rule author. It would be desirable to abstract away from the low-level syntactic AST representation, and instead treat these AST nodes as what they semantically represent – parsers.

Instead, fig. 1.1 shows how parser terms can be represented as an algebraic data type ADT, in the same way `parsley` itself uses a deep embedding to represent parsers as pure data objects. The reasoning behind this approach is the same as that for PARSLEY – this representation allows parsers to be easily inspected and analysed via pattern matching on constructors.

```scala
trait Parser
case class NonTerminal(ref: Symbol) extends Parser
case class Pure(f: Function) extends Parser
case object Empty extends Parser
case class Choice(p: Parser, q: Parser) extends Parser
case class Ap(p: Parser, q: Parser) extends Parser
...
```

Fig. 1.1: A subset of the core combinators in the `Parser` ADT.

This section explores the motivation behind this and the design choices made in the implementation. Use the left-recursion factoring (**??**) rule as a basis/context to demonstrate the utility of this representation.

Running example   The left-recursion factoring rule (??) performs the most complex analyses and transformations on parsers in parsley-garnish. Thus, it is a good example to motivate the design requirements for the parser representation. The following left-recursive parser and its transformation into its postfix form will serve as a running example for this section:

```scala
lazy val expr: Parsley[String] = (expr, string("a")).zipped(_ + _) | string("b")
```

1.1.1 Detecting Named Parsers

Before any analysis on parsers can be performed, it is first necessary to identify which AST nodes correspond to parsers. parsley-garnish builds a map of all parsers defined within a source file, indexed by the unique symbol of its name.

Identifying these AST nodes of interest involves pattern matching on val, var, and def definitions with a type inferred to be some Parsley[_] – this information is accessed by querying the Scalafix semantic API for the node's symbol information. Consider the labelled ast structure of the expr parser:

```scala
Defn.Val(
  mods = List(Mod.Lazy()),
  pats = List(Pat.Var(Term.Name("expr"))),
  decltpe = Some(
    Type.Apply(Type.Name("Parsley"), Type.ArgClause(List(Type.Name("String"))))
  ),
  rhs = Term.ApplyInfix(...)
)
```

The qualified symbol expr is used as the key in the map, and the rhs term is lifted the intermediate parser representation for analysis. A reference to the original AST node is also kept so any lint diagnostics or code rewrites can be applied to the correct location in the source file. Thus, a full traversal through the source file builds a map of all named parsers, representing all non-terminals in the grammar defined within that file.

1.1.2 Converting Scalameta Terms to the Parser ADT

Having identified the AST nodes which represent parsers, they need to be transformed into the appropriate Parser representation. This involves pattern matching on the scala.meta.Term to determine which parser combinator it represents, and then constructing the appropriate Parser instance.

Each Parser defines a partial function, fromTerm, which creates an instance of that parser from the appropriate scala.meta.Term. These fromTerm methods are combined to define a toParser extension method on scala.meta.Term – this is where AST nodes are lifted to their corresponding Parser representation.

The top-level combinator that makes up expr's definition is the choice combinator, |. Scalameta represents this infix application of the | operator as so:

```scala
Term.ApplyInfix(
  lhs = Term.Apply(...), // AST node for (expr, string("a")).zipped(_ + _)
  op = Term.Name("|"),
  targClause = Type.ArgClause(List()),
  argClause = Term.ArgClause(
    List(
      Term.Apply(
        Term.Name("string"),
        Term.ArgClause(List(Lit.String("b")), None)
```

```
      )
    ),
    None
  )
)
```

This structure therefore guides the implementation of the pattern match in `Choice.fromTerm`:

```scala
object Choice {
  val matcher = SymbolMatcher.normalized("parsley.Parsley.`|`", "parsley.Parsley.`<|>`")

  def fromTerm(implicit doc: SemanticDocument): PartialFunction[Term, Choice] = {
    case Term.ApplyInfix(p, matcher(_), _, Term.ArgClause(List(q), _)) =>
      Choice(p.toParser, q.toParser)
  }
}
```

The definition of this method is fairly self-explanatory: it matches on a `ApplyInfix` term where the operator is the | combinator, and recursively applies `toParser` to its LHS and RHS nodes. Finishing off, the `expr` parser is therefore converted to the following `Parser` instance:

```scala
Choice(
  Zipped(Function(_ + _), List(NonTerminal(expr), Str(a))),
  Str(b)
)
```

The exact representation of the `Function` is not important at this momenet – this is covered in the next section. For brevity, the remaining code snippets in this section will simplify the function representations and continue to grey them out.

1.1.3   Building New Parsers From Existing Parsers

Now that raw AST terms can be lifted to the higher-level parser representation, it is easy to build new parsers from existing parsers. This is crucial for left-recursion factoring, which "unfolds" parsers into separate parsers representing the left-recursive and non-left-recursive parts. These are then recombined to form parsers which are free from left recursion.

Smart constructors are used to make manipulating parser terms resemble writing parsley code itself. These are defined as infix operators, which are written as extension methods on the `Parser` trait:

```scala
implicit class ParserOps(private val p: Parser) extends AnyVal {
  def <*>(q: Parser): Parser = Ap(p, q)
  def <|>(q: Parser): Parser = Choice(p, q)
  def map(f: Function): Parser = FMap(p, f)
}
```

Parser terms can now be manipulated in a manner that looks almost indistinguishable from writing parsley code. For example, the unfold method on the `Ap` parser contains this snippet, where `pl`, `ql`, and `q` are parsers (`pe` is not a parser, but rather an `Option` value):

```scala
val lefts = {
  val llr = pl.map(flip) <*> q
  val rlr = pe.map(f => ql.map(composeH(f))).getOrElse(Empty)
  llr <|> rlr
}
```

Other than the capitalised `Empty` constructor, this would be perfectly valid `parsley` code.

1.1.4 Simplifying Parsers Using Parser Laws

Recombining unfolded parsers during left-recursion factoring introduces many necessary, but extraneous "glue" combinators. Even though the transformed parser is semantically correct, it ends up very noisy syntactically. Consider the resulting parser from factoring out the left-recursion in `expr`:

```
lazy val expr: Parsley[String] = chain.postfix(
  empty | (empty.map(a => b => a + b) | empty <*> expr) <*> string("a")
    | string("b") | empty
)(
  (empty.map(flip) <*> expr | pure(identity).map(compose(a => b => a + b)))
    .map(flip) <*> string("a")
    | empty | empty
)
```

The intent of this parser is completely obfuscated – it would be unacceptable for the output of the transformation to be left in this form. For human readability, this parser term must be simplified as much as possible, using domain-specific knowledge about parser combinators. This is where the deep embedding approach comes to shine; simplifications are easily expressed by pattern matching on `Parser` constructors.

Willis, Wu, and Pickering [2020] note that parser combinators are subject to *parser laws*, which often form a natural simplification in one direction. In Haskell `parsley`, Willis [2023] uses these parser laws as the basis for high-level optimisations to simplify the structure of the combinator tree. `parsley-garnish` uses the same principles to simplify the parser term to become more human-readable. The two only differ in the purpose of the simplification: whereas Haskell `parsley` does this to produce an optimised AST to be compiled as code, `parsley-garnish` simplifies the parser AST to be pretty-printed as text.

Fig. 1.2 shows the subset of parser laws utilised by `parsley-garnish` for parser simplification. Most of the laws in fig. 1.2 have already been shown to hold for Parsley by Willis and Wu [2018]; an additional proof for eq. (1.8) can be found in **??**.

$$p.map(f).map(g) = p.map(g \ compose \ f) \tag{1.1}$$
$$pure(f) <*> pure(x) = pure(f(x)) \tag{1.2}$$
$$pure(f) <*> x = x.map(f) \tag{1.3}$$
$$empty \ | \ u = u \tag{1.4}$$
$$u \ | \ empty = u \tag{1.5}$$
$$pure(x) \ | \ u = pure(x) \tag{1.6}$$
$$empty <*> u = empty \tag{1.7}$$
$$empty.map(f) = empty \tag{1.8}$$

Fig. 1.2: Functor (1.1), Applicative (1.2, 1.3), and Alternative (1.4–1.8) laws.

In the previous example, it is evident that the most noise results from the `empty` combinators. These can be eliminated using eqs. (1.4), (1.5), (1.7), and (1.8):

```
lazy val expr: Parsley[String] = chain.postfix(string("b"))(
  (pure(identity).map(compose(a => b => a + b))).map(flip) <*> string("a")
)
```

The complicated term in the postfix operator can then be simplified as follows:

```
(pure(identity).map(compose(a => b => a + b))).map(flip) <*> string("a")
```

=   { eqs. (1.2) and (1.3) }

```
pure(compose(a => b => a + b)(identity)).map(flip) <*> string("a")
```

=   { eqs. (1.2) and (1.3) }

```
pure(flip(compose(a => b => a + b)(identity))) <*> string("a")
```

=   { eq. (1.3) }

```
string("a").map(flip(compose(a => b => a + b)(identity)))
```

This results in the most simplified form of the parser:

```
val f: Function = flip(compose(a => b => a + b)(identity))
lazy val expr: Parsley[String] = chain.postfix(string("b"))(string("a").map(f))
```

Encapsulating boilerplate   Lawful simplifications are applied akin to peephole optimisations on the recursively defined `Parser` ADT. There are many instances of parsers, which inevitably leads to repetitive and error-prone boilerplate code which exists to simply recurse through each case. To avoid this, the recursive traversal itself is decoupled from the application of the transformation function. Although the traversal is still hand-written, the implementation is inspired by the generic traversal patterns offered by Haskell's `uniplate` library [Mitchell and Runciman 2007].

This is realised as a `transform` method on the `Parser` trait, which takes a partial function and applies it to nodes where it is defined. The transformation is applied via a bottom-up traversal:

```
def transform(pf: PartialFunction[Parser, Parser]): Parser = {
  val p = this match {
    case Ap(p, q)     => Ap(p.transform(pf), q.transform(pf))
    case Zipped(f, ps) => Zipped(f, ps.map(_.transform(pf)))
    case Pure(f)       => Pure(f)
    ...
  }
  if (pf.isDefinedAt(p)) pf(p) else p
}
```

A `rewrite` method can then be defined in terms of `transform`, applying the partial function everywhere and re-applying it until it no longer makes a change. This has the effect of applying a transformation exhaustively until a normal form is reached.

```
def rewrite(pf: PartialFunction[Parser, Parser]): Parser = {
  def pf0(p: Parser) = if (pf.isDefinedAt(p)) pf(p).rewrite(pf) else p
  this.transform(pf0)
}
```

Therefore, any transformation on parsers can be defined without having to worry about recursion boilerplate: the act of traversal itself is fully abstracted away and encapsulated within the `transform` method. Using `rewrite`, parser simplification can then be expressed in a clean and maintainable manner:

```
def simplify: Parser = this.rewrite {
  // p.map(f).map(g) == p.map(g compose f)
  case FMap(FMap(p, f), g) => FMap(p, composeH(g, f))
```

```
  // u <|> empty == u
  case Choice(u, Empty) => u
  // pure(f) <|> u == pure(f)
  case Choice(Pure(f), _) => Pure(f)
  ...
}
```

Further design considerations are made to ensure the extensibility and safety of this approach: the `Parser` trait is sealed, which enables compiler warnings if a new `Parser` case is added and the `transform` method is not updated. Since the traversal is still written by hand rather than generically derived, it is still more prone to error The traversal could be generically derived rather than written by hand, but this would require the use of an external dependency such as `shapeless`[1], which is overkill for the complexity of the `Parser` ADT.

1.1.5   Converting Parsers Back to Scalameta Terms

After parsers have been transformed and simplified, the last step is to convert them back to a textual representation to be applied as a Scalafix patch. Parsers can be lowered back to `scala.meta.Term` nodes by the inverse of the original `fromTerm` transformation. The `Parser` trait defines this transformation as the method `term`, using quasiquotes to simplify the construction of the `scala.meta.Term` nodes.

```
case class Zipped(func: Function, parsers: List[Parser]) extends Parser {
  val term: Term = q"(..${parsers.map(_.term)}).zipped(${func.term})"
}
```

This term can then be pretty-printed into a string, and applied as a Scalafix patch.

Summary

1.2   Representing and Normalising Expressions

Section 1.1 showed that it is useful to lift Scala AST nodes to a specialised `Parser` AST, making it easier to manipulate and inspect parsers. Crucially, this allowed us to simplify parsers via term-rewriting rules based on parser laws. Section 1.1.4 demonstrated why this is necessary for parsley-garnish: transformations such as left-recursion factoring ?? result in complex parser terms that must be simplified to be readable.

However, so far, the functions passed to lifting combinators such as `map` have been treated as opaque, and have not been subject to the same simplifications as parsers. This is evident from where we left off in the example from the previous section:

```
val f = flip(compose(a => b => a + b)(identity))
// f is equivalent to (a => b => b + a)
```

This mess is an artefact of the left-recursion factoring transformation – recombination of unfolded parsers requires using higher-order functions such as `flip` and `compose`. Yet again, any user would find it unacceptable if parsley-garnish gave this as the output of a transformation. Therefore, these functions must be *normalised* into a semantically equivalent but syntactically simpler form.

This section explores how function term normalisation can be achieved.

---

[1]https://github.com/milessabin/shapeless

1.2.1   The n-ary Lambda Calculus

Once again, the complexity of manipulating raw Scalameta AST nodes can be avoided by building a new intermediate AST representation for function terms.

Scala, as a functional programming language, uses an extension of the λ-calculus [Church 1936] as its theoretical foundations [Cremet et al. 2006; Amin et al. 2016]. The function terms that we want to normalise are equivalent to λ-terms, just with extra syntactic sugar. In the standard λ-calculus, each function only takes one argument, and multi-argument functions are represented as a chain of single-argument functions: this is known as *currying*. Scala supports curried functions using multiple parameter lists, but uncurried functions are preferred for performance reasons. Since these functions will be transformed from Scala code and back, it is desirable to maintain a high-level equivalence between these two representations. Thus, our representation will be based on fig. 1.3, which extends the λ-calculus to support proper multi-argument functions using n-ary abstraction and application.

β-Reduction and α-Conversion

In the λ-calculus, terms are evaluated via β-reduction: fig. 1.4 shows how this can be defined for the n-ary λ-calculus. Unlike the standard λ-calculus, reduction will only take place if the expected number of arguments in $\overline{\mathbf{x}}$ are equal to the number of arguments in $\overline{\mathbf{N}}$; otherwise, evaluation is stuck.

The syntax $M[N/x]$ denotes term substitution, where all free occurrences of x in M are replaced with N. Substitution must avoid *variable capture*, when N contains free variables that are bound in the scope where x is found [van Bakel 2022]. This requires performing α-conversion, which renames bound variables to avoid capture. In the λ-calculus, two terms are considered α-equivalent if they can be transformed into each other by renaming bound variables: the term $\lambda x.x$ is equivalent to $\lambda y.y$.

For example, substitution without α-conversion incorrectly β-reduces the following term:

$$(\lambda x.\lambda y.xy)y \rightarrow_\beta (\lambda y.xy)\,[y/x]$$
$$= \lambda y.yy$$

The y we wanted to substitute was originally a free variable, distinct from the y bound in the lambda $\lambda y.xy$. However, after substitution, it became captured under the lambda, where the two y terms are now indistinguishable in the incorrect expression $\lambda y.yy$. The correct β-reduction with capture-avoiding substitution would instead proceed as follows:

$$(\lambda x.\lambda y.xy)y \rightarrow_\beta (\lambda y.xy)\,[y/x]$$
$$=_\alpha (\lambda z.xz)\,[y/x]$$
$$= \lambda z.yz$$

$$
\begin{array}{lll}
M, N ::= & x & \text{variable} \\
\mid & (\lambda\overline{\mathbf{x}}.\, M) & n\text{-ary abstraction, where } \overline{\mathbf{x}} = (x_1, \dots, x_n) \\
\mid & (M\,\overline{\mathbf{N}}) & n\text{-ary application, wher } \overline{\mathbf{N}} = (N_1, \dots, N_n)
\end{array}
$$

Fig. 1.3: Syntax for the untyped λ-calculus extended with n-ary abstraction and application.

$$(\lambda\overline{\mathbf{x}}.\ M)\ \overline{\mathbf{N}} \rightarrow_\beta M[\overline{\mathbf{N}}/\overline{\mathbf{x}}] \qquad (\text{if } |\overline{\mathbf{x}}| = |\overline{\mathbf{N}}|)$$

Fig. 1.4: The β-reduction rule for the n-ary lambda calculus.

1.2.2   Representing Names

There exists a plethora of approaches to implementing the λ-calculus, mostly differing in how they represent variable names. This affects how variable capture is handled, and also how α-equivalence of two terms can be determined. For parsley-garnish, cheap α-equivalence is desirable to help check equivalence of parser terms, which is useful for some transformations.

Naïve capture-avoiding substitution   Representing variable names as strings is the most straightforward approach in terms of understandability. The example below shows how the simply-typed λ-calculus can be represented as a generalised algebraic data type (GADT) [Cheney and Hinze 2003] in Scala:

```scala
type VarName = String

trait Lambda
case class Abs[A, B](x: Var[A], f: Lambda[B]) extends Lambda[A => B]
case class App[A, B](f: Lambda[A => B], x: Lambda[A]) extends Lambda[B]
case class Var[A](name: VarName) extends Lambda[A]

// λf. λx. f x
val f = Var("f")
val x = Var("x")
val expr = Abs(f, Abs(x, App(f, x)))
```

Naïvely substituting these terms, however, seems logically simple but can be very tricky to get right. This approach requires calculating the free variables in a scope before performing substitution, renaming bound variables if it would lead to variable capture. This is not used in any real implementation of the λ-calculus due to its inefficiency, as it requires traversing the whole term tree multiple times. Checking α-equivalence is also tedious, requiring another full traversal of the term tree to compare variable names.

Barendregt's convention   Renaming all bound variables to be unique satisfies *Barendregt's convention* [Barendregt 1984], which removes the need to check for variable capture during substitution. However, to maintain this invariant, variables must also be renamed during substitution – this administrative renaming has relatively high performance overhead and chews through a scarily large number of fresh variable names. The approach has been successfully optimised to very impressive performance, though: the Haskell GHC compiler uses Barendregt's convention with a technique dubbed "the Rapier" [Peyton Jones and Marlow 2002], maintaining invariants to avoid renaming on substitution when unnecessary. Unfortunately, maintaining the invariants to keep this transformation correct becomes very difficult [Maclaurin, Radul, and Paszke 2023].

parsley-garnish originally used a named approach with Barendregt's convention, generating fresh variable names using an atomic counter. However, this required an extra α-conversion pass to clean up variable names before pretty-printing the term, since the fresh variable names were very ugly.

Nameless and hybrid representations   Nameless representations like *De Bruijn indices* [de Bruijn 1972] eschew names entirely, instead representing variables as the number of binders between the variable and its binding site. This makes α-equivalence trivial to check, as it is just a matter of comparing the indices. Although

an elegant representation, De Bruijn terms are notoriously difficult to work with, as they are not easily human-readable. Furthermore, performing substitutions with De Bruijn terms has an overhead as variable positions have to be shifted – this is undesirable given that the purpose of our ADT is to normalise λ-terms. To avoid this, hybrid representations combining named and nameless representations exist [McBride and McKinna 2004; Charguéraud 2012], but they become rather complex solutions for what should be a relatively simple λ-calculus implementation for `parsley-garnish`'s needs.

Higher-order abstract syntax   Using *higher-order abstract syntax* (HOAS) [Pfenning and Elliott 1988] sidesteps variable binders entirely by borrowing substitution from the meta-language. This makes it the meta-language's responsibility to handle variable capture instead. In contrast, the previous techniques were examples of first-order abstract syntax, which represents variables and unknowns with identifiers (whether with names or indices). A HOAS approach does not name bound variables, instead representing them as bindings in the meta-language:

```scala
trait HOAS
case class Abs[A, B](f: HOAS[A] => HOAS[B]) extends HOAS[A => B]
case class App[A, B](f: HOAS[A => B], x: HOAS[A]) extends HOAS[B]

// λf. λx. f x
val expr = Abs(f => Abs(x => App(f, x)))
```

Therefore, this representation performs substitution through Scala's function application, which makes it extremely fast compared to the other approaches. However, since lambda abstractions are represented as lambda expressions within Scala itself, the function body becomes wrapped under Scala's variable binding, making them difficult to work with.

1.2.3   Implementing Normalisation

One remaining hurdle stands before deciding on an ADT representation: how normalisation will be implemented. Partial evaluation and normalisation are related concepts – it is useful to view normalisation as statically evaluating as many terms as possible, but since not all terms have known values, the expression cannot be fully evaluated to a result value. Normalisation can thus be viewed simply as a process of evaluation, but in the presence of unknown terms. This section briefly explains the traditional notion of reduction-based normalisation, before introducing normalisation by evaluation as a more elegant and efficient alternative.

Reduction-Based Normalisation

The β-reduction rule is a *directed* notion of reduction, which can be implemented as a syntax-directed term-rewriting system, in a similar way to how `Parser` terms are simplified. The goal is to achieve beta normal form (β-NF) by allowing β-reduction to occur deep inside λ-terms, in all redexes of a term, until no more reductions can be made.

Normalisation by Evaluation

An interesting alternative approach stems from a notion of *reduction-free* normalisation, based on an undirected notion of term equivalence, rather than directed reduction. *Normalisation by Evaluation* (NBE) [Filinski and Korsholm Rohde 2004] achieves this by evaluating syntactical terms into a semantic model, then *reifying* them back into the syntactic domain. The denotational model (denoted by $[\![\ -\ ]\!]$) generally requires implementing a separate datatype from the syntactic AST representation of functions. The semantics is specifically constructed to

be *residualising*, meaning that terms can be extracted out into the original syntactic representation. Normalisation is then just defined as the composition of these two operations, as illustrated in fig. 1.5.

1.2.4   The Function ADT

parsley-garnish opts to implement normalisation with NBE, which results in a two-staged representation of function terms:

1. Scalameta terms are lifted to the `Function` ADT, which represents the syntax of terms using a simple named approach.

2. `Sem` uses HOAS to leverage Scala's semantics as the denotational model. During normalisation, `Function` terms are evaluated into `Sem`, then reified back into `Function`.

This helps achieve the following:

- The syntactic `Function` ADT is represented in a simple manner, which is easy to construct and manipulate as opposed to a HOAS representation. This allows function terms to be pattern matched on, as part of parser simplifications.

- Lifting the syntactic constructs to Scala semantics with HOAS allows for extremely efficient normalisation, and easier guarantees of correctness with respect to variable capture.

- Reifying `Sem` terms back into syntactic `Function` terms automatically α-converts names, giving us α-equivalence for free.

Lambda Calculus

The syntactic representation of functions is thus represented as the uni-typed `Function` ADT in fig. 1.6. `Defunc` is the higher-level representation of defunctionalised higher-order functions, which are desugared into `Lambda` terms for normalisation. `Lambda` represents *n*-ary λ-terms, extended with the following:

- Optional explicit type annotations for variables – these are not used for type-checking, but are there to preserve Scala type annotations originally written by the user.

- `Translucent` terms to encapsulate open terms holding a `scala.meta.Term` which cannot be normalised further. These carry an environment of variable bindings to substitute back in during pretty-printing – this is analogous to splicing into quoted expressions in the meta-programming world.

The original intention was to represent `Function` as a type-parameterised ADT for improved type safety, where `Lambda` would use a variant of the simply typed λ-calculus. This would've also allowed `Parser` to be parameterised by the result type of the parser. However, attempting to implement this ran into two main hurdles:

$$\text{Syntactic domain} \xrightarrow[\text{reify}]{\llbracket - \rrbracket} \text{Semantic domain}$$

$$normalise = reify \circ \llbracket - \rrbracket$$

Fig. 1.5: Normalisation by evaluation in a semantic model.

```scala
trait Function

// Core expression language
trait Lambda extends Function
case class Abs(xs: List[Var], f: Function) extends Lambda
case class App(f: Function, xs: List[Function]) extends Lambda
case class Var(name: VarName, displayType: Option[scala.meta.Type]) extends Lambda
case class Translucent(t: Term, env: Map[VarName, Function]) extends Lambda

// Defunctionalised higher-order functions
trait Defunc extends Function
case object Identity extends Defunc
case object Flip extends Defunc
case object Compose extends Defunc
...
```

Fig. 1.6: The `Function` ADT for representing functions.

- `Var` and `Translucent` terms would need to be created with concrete type parameters of their inferred types. Scalafix's semantic API is not powerful enough to guarantee that all terms can be queried for their inferred types – in fact, the built-in Scalafix rule *Explicit Result Types* calls the Scala 2 presentation compiler to extract information like this[2]. This solution is not ideal as it is brittle and breaks Scalafix's cross-compatibility promises.

- Scala 2's type inference for GADTs is less than ideal, requiring extra type annotations and unsafe casts which ultimately defeat the original purpose of type safety. This situation is improved, although not completely solved, in Dotty [Parreaux, Boruch-Gruszecki, and Giarrusso 2019] – but Scalafix does not yet support Scala 3.

Evaluating Performance of Normalisation Strategies

1.2.5   Converting Scalameta Terms to the Function ADT

There are three cases to consider when converting `scala.meta.Term` nodes into `Function` terms.

Lambda Expressions   Writing parsers often involves defining simple lambda expressions used to glue together parsers, or to transform the result of a parser:

```scala
val asciiCode: Parsley[Int] = item.map(char => char.toInt)
```

These lambda expressions are represented in the Scalameta AST as `Term.Function` nodes, which are recursively traversed to collect all parameter lists. This is folded into a chain of *n*-ary abstractions, with the final term being the body of the lambda, which is wrapped into a `Translucent` term. To ensure that the parameter names in the `Translucent` body term are unique, the parameters are α-converted to fresh names. Take the following example to see why this is helpful:

```scala
a => (a, b) => a + b
```

Although any sane Scala programmer would not write this, this convoluted example shows how variable shadowing can occur – the a in the function body refers to the a in the second parameter list, as it shadows the a in the first parameter list. The resulting `Function` term would then resemble the following λ-calculus expression:

---

[2]https://github.com/scalacenter/scalafix/issues/1583

```
λ(_l1). λ(_l2, _l3). Translucent(_l2 + _l3, env = {_l1 → _l1, _l2 → _l2, _l3 → _l3})
```

Values shown in bold are `scala.meta.Term` nodes, so the lambda body's environment maps `Term.Name` nodes to their corresponding variable terms. When the term is pretty-printed, the `Term.Name` nodes are replaced with the corresponding `Function` terms – this is similar to the splicing operation on quasiquotes.

Placeholder Syntax   Scala supports a placeholder syntax using underscores to make lambda expressions more concise, so the earlier parser can be rewritten as:

```
val asciiCode: Parsley[Int] = item.map(_.toInt)
```

Scalameta differentiates between regular lambda expressions and those using placeholder syntax, representing the latter as `Term.AnonymousFunction` nodes. This makes it easy to identify which approach to be taken during conversion. To convert this case, each successive underscore in the expression body is replaced with a fresh variable name. Using placeholder syntax only results in a fully uncurried function with a single parameter list[3] – this is filled with the freshly generated variable names in order of their occurrence in the expression body.

Eta-Expansion   If the term is not a lambda expression, `parsley-garnish` attempts to η-expand the term if possible. For example, an idiomatic parser written using the *Parser Bridges* pattern [Willis and Wu 2022] could resemble the following:

```
case class AsciiCode(code: Int)
object AsciiCode extends ParserBridge1[Char, AsciiCode] {
  def apply(char: Char): AsciiCode = AsciiCode(char.toInt)
}
val asciiCode = AsciiCode(item)
```

When `parsley-garnish` converts `asciiCode` to a `Parser`, it desugars the bridge constructor into something resembling `item.map(AsciiCode.apply)`. The η-expanded form of `AsciiCode.apply` would be as follows:

```
(char: Char) => AsciiCode.apply(char)
```

To η-expand `scala.meta.Term` nodes, `parsley-garnish` attempts to look up the method signature of its symbol using Scalafix's semantic API. This is not always possible – in that case, the term can't be statically inspected any further and is just wrapped in a `Translucent` term.

1.2.6   Normalising Function Terms

the code below is outdated

Reflection

```
trait Function {
  def reflect: HOAS = {
    def reflect0(func: Function, boundVars: Map[Var, HOAS]): HOAS = func match {
      case Abs(xs, f) =>
        HAbs(xs.size, vs => reflect0(f, boundVars ++ xs.zip(vs)))
      case App(f, xs) =>
        HApp(reflect0(f, boundVars), xs.map(reflect0(_, boundVars)))
      case v @ Var(name) =>
        boundVars.getOrElse(v, HVar(name))
```

---

[3] https://www.scala-lang.org/files/archive/spec/2.13/06-expressions.html#anonymous-functions
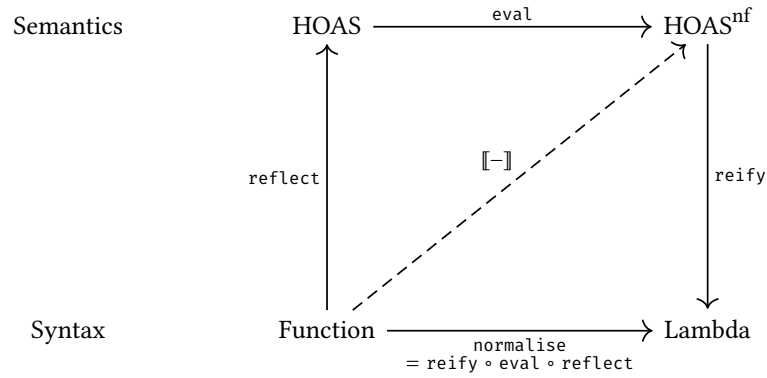
Fig. 1.7: Normalisation by evaluation for the `Function` datatype.

```
    case Translucent(term, env) =>
      HTranslucent(term, env.mapValues(reflect0(_, boundVars)))

    case Id => HAbs(f => f)
    case Flip => HAbs(f => HAbs(x => HAbs(y => HApp(HApp(f, y), x))))
    case Compose => HAbs(f => HAbs(g => HAbs(x => HApp(f, HApp(g, x)))))
  }

  reflect0(this, Map.empty)
  }
}
```

Evaluation

```
trait HOAS {
  def eval: HOAS = this match {
    case HAbs(n, f) => HAbs(n, x => f(x).eval)
    case HApp(f, x) => f.whnf match {
      case HAbs(_, g) => g(x).eval
      case g          => HApp(g.eval, x.map(_.eval))
    }
    case HTranslucent(t, env) => HTranslucent(t, env.map { case (k, v) => k -> v.eval })
    case HTranslucent(t, env) => HTranslucent(t, env.mapValues(_.eval))
    case _ => this
  }

  private def whnf: HOAS = this match {
    case HApp(f, x) => f.whnf match {
      case HAbs(_, g) => g(x).whnf
      case g          => HApp(g, x)
    }
    case HTranslucent(t, env) => HTranslucent(t, env.mapValues(_.whnf))
    case _ => this
  }
}
```

Reification

```scala
trait HOAS {
  def reify: Function = this match {
    case HAbs(n, f) =>
      val params = (1 to n).map(_ => Var.fresh()).toList
      Abs(params, f(params.map(x => HVar(x.name))).reify)
    case HApp(f, xs) => App(f.reify, xs.map(_.reify))
    case HVar(name) => Var(name)
    case HTranslucent(t, env) => Translucent(t, env.mapValues(_.reify))
  }
}
```

1.2.7   Converting Functions back to Scalameta Terms

Surprise bitches same shit again, quasiquotes ftw

1.2.8   Further Work?

Eta reduction – this is more complicated than in Haskell since Scala has special syntax Partial evaluation, not just normalisation (if we reduce to fully closed terms 1+1 can we get it to evaluate to 2? – except currently this would be a Translucent term)

TODO: Abstraction built over scalafix/meta ASTs to represent functions. Allows us to statically evaluate function composition/flipping etc, so it doesn't turn into one big mess – again, human readability of the transformed output is the goal. Abstraction is again an ADT as a lambda calculus, but with parameter lists so not everything is curried. îdk, this is still a work-in-progress. Seems that there might not be enough time to uncurry the leftrec analysis so this design decision might not be super important. Representation as a lambda calc has allocation overhead, but greatly simplifies function evaluation via beta reduction, instead of having to deal with high-level representations of compose/id (not too bad tbh) and flip (annoying). Also attempted to make it typed but that didn't go so well with Scala's limitations on type inference.

* Extracting method arguments (alongside their types) is very painful * Need to unify information from signature (within symbolinformation) and synthetics * synthetics exist in certain cases: .apply methods, showing the concrete type of a generic argument, implicit conversions * from https://scalacenter.github.io/scalafix/docs/developers/semantic-tree.html: SemanticTree is a sealed data structure that encodes tree nodes that are generated by the compiler from inferred type parameters, implicit arguments, implicit conversions, inferred .apply and for-comprehensions.

* Don't have full access to type information - can do more work here theoretically, but its difficult and error-prone * So we don't model a typed lambda calculus, just have it untyped

Approaches - AVOIDING capture via substitution * Substitution approaches * De Bruijn indices - inefficient to open/close terms so much - De Bruijn levels as an alternative * HOAS * Normalisation by evaluation