Safely Composable Type-Specific Languages

Cyrus Omar, Darya Kurilova, Ligia Nistor, Benjamin Chung,
Alex Potanin[1], and Jonathan Aldrich

Carnegie Mellon University and [1]Victoria University of Wellington
{comar, darya, lnistor, bwchung, aldrich}@cs.cmu.edu and [1]alex@ecs.vuw.ac.nz

Abstract. Programming languages often include specialized notation for common datatypes (e.g. lists) and some also build in support for specific specialized datatypes (e.g. regular expressions), but user-defined types must use general-purpose notations. Frustration with this causes developers to use strings, rather than structured representations, with alarming frequency, leading to correctness, performance, security and usability issues. Allowing library providers to modularly extend a language with new notations could help address these issues. Unfortunately, prior mechanisms either limit expressiveness or are not safely composable: individually unambiguous extensions can still lead to ambiguities when used together. We introduce *type-specific languages* (TSLs): logic associated with a type that determines how the bodies of *generic literals*, able to contain arbitrary syntax, are parsed and elaborated, hygienically. The TSL for a type is invoked only when a literal appears where a term of that type is expected, guaranteeing non-interference. We give evidence supporting the applicability of this approach and formally specify it with a bidirectionally typed elaboration semantics for the Wyvern language.

Keywords: extensible languages; parsing; bidirectional typechecking

1 Motivation

Many data types can be seen, semantically, as modes of use of recursive product and sum types. For example, lists can be defined using a more general mechanism for defining recursive sum types by observing that a list can either be empty, or be broken down into a product of the *head* element and the *tail*, another list. In an ML-like functional language, sums are exposed as datatypes and products as tuples and records, so list types can be defined as follows:

```
datatype 'a list = Nil | Cons of 'a * 'a list
```

In class-based object-oriented language, objects are products of their fields and class-based method dispatch exposes a form of sum type [15].

By defining types like `int list` in terms of these general purpose semantic mechanisms, we immediately benefit from powerful reasoning principles (e.g. in ML, structural induction) and language support (e.g. pattern matching). They are already well optimized by compilers and benefit from general-purpose editor support as well. While this is all quite useful, this generality can sometimes come at a cost: general-purpose *syntax* is often a burden even when a general-purpose

semantics is useful. For example, few would claim that writing a list of numbers as a sequence of `Cons` cells is convenient:

```
Cons(1, Cons(2, Cons(3, Cons(4, Nil))))
```

Because lists are a common data structure, many languages include *literal notation* for introducing them, e.g. `[1, 2, 3, 4]`. This notation is semantically equivalent to the general-purpose notation shown above, but brings cognitive benefits by drawing attention to the content of the list, rather than the nature of the encoding. Using terminology from Green's cognitive dimensions of notations [14], it is more *terse*, *visible* and *maps more closely* to the intuitive notion of a list. Stoy, in discussing the value of good notation, writes [**?**]:

> A good notation thus conceals much of the inner workings behind suitable abbreviations, while allowing us to consider it in more detail if we require: matrix and tensor notations provide further good examples of this. It may be summed up in the saying: "A notation is important for what it leaves out."

Although list, number and string literals are nearly ubiquitous features of modern languages, some languages additionally provide specialized notation for other common data structures (like maps, sets, vectors and matrices), data formats (like XML and JSON), query languages (like regular expressions and SQL), markup languages (like HTML and Markdown) and many other types of data. For example, a language with built-in notation for HTML and SQL, supporting type safe *splicing* via curly braces, might define:

```
let webpage : HTML = <html><body><h1>Results for {keyword}</h1>
  <ul id="results">{to_list_items(query(db,
    SELECT title, snippet FROM products WHERE {keyword} in title)}
  </ul></body></html>
```

as shorthand for:

```
let webpage : HTML = HTMLElement(Dict.empty(), [BodyElement(Dict.empty(),
  [H1Element(Dict.empty(), [TextNode("Results for " + keyword)]),
  ULElement((Dict.add Dict.empty() ("id", "results")), to_list_items(query(db,
    SelectStmt(["title", "snippet"], "products",
      [WhereClause(InPredicate(StringLit(keyword), "title"))]))))]]]
```

When a specialized notation like this is not available, but the equivalent general-purpose notation is too cognitively demanding for comfort, developers typically turn to run-time mechanisms to make constructing data structures more convenient. Among the most common strategies across language paradigms in these situations is to simply use a string representation parsed at run-time:

```
let webpage : HTML = parse_html("<html><body><h1>Results for "+keyword+"</h1>
  <ul id=\"results\">" + to_string(to_list_items(query(db, parse_sql(
    "SELECT title, snippet FROM products WHERE '"+keyword+"' in title")))) +
  "</ul></body></html>")
```

Though recovering some of the notational convenience of the literal version, it is still more awkward to write, requiring explicit conversions to and from structured representations (`parse_html` and `to_string`, respectively) and escaping when the syntax of the language interferes with the syntax of string literals (line

2). Code like this also causes a number of problems beyond cognitive load. Because parsing occurs at run-time, syntax errors will not be discovered statically, causing potential problems in production scenarios. Run-time parsing also incurs performance overhead, particularly relevant when code like this is executed often (as on a heavily-trafficked website). But the most serious issue with this code is that it is fundamentally insecure: it is vulnerable to cross-site scripting attacks (line 1) and SQL injection attacks (line 3). For example, if a user entered the keyword `'; DROP TABLE products --`, the entire product database could be erased. These attack vectors are considered to be two of the most serious security threats on the web today [3]. Although developers are cautioned to sanitize their input, it can be difficult to verify that this was done correctly throughout a codebase. The best way to avoid these problems today is to avoid strings and insist on structured representations, despite the inconvenient notation.

Unfortunately, situations like this, where maintaining strong correctness, performance and security guarantees entails significant syntactic overhead, causing developers to turn to worse solutions that are more convenient, are quite common. To emphasize this, let us return to our running example of pattern literals. A small regular expression like `(\d\d):(\d\d)\w*((am)|(pm))` might be written using general-purpose notation as:

```
1  Seq(Group(Seq(Digit, Digit), Seq(Char(":"), Seq(Group(Seq(Digit, Digit)),
2    Seq(ZeroOrMore(Whitespace), Group(Or(Group(Seq(Char("a"), Char("m"))),
3    Group(Seq(Char("p"), Char("m"))))))))))
```

This is clearly more cognitively demanding, both when writing the regular expression and when reading it. Among the most common strategies in these situations, for users of any kind of language, is again to simply use a string representation that is parsed at run-time:

```
1  rx_from_str("(\\d\\d):(\\d\\d)\\w*((am)|(pm))")
```

This is problematic, for all of the reasons described above: excessive conversions between representations, interference issues with string syntax, correctness problems, performance overhead and security issues.

Today, supporting new literal notations within an existing language requires the cooperation of the language designer. This is primarily because, with conventional parsing strategies, not all notations can unambiguously coexist, so a designer is needed to make choices about which syntactic forms are available and what their semantics should be. For example, conventional notations for sets and maps are both delimited by curly braces. When Python introduced set literals, it chose to distinguish them based on whether the literal contained only elements (e.g. `{3}`), or key-element pairs (`{"x": 3}`). But this causes an ambiguity with the syntactic form `{ }` – should it mean an empty set or an empty map (called a dictionary in Python)? The designers of Python chose the latter interpretation (for backwards compatibility reasons).

So although languages that allow providers to introduce new syntax from within libraries appear to hold promise for the reasons described above, enabling this form of extensibility is non-trivial because there is no longer a central designer making decisions about such ambiguities. In most existing related work,

the burden of resolving ambiguities falls to clients who have the misfortune of importing conflicting extensions. For example, SugarJ [12] and other extensible languages generated by Sugar* [**?**] allow providers to extend the base syntax of the host language with new forms, like set and map literals. These new forms are imported transitively throughout a program. To resolve syntactic ambiguities that arise, clients must manually augment the composed grammar with new rules that allow them to choose the correct interpretation explicitly. This is both difficult to do, requiring a reasonably thorough understanding of the underlying parser technology (in Sugar*, generalized LR parsing) and increases the cognitive load of using the conflicting notations (e.g. both sets and dictionaries) in the same file because disambiguation tokens must be used. These kinds of conflicts occur in a variety of circumstances: HTML and XML, different variants of SQL, JSON literals and dictionaries, or simply different implementations ("desugarings") of the same specialized syntax (e.g. two regular expression engines). Techniques that limit the kinds of syntax extensions that can be expressed, to guarantee that ambiguities cannot occur, simply cannot express these kinds of examples as-is (e.g. [**?**]).

In this work, we will describe an alternative parsing strategy that avoids these problems by shifting responsibility for parsing certain *generic literal forms* into the typechecker. The typechecker, in turn, defers responsibility to user-defined types, by treating the body of the literal as a term of the *type-specific language (TSL)* associated with the type it is being checked against. The TSL is responsible for rewriting this term to ultimately use only general-purpose notation. This strategy avoids the problem of conflicting syntax, because neither the base language nor TSLs are ever extended directly. It also permits semantic flexibility – the meaning of a form like `{ }` can differ depending on its type, so it is safe to use it for empty sets, maps and other data structures, like JSON literals. This frees these common notations from being tied to the variant of a data structure built into a language's standard library, which sometimes does not provide the exact semantics that a programmer needs (for example, Python dictionaries do not preserve key insertion order).

We develop our work as a variant of an emerging programming language called Wyvern [22]. To allow us to focus on the essence of our proposal, the variant of Wyvern we develop here is simpler than the variant previously described: it is purely functional (there are no effects other than non-termination) and it does not enforce a uniform access principle for objects (fields can be accessed directly). It also adds recursive sum types, which we call *case types*, similar to those found in ML. One can refer to our version of the language as *TSL Wyvern* when the variant being discussed is not clear. Our work substantially extends and makes concrete a mechanism sketched in an earlier short workshop paper [23]. We make the following novel contributions **[TODO: condense this per reviewer comments]**:

1. We specify a more complete layout-sensitive concrete syntax and show how it can be written without the need for a context-sensitive lexer or parser using an Adams grammar. It now includes a variety of inline literals, provides a

full specification for the whitespace-delimited literal form introduced by a forward reference, ~, and provides other forms of forward-referenced forms.
2. We develop a general mechanism for associating metadata with a type. A TSL is then implemented by associating a parser (of type `Parser`) with a type. The parser is responsible for rewriting tokenstreams (of type `Tokenstream`) into Wyvern ASTs (of type `Exp`). These types are defined in the standard library.
3. This lower-level mechanism is general, but writing a hand-written parser and manipulating syntax trees manually is cognitively demanding. We observe that *grammars* and *quasiquotes* can both be seen as TSLs for parsers and ASTs respectively and discuss how to implement them as such.
4. A naïve rewriting strategy would be *unhygienic* – it could allow for the inadvertent capture of local variables. We show a novel mechanism that ensures hygiene by requiring that the generated AST is closed except for subtrees derived from portions of the user's tokenstream that are interpreted as nested Wyvern expressions. We also show how to explicitly refer to local values available in the parser definition (e.g. helper functions) in a safe way.
5. We formalize the static semantics and literal parsing rules of TSL Wyvern as a bidirectional type system. By distinguishing locations where an expression synthesizes a type from locations where an expression is being analyzed against a previously synthesized type, we can precisely state where generic literals can appear. This also formalizes the hygiene mechanism.
6. We provide several examples of TSLs throughout the paper, but to examine how broadly applicable the technique is, we conduct a simple corpus analysis, finding that string languages are used ubiquitously.

2 Type-Specific Languages in Wyvern

We begin with an example in Fig. 1 showing several different TSLs being used to define a fragment of a web application showing search results from a database. We will review this example below to develop intuitions about TSLs in Wyvern; a formal and more detailed description will follow in the subsequent sections.

2.1 Wyvern

We develop our work as a variant of a new programming language being developed by our group called Wyvern [22]. To allow us to focus on the essence of our proposal, the variant of Wyvern we will describe in this thesis is simpler than the variant previously described: it is purely functional (there are no effects other than non-termination) and it does not enforce a uniform access principle for objects (fields can be accessed directly). Objects can thus be thought of as recursive labeled products with support for simple methods (functions that are automatically given a self-reference) for convenience. We also add recursive labeled sum types, which we call *case types*, that are quite similar to datatypes in ML. One can refer to the version of the language described in this thesis as *TSL Wyvern*. TSL Wyvern has a layout-sensitive syntax, for reasons we will discuss.

2.2 Example: Web Search

We begin in Fig. 1 with an example showing several different TSLs being used to define a fragment of a web application showing search results from a database.

```
1  let imageBase : URL = <images.example.com>
2  let bgImage : URL = <%imageBase%/background.png>
3  new : SearchServer
4    def resultsFor(searchQuery, page)
5      serve(~) (* serve : HTML -> Unit *)
6        >html
7          >head
8            >title Search Results
9            >style ~
10             body { background-image: url(%bgImage%) }
11             #search { background-color: %darken(`#aabbcc`, 10pct)% }
12         >body
13           >h1 Results for < HTML.Text(searchQuery)
14           >div[id="search"]
15             Search again: < SearchBox("Go!")
16           < (* fmt_results : DB * SQLQuery * Nat * Nat -> HTML *)
17              fmt_results(db, ~, 10, page)
18                SELECT * FROM products WHERE {searchQuery} in title
```

Fig. 1: Wyvern Example with Multiple TSLs

```
<literal body here, <inner angle brackets> must be balanced>
{literal body here, {inner braces} must be balanced}
[literal body here, [inner brackets] must be balanced]
`literal body here, ``inner backticks`` must be doubled`
'literal body here, ''inner single quotes'' must be doubled'
"literal body here, ""inner double quotes"" must be doubled"
12xyz (* no delimiters necessary for number literals; suffix optional *)
```

Fig. 2: Inline Generic Literal Forms

Note that for clarity of presentation, we color each character according to the TSL it is governed by. Black represents the base language and comments are in italics.

2.3 Inline Literals

Our first TSL appears on the right-hand side of the variable binding on line 1. The variable `imageBase` is annotated with its type, `URL`. This is a named object type[1] declaring several fields representing the components of a URL: its protocol, domain name, port, path and so on (not shown). We could have created a value of type `URL` using general-purpose notation:

```
1  let imageBase : URL = new
2    val protocol = "SUShttpEUS"
3    val subdomain = "SUSimagesEUS"
4    (* ... *)
```

This is tedious. Because the `URL` type has a TSL associated with it, we can instead introduce precisely this value using conventional notation for URLs by placing it in the *body* of a *generic literal*, `<images.example.com>`. Any other delimited form in Fig. 2 could equivalently be used if the constraints shown are obeyed. The type annotation on `imageBase` implies that this literal's *expected type* is `URL`, so the

[1] We do not currently support polymorphic types in Wyvern, so in our discussion types and type constructors are indistinguishable, and we will simply use "types" for concision.

body of the literal (the characters between the angle brackets, in blue) will be governed by the `URL` TSL during the typechecking phase. This TSL will parse the body (at compile-time) to produce a Wyvern abstract syntax tree (AST) that explicitly instantiates a new object of type `URL` using general-purpose notation as if the above had been written directly. We will return to how this works shortly.

In addition to supporting conventional notation for URLs, this TSL supports *splicing* another Wyvern expression of type `URL` to form a larger URL. The spliced term is delimited by percent signs, as seen on line 2 of Fig. 1. The TSL parses code between percent signs as a Wyvern expression, using its abstract syntax tree (AST) to construct an AST for the expression as a whole. A string-based representation of the URL is never used at run-time. Note that the delimiters used to go from Wyvern to a TSL are controlled by Wyvern while the TSL controls how to return to Wyvern.

2.4 Layout-Delimited Literals

On line 5 of Fig. 1, we see a call to a function `serve` (not shown) which has type `HTML -> Unit`. Here, `HTML` is a user-defined *case type*, having cases for each HTML tag as well as some other structures, like text nodes and sequencing. Declarations of some of these cases can be seen on lines 2-6 of Fig. 3 (note that TSL Wyvern also includes simple product types for convenience, written `T1 * T2`). We could again use Wyvern's general-purpose introductory form for case types, e.g. `HTML.BodyElement((attrs, child))` (unlike in ML, in Wyvern we must explicitly qualify constructors with the case type they are part of when they are used. This is largely to make our formal semantics simpler and for clarity of presentation.) But, as discussed above, using this syntax can be inconvenient and cognitively demanding. Thus, we associate a TSL with `HTML` that provides a simplified notation for writing HTML, shown being used on lines 6-18 of Fig. 1. This literal body is layout-delimited, rather than delimited by explicit tokens as in Fig. 2, and introduced by a form of *forward reference*, written ~ ("tilde"), on the previous line. Because the forward reference occurs in a position where the expected type is `HTML`, the literal body is governed by that type's TSL. The forward reference will be replaced by the general-purpose term, of type `HTML`, generated by the TSL during typechecking. Because layout was used as a delimiter, there are no syntactic constraints on the body, unlike with inline forms (Fig. 2). For HTML, this is quite useful, as all of the inline forms impose constrains that would cause conflict with some valid HTML.

2.5 Implementing a TSL

Portions of the implementation of the TSL for `HTML` are shown on lines 8-15 of Fig. 3. A TSL is associated with a named type, forming an *active type*, using a more general mechanism for associating a pure, static value with a named type, called its *metadata*. Metadata is introduced as shown on line 8 of Fig. 3. Type metadata, in this context, is comparable to class annotations in Java or attributes in C#/F# and internalizes the practice of writing metadata using comments, so that it can be checked by the language and accessed programmatically more

```
casetype HTML
  Empty
  Seq of HTML * HTML
  Text of String
  BodyElement of Attributes * HTML
  StyleElement of Attributes * CSS
  (* ... *)
  metadata = new : HasTSL
    val parser = ~
      start <- '>body'= attributes start>
        fn attrs, child => `HTML.BodyElement((%attrs%, %child%))`
      start <- '>style'= attributes EXP>
        fn attrs, e => `HTML.StyleElement((%attrs%, %e%))`
      start <- '<'= EXP>
        fn e => `%e% : HTML`
```

Fig. 3: A Wyvern case type with an associated TSL.

```
objtype HasTSL
  val parser : Parser
objtype Parser
  def parse(ps : ParseStream) : ParseResult
  metadata : HasTSL = new
    val parser = (* parser generator *)
casetype ParseResult
  OK of Exp * ParseStream
  Error of String * Location
```

```
casetype Exp
  Var of ID
  Lam of ID * Type * Exp
  Ap of Exp * Exp
  Tuple of Exp * Exp
  ...
  Spliced of ParseStream
  metadata : HasTSL = new
    val parser = (* quasiquotes *)
```

Fig. 4: Some of the types included in the Wyvern prelude. They are mutually defined.

easily. This can be used for a variety of purposes – to associate documentation with a type, to mark types as being deprecated, and so on.

For the purposes of this work, metadata values will always be of type `HasTSL`, an object type that declares a single field, `parser`, of type `Parser`. The `Parser` type is an object type declaring a single method, `parse`, that transforms a `ParseStream` extracted from a literal body to a Wyvern AST. An AST is a value of type `Exp`, a case type that encodes the abstract syntax of Wyvern expressions. Fig. 4 shows portions of the declarations of these types, which live in the Wyvern *prelude* (a collection of types that are automatically loaded before any other).

Notice, however, that the TSL for `HTML` is not provided as an explicit `parse` method but instead as a declarative grammar. A grammar is a specialized notation for defining a parser, so we can implement a more convenient grammar-based parser generator as a TSL associated with the `Parser` type. We chose the layout-sensitive formalism developed by Adams [4] – Wyvern is itself layout-sensitive and has a grammar that can be written down using this formalism, so it is sensible to expose it to TSL providers as well. Most aspects of this formalism are completely conventional. Each non-terminal (e.g. `start`) is defined by a number of disjunctive productions, each introduced using `->`. Each production defines a sequence of terminals (e.g. `'>body'`) and non-terminals (e.g. `start`, or one of the built-in non-terminals `ID`, `EXP` or `TYPE`, representing Wyvern identifiers, expressions and types, respectively). Unique to Adams grammars is that each terminal

and non-terminal in a production can also have an optional *layout constraint* associated with it. The layout constraints available are `=` (meaning that the leftmost column of the annotated term must be aligned with that of the parent term), `>` (the leftmost column must be indented further) and `>=` (the leftmost column *may* be indented further). We will discuss this formalism further when we formally specify Wyvern's layout-sensitive concrete syntax.

Each production is followed, in an indented block, by a Wyvern function that generates a value given the values recursively generated by each of the n non-terminals it contains, ordered left-to-right. For the starting non-terminal, always called `start`, this function must return a value of type `Exp`. User-defined non-terminals might have a different type associated with them (not shown). Here, we show how to generate an AST using general-purpose notation for `Exp` (lines 13-15) as well as a more natural *quasiquote* style (lines 11 and 18). Quasiquotes are expressions that are not evaluated, but rather reified into syntax trees. We observe that quasiquotes too fall into the pattern of "specialized notation associated with a type" – quasiquotes for expressions, types and identifiers are simply TSLs associated with `Exp`, `Type` and `ID` (Fig. 4). They support the full Wyvern concrete syntax as well as an additional delimited form, written with `%s`, that supports "unquoting": splicing another AST into the one being generated. Again, splicing is safe and structural, rather than based on string interpolation.

We have now seen several examples of TSLs that support splicing. The question then arises: what type should the spliced Wyvern expression be expected to have? This is determined by placing the spliced value in a place in the generated AST where its type is known – on line 11 of Fig. 3 it is known to be `HTML` and on line 13 it is known to be `CSS` by the declaration of `HTML`, and on line 15, it is known to be `HTML` by the use of an explicit ascription. When these generated ASTs are recursively typechecked during compilation, any use of a nested TSL at the top-level (e.g. the CSS TSL in Fig 1) will operate as intended.

2.6 Implementing Interpolation

We have now seen several examples of interpolation. Within the TSL for `HTML`, we see it used in several ways:

HTML Interpolation At any point where a tag should appear, we can also interpolate a Wyvern expression of type `HTML` by enclosing it within curly braces (e.g. on line 13, 15 and 16-19 of Fig. 1). This is implemented on lines 17 and 18 of Fig. 3. The special non-terminal `EXP[T]` signals a switch into parsing a Wyvern expression. The tokenstream will be parsed as a Wyvern expression until a `T` token is encountered *that would otherwise trigger a parse error*. In other words, the Wyvern grammar binds more tightly to itself than to any surrounding TSL. The AST for the parsed Wyvern expression is given an expected type, `HTML`, by simply surrounding it with an ascription (line 18). Because interpolation must be structured (a string cannot be interpolated directly), injection and cross-site scripting attacks cannot occur. Safe string interpolation (which escapes any inner HTML) could be implemented using another delimiter.

CSS Interpolation After the `:style` tag appears (e.g. on line 9 of Fig. 1), instead of hard-coding CSS syntax into the HTML DSL, we instead wish to use the TSL associated with a type representing a CSS stylesheet: `CSS`. We do this by again interpolating a Wyvern expression (lines 12-15 of Fig. 3), making sure that it appears in a position where the expected type is `CSS` (the second piece of data associated with the `StyleElement` constructor, in this case). Wyvern is given control until a full expression has been read and an unexpected newline appears (that is, a newline that does not introduce a layout-delimited block).

Interpolation within the CSS TSL The TSL for `CSS` itself has support for interpolation in a similar manner, choosing `%` as the delimiter. It chooses the type based on the semantics of the surrounding CSS form. For example, when a Wyvern expression appears inside `url`, as on line 10 of Fig. 1, it must be of type `URL`. When a Wyvern expression appears where a color is needed, the `Color` type is used. This type itself has a TSL associated with it that interprets CSS color strings, showing that TSLs can be used within TSLs by simply escaping out to Wyvern, the host language, and then back in. In this case, we emphasize that TSLs produced structured values by calling the `darken` method on it to produce a new color. This method itself takes a `Percentage` as an argument. The TSL for this type accepts literal bodies containing numbers followed by `pct`, or simply a real number without a suffix. These literal bodies, because they begin with a number (and no other form in Wyvern can), does not require delimiters (Fig. 2).

Interpolation within the SQLQuery TSL The TSL used for SQL queries on line 18 of Fig. 1 follows an identical pattern, allowing strings to be interpolated into portions of a query in a safe manner. This prevents SQL injection attacks.

3 Syntax

3.1 Concrete Syntax

We will now describe the concrete syntax of Wyvern declaratively, using the same layout-sensitive formalism that we have introduced for TSL grammars, developed recently by Adams [4]. Such a formalism is useful because it allows us to implement layout-sensitive syntax, like that we've been describing, without relying on context-sensitive lexers or parsers. Most existing layout-sensitive languages (e.g. Python and Haskell) use hand-rolled context-sensitive lexers or parsers (keeping track of, for example, the indentation level using special `INDENT` and `DEDENT` tokens), but these are more problematic because they cannot be used to generate editor modes, syntax highlighters and other tools automatically. In particular, we will show how the forward references we have described can be correctly encoded without requiring a context-sensitive parser or lexer using this formalism. It is also useful that the TSL for `Parser`, above, uses the same parser technology as the host language, so that it can be used to generate quasiquotes.

Wyvern's concrete syntax, with a few minor omissions for concision, is shown in Figure 5. We first review Adams' formalism in some additional detail, then describe some key features of this syntax.

3.2 Background: Adams' Formalism

For each terminal and non-terminal in a rule, Adams proposed associating with them a relational operator, such as $=$, $>$ and $\geq$ to specify the indentation at which those terms need to be with respect to the non-terminal on the left-hand side of the rule. The indentation level of a term can be identified as the column at which the left-most character of that term appears (not simply the first character, in the case of terms that span multiple lines). The meaning of the comparison operators is akin to their mathematical meaning: $=$ means that the term on the right-hand side has to be at exactly the same indentation as the term on the left-hand side; $>$ means that the term on the right-hand side has to be indented strictly further to the right than the term on the left-hand side; $\geq$ is like $>$, except the term on the right could also be at the same indentation level as the term on the left-hand side. For example, the production rule of the form $\texttt{A} \rightarrow \texttt{B}^{=}\ \texttt{C}^{\geq}\ \texttt{D}^{>}$ approximately reads as: "Term `B` must be at the same indentation level as term `A`, term `C` may be at the same or a greater indentation level as term `A`, and term `D` must be at an indentation level greater than term `A`'s." In particular, if `D` contains a `NEWLINE` character, the next line must be indented past the position of the left-most character of `A` (typically constructed so that it must appear at the beginning of a line). There are no constraints relating `D` to `B` or `C` other than the standard sequencing constraint: the first character of `D` must be to further in the file than the others. Using Adam's formalism, the grammars of real-world languages like Python and Haskell can be written declaratively. This formalism can be integrated into LR and LALR parser generators.

3.3 Programs

An example Wyvern program showing several unique syntactic features of TSL Wyvern is shown in Fig. 1. The top level of a program (the `p` non-terminal) consists of a series of type declarations – object types using **`objtype`** or case types using **`casetype`** – followed by an expression, `e`. Each type declaration contains associated declarations – signatures for fields and methods in `objdecls` and case declarations in `casedecls` (not shown on the figure). Each also can also include a metadata declaration. Metadata is simply an expression associated with the type, used to store TSL logic (and in future work, other logic). Sequences of top-level declarations use the form $\texttt{p}^{=}$ to signify that all the succeeding `p` terms must begin at the same indentation.

3.4 Forward Referenced Blocks

Wyvern makes extensive use of forward referenced blocks to make its syntax clean. In particular, layout-delimited TSLs, the general-purpose introductory form for object types and the elimination form for case types and product types all use forward referenced blocks. Fig. 6 shows all of these in use (assuming suitable definitions of casetypes `Nat` and `HTML`, not included). In the grammar, note particularly the rules for **`let`** and that inline literals, even those containing nested expressions with forward references, can be treated as expressions not containing forward references – *in the initial phase of parsing, before typechecking commences, all literal forms are left unparsed.*

```
(* programs *)
p → 'objtype'= ID> NEWLINE> objdecls> NEWLINE> metadatadecl> p=
p → 'casetype'= ID> NEWLINE> casedecls> NEWLINE> metadatadecl> p=
p → e=

metadatadecl → ε | 'metadata'= '='> e> NEWLINE>

(* expressions *)
e → ē=
e → ẽ['~']= NEWLINE> chars>
e → ẽ['new']= NEWLINE> d>
e → ẽ['case(' ē ')']= NEWLINE> c>

(* object definitions *)
d → ε
d → 'val'= ID> ':'> type> '='> e> NEWLINE> d=
d → 'def'= ID> '('> argsig> ')'> ':'> type> '='> e> NEWLINE> d=

(* cases *)
c → cc | cp
cc → ID= '('> ID> ')'> '=>'> e>
cc → ID= '('> ID> ')'> '=>'> e> NEWLINE> cs=
cp → '('= ID> ','> ID> ')'> '=>'> e>

(* expressions not containing forward references *)
ē → ID=
ē → 'fn'= ID> ':'> type> '=>'> ē>
ē → ē= '('> āl> ')'>
ē → '('= ē> ','> ē> ')'>
ē → 'let'= ID> ':'> type> '='> e> NEWLINE> ē=
ē → ē= '.'> ID>
ē → type= '.'> ID> '('> ē> ')'>
ē → ē= ':'> type>
ē → 'valAST'= '('> ē> ')'>
ē → type= '.'> 'metadata'>
ē → inlinelit=

āl → ε | āl_nonempty=
āl_nonempty → ē= | ē= ','> āl_nonempty>

inlinelit → chars1['‘']= | chars1['’']= | chars1['"']= | ...
inlinelit → chars2['{', '}']= | chars2['<', '>']= | chars2['[', ']']= | ...
inlinelit → numlit=

(* expressions containing exactly one forward reference *)
ẽ[fwd] → fwd=
ẽ[fwd] → 'fn'= ID> ':'> type> '=>'> ẽ[fwd]>
ẽ[fwd] → ẽ[fwd]= '('> āl> ')'>
ẽ[fwd] → '('= ē> ','> ẽ[fwd]> ')'>
ẽ[fwd] → '('= ẽ[fwd]> ','> ē> ')'>
ẽ[fwd] → 'let'= ID> ':'> type> '='> e> NEWLINE> ẽ[fwd]=
ẽ[fwd] → ē= '('> ãl[fwd]> ')'>
ẽ[fwd] → ẽ[fwd]= '.'> ID>
ẽ[fwd] → type= '.'> ID> '('> ẽ[fwd]> ')'>
ẽ[fwd] → ẽ[fwd]= ':'> type>
ẽ[fwd] → 'valAST'= '('> ẽ[fwd]> ')'>

ãl[fwd] → ẽ[fwd]= | ẽ[fwd]= ','> āl_nonempty> | ē= ','> ãl[fwd]>
```

Fig. 5: Concrete Syntax (a few simple productions have been omitted)

```
1  objtype T
2    val y : HTML
3  let page : HTML->HTML = fn x:HTML => ~
4    :html
5      :body
6        {x}
7  page(case(5 : Nat))
8    Z(_) => (new : T).y
9      val y : HTML = ~
10       :h1 Zero!
11   S(x) => ~
12     :h1 Successor!
```

$$
\begin{array}{l}
\textbf{objtype}\ T\ \{ \\
\quad \textbf{val}\ y : HTML, \\
\quad \textbf{metadata} = (\textbf{new}\ \{\}) : Unit\ \}; \\
(\boldsymbol{\lambda} page : HTML \to HTML. \\
\ page(\textbf{case}(\lfloor 5 \rfloor : Nat)\ \{ \\
\quad\ Z(\_) \Rightarrow ((\textbf{new}\ \{ \\
\quad\quad \textbf{val}\ y : HTML = \lfloor : h1\ Z! \rfloor\}) : T).y \\
\quad\ |S(x) \Rightarrow \lfloor : h1\ S! \rfloor\})) \\
\quad\ (\boldsymbol{\lambda} x : HTML.\ \lfloor : html \\
\quad : body \\
\quad\ \{x\} \rfloor)
\end{array}
$$

Fig. 6: An example Wyvern program demonstrating forward references. The corresponding abstract syntax, where forward references are inlined, is on the right.

3.5 Abstract Syntax

The concrete syntax of a Wyvern program, `p`, is parsed to produce a program in the abstract syntax, ρ, shown on the left side of Fig. 7. Forward references are internalized. In particular, note that all literal forms are unified into the abstract literal form $\lfloor body \rfloor$, including the layout-delimited form and number literals. The abstract syntax contains a form, $\textbf{fromTS}(e)$, that has no analog in the concrete syntax. This will be used internally to ensure hygiene, as we will discuss in the next section.

4 Bidirectional Typechecking and Literal Rewriting

We will now specify a type system for the abstract syntax in Fig. 7. Conventional type systems are specified using a typechecking judgement like $\Delta; \Gamma \vdash e : \tau$, where the variable context, Γ, tracks the types of variables, and the type context, Δ, tracks types and their signatures. However, this conventional formulation does not separately consider how, when deriving this judgement, it will be considered algorithmically – will a type be provided, so that we simply need to check e against it, or do we need to synthesize a type for e? For our system, this distinction is crucial: a generic literal can only be used in the first situation.

Bidirectional type systems, as presented by Lovas and Pfenning [20], make this distinction clear by specifying the type system instead using two simultaneously defined judgements: one for expressions that can *synthesize* a type based on the surrounding context (e.g. variables and elimination forms), and another for expressions for which we know what type to *check* or *analyze* the term against (e.g. generic literals and some introductory forms). Our work builds upon this work, making the following core additions: the tye context Δ now tracks the metadata in addition to type signatures, and as we typecheck, we need to also perform literal rewriting by calling the parser associated with the type that a literal is being analyzed against, typechecking the AST it produces and ensuring that hygiene is maintained.

$$
\begin{array}{lll}
\rho ::= \mathbf{objtype}[T, \omega, e]; \rho & \omega ::= \emptyset \mid \mathbf{membdecl}[\ell, \tau]; \omega & \\
\quad \mid \mathbf{casetype}[T, \chi, e]; \rho & \chi ::= \emptyset \mid \mathbf{casedecl}[C, \tau]; \chi & \\
\quad \mid \mathbf{expr}[e] & \tau ::= \mathbf{named}[T] \mid \mathbf{arrow}[\tau, \tau] &
\end{array}
$$

$$
\begin{array}{lll}
e ::= x & \hat{\imath} ::= x & i ::= x \\
\quad \mid \mathbf{easc}[\tau](e) & \quad \mid \mathbf{hasc}[\tau](\hat{\imath}) & \quad \mid \mathbf{iasc}[\tau](i) \\
\quad \mid \mathbf{elam}(x.e) & \quad \mid \mathbf{hlam}(x.\hat{\imath}) & \quad \mid \mathbf{ilam}(x.i) \\
\quad \mid \mathbf{eap}(e; e) & \quad \mid \mathbf{hap}(\hat{\imath}; \hat{\imath}) & \quad \mid \mathbf{iap}(i; i) \\
\quad \mid \mathbf{enew}\,\{m\} & \quad \mid \mathbf{hnew}\,\{\hat{d}\} & \quad \mid \mathbf{inew}\,\{\dot{d}\} \\
\quad \mid \mathbf{eprj}[\ell](e) & \quad \mid \mathbf{hprj}[\ell](\hat{\imath}) & \quad \mid \mathbf{iprj}[\ell](i) \\
\quad \mid \mathbf{einj}[C](e) & \quad \mid \mathbf{hinj}[C](\hat{\imath}) & \quad \mid \mathbf{iinj}[C](i) \\
\quad \mid \mathbf{ecase}(e)\,\{r\} & \quad \mid \mathbf{hcase}(\hat{\imath})\,\{\hat{r}\} & \quad \mid \mathbf{icase}(i)\,\{\dot{r}\} \\
\quad \mid \mathbf{etoast}(e) & \quad \mid \mathbf{htoast}(\hat{\imath}) & \quad \mid \mathbf{itoast}(i) \\
\quad \mid \mathbf{emetadata}[T] & \quad \mid \mathbf{hmetadata}[T] & \quad \mid \mathbf{imetadata}[T] \\
\quad \mid \mathbf{lit}[body] & \quad \mid \mathbf{spliced}[e] &
\end{array}
$$

$$
\begin{array}{lll}
d ::= \emptyset & \hat{d} ::= \emptyset & \dot{d} ::= \emptyset \\
\quad \mid \mathbf{eval}[\ell](e); d & \quad \mid \mathbf{hval}[\ell](\hat{\imath}); \hat{d} & \quad \mid \mathbf{ival}[\ell](i); \dot{d} \\
\quad \mid \mathbf{edef}[\ell](x.e); d & \quad \mid \mathbf{hdef}[\ell](x.\hat{\imath}); \hat{d} & \quad \mid \mathbf{idef}[\ell](x.i); \dot{d} \\
r ::= \emptyset & \hat{r} ::= \emptyset & \dot{r} ::= \emptyset \\
\quad \mid \mathbf{erule}[C](x.e); r & \quad \mid \mathbf{hrule}[C](x.\hat{\imath}); \hat{r} & \quad \mid \mathbf{irule}[C](x.i); \dot{r}
\end{array}
$$

Fig. 7: Abstract Syntax of TSL Wyvern. T ranges over type name, ℓ ranges over object member labels, C ranges over case labels, x ranges over variables and *body* ranges over literal bodies. Tuples can be introduced as derived forms: $(i_1, i_2) := \mathbf{inew}\,\{\mathbf{ival}[\ell_1](i_1); \mathbf{ival}[\ell_2](i_2)\}$.

The judgement $\boxed{\Delta; \Gamma'; \Gamma \vdash e \Rightarrow \tau \rightsquigarrow \hat{e}}$ means that from the type context Δ, the *surrounding variable context*, Γ', and the *local variable context*, Γ, we synthesize the type τ for e and rewrite it to $\hat{e}$ (which does not contain literals or the special form $\mathbf{fromTS}(e)$, which makes the surrounding context available; see below). The judgement $\boxed{\Delta; \Gamma'; \Gamma \vdash e \Leftarrow \tau \rightsquigarrow \hat{e}}$ similarly means that we check e against the type τ and the expression e is rewritten into the expression $\hat{e}$. The forms of Γ and Δ are given in Fig. 7. Note that Δ carries the type's signature as well as the rewritten form of the metadata. The rules for these judgements, as well as key rules for several auxiliary judgements that are needed in their premises, are given in Figs. 8-11.

These rules assume that a collection of built-in types are included by default at the top of programs (e.g. `Unit`, `Parser`, `Exp` already mentioned, and a few others), captured by an intial type context Δ_0. We show the concrete syntax for the two key ones in Fig. 7. The static semantics and the dynamic semantics (defined for $\hat{e}$ only) are that of a conventional functional language with functions, inductive datatypes, products and records with the addition of a few new forms. The key new dynamic semantics rules are described in Figs. 12 and 13. We will now describe how some of the novel rules that support TSLs work below. We refer the reader to [20] and texts on type systems, e.g. [15, 26], for the remainder.

```
objtype Parser
  def parse(ts : TokenStream) : (Exp *
    TokenStream)
  metadata = new
    val parser : Parser = new
      val parse(ts : TokenStream) : (
          Exp * TokenStream) =
        (* parser generator based
           on Adams' formalism *)
```

```
casetype Exp
    Var of ID
  | Lam of ID * Type * Exp
  | App of Exp * Exp
  ...
  | FromTS of Exp * Exp
  | Literal of TokenStream
  | Error of ErrorMessage
  metadata = (* quasiquotes *)
```

Fig. 8: Two of the built-in types included in Δ_0 (concrete syntax).

$\boxed{\vdash_\Sigma \rho \rightsquigarrow i}$ $\quad \Sigma ::= \emptyset \mid \Sigma, T[\delta, \mu] \quad \delta ::= \ ? \mid \omega \mid \chi \quad \mu ::= \ ? \mid i : \tau$

$$\frac{T \notin \mathrm{dom}(\Sigma) \quad \vdash_{\Sigma, T[?,?]} \omega \quad \emptyset \vdash_{\Sigma, T[\omega,?]} e_m \rightsquigarrow i_m \Rightarrow \tau_m \quad \vdash_{\Sigma, T[\omega, i_m : \tau_m]} \rho \rightsquigarrow i}{\vdash_\Sigma \mathbf{objtype}[T, \omega, e_m]; \rho \rightsquigarrow i} \ \textit{P-OT}$$

$$\frac{T \notin \mathrm{dom}(\Sigma) \quad \vdash_{\Sigma, T[?,?]} \chi \quad \emptyset \vdash_{\Sigma, T[\chi,?]} e_m \rightsquigarrow i'_m \Rightarrow \tau_m \quad \vdash_{\Sigma, T[\chi, i'_m : \tau_m]} \rho \rightsquigarrow i_m}{\vdash_\Sigma \mathbf{casetype}[T, \chi, e_m]; \rho \rightsquigarrow i} \ \textit{P-CT}$$

$$\frac{\emptyset \vdash_\Sigma e \rightsquigarrow i \Rightarrow \tau}{\vdash_\Sigma \mathbf{expr}[e] \rightsquigarrow i} \ \textit{P-E}$$

$\boxed{\vdash_\Sigma \omega}$

$$\frac{}{\vdash_\Sigma \emptyset} \ \textit{M-emp} \qquad \frac{\ell \notin \mathrm{dom}(\omega) \quad \vdash_\Sigma \tau \quad \vdash_\Sigma \omega}{\vdash_\Sigma \mathbf{membdecl}[\ell, \tau]; \omega} \ \textit{M-decl}$$

$\boxed{\vdash_\Sigma \chi}$

$$\frac{}{\vdash_\Sigma \emptyset} \ \textit{C-emp} \qquad \frac{C \notin \mathrm{dom}(\chi) \quad \vdash_\Sigma \tau \quad \vdash_\Sigma \chi}{\vdash_\Sigma \mathbf{casedecl}[C, \tau]; \chi} \ \textit{C-decl}$$

Fig. 9: Statics for programs, ρ. Note that type declarations can be recursive but not mutually recursive as written. Our prelude, Σ_0, requires mutually recursive type definitions, and we show how to add this in the technical report.

4.1 Defining a TSL Manually

In the example in Fig. 3, we showed a TSL being defined using a parser generator based on Adams' formalism. A parser generator is itself merely a TSL for a parser, and a parser is the fundamental construct that becomes associated with a type to form a TSL. The signature for the built-in type `Parser` is shown in Fig. 7. It is an object type with a `parse` function taking in a `TokenStream` and producing an AST of a Wyvern expression, which is of type `Exp`. This built-in type is shown also in Fig. 7. Note that there is a form for each form in the abstract syntax, e, as well as an `Error` form for indicating error messages (in the theory, nothing is done with these messages). As previously mentioned, quasiquotes are merely a TSL that allows one to construct the abstract syntax, represented as this case type, using concrete syntax, with the addition of an unquote mechanism.

The `parse` function for a type t is called when checking a literal form against that type. This is seen in the key rule of our statics: *T-lit*, in Fig. 9. The premises of these rules operate as follows:

1. This rule uses some built-in types. We first ensure they are available.

$\boxed{\Gamma \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau}$ $\boxed{\Gamma \vdash_\Sigma e \rightsquigarrow i \Rightarrow \tau}$ $\quad \Gamma ::= \emptyset \mid \Gamma, x : \tau$

$$\frac{\Gamma \vdash_\Sigma e \rightsquigarrow i \Rightarrow \tau}{\Gamma \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau}\ \textit{T-syn-to-ana} \qquad \frac{\vdash_\Sigma \tau \quad \Gamma \vdash_\Sigma e \Leftarrow \tau \rightsquigarrow i}{\Gamma \vdash_\Sigma \mathbf{easc}[\tau](e) \rightsquigarrow \mathbf{iasc}[\tau](i) \Rightarrow \tau}\ \textit{T-asc}$$

$$\frac{x : \tau \in \Gamma}{\Gamma \vdash_\Sigma x \rightsquigarrow x \Rightarrow \tau}\ \textit{T-var} \qquad \frac{\Gamma, x : \tau_1 \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau_2}{\Gamma \vdash_\Sigma \mathbf{elam}(x.e) \rightsquigarrow \mathbf{ilam}(x.i) \Leftarrow \mathbf{arrow}[\tau_1, \tau_2]}\ \textit{T-abs}$$

$$\frac{\Gamma \vdash_\Sigma e_1 \rightsquigarrow i_1 \Rightarrow \tau_1 \to \tau_2 \quad \Gamma \vdash_\Sigma e_2 \rightsquigarrow i_2 \Leftarrow \tau_1}{\Gamma \vdash_\Sigma \mathbf{eap}(e_1; e_2) \rightsquigarrow \mathbf{iap}(i_1; i_2) \Rightarrow \tau_2}\ \textit{T-ap}$$

$$\frac{T \neq \mathit{ParseStream} \quad T[\omega, \mu] \in \Sigma \quad \Gamma \vdash^t_\Sigma d \rightsquigarrow \dot{d} \Leftarrow \omega}{\Gamma \vdash_\Sigma \mathbf{enew}\,\{d\} \rightsquigarrow \mathbf{inew}\,\{\dot{d}\} \Leftarrow \mathbf{named}[T]}\ \textit{T-obj-intro}$$

$$\frac{\Gamma \vdash_\Sigma e \rightsquigarrow i \Rightarrow \mathbf{named}[T] \quad T[\omega, \mu] \in \Sigma \quad \mathbf{membdecl}[\ell, \tau] \in \omega}{\Gamma \vdash_\Sigma \mathbf{eprj}[\ell](e) \rightsquigarrow \mathbf{iprj}[\ell](i) \Rightarrow \tau}\ \textit{T-obj-elim}$$

$$\frac{T[\chi, \mu] \in \Sigma \quad \mathbf{casedecl}[C, \tau] \in \chi \quad \Gamma \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau}{\Gamma \vdash_\Sigma \mathbf{einj}[C](e) \rightsquigarrow \mathbf{iinj}[C](i) \Leftarrow \mathbf{named}[T]}\ \textit{T-case-intro}$$

$$\frac{\Gamma \vdash_\Sigma e \rightsquigarrow i \Rightarrow \mathbf{named}[T] \quad T[\chi, \mu] \in \Sigma \quad \Gamma \vdash_\Sigma r \rightsquigarrow \dot{r} \Leftarrow \chi \Rightarrow \tau}{\Gamma \vdash_\Sigma \mathbf{ecase}(e)\,\{r\} \rightsquigarrow \mathbf{icase}(i)\,\{\dot{r}\} \Rightarrow \tau}\ \textit{T-case-elim}$$

$$\frac{\Sigma_0 \subset \Sigma \quad \Gamma \vdash_\Sigma e \rightsquigarrow i \Rightarrow \tau}{\Gamma \vdash_\Sigma \mathbf{etoast}(e) \rightsquigarrow \mathbf{itoast}(i) \Rightarrow \mathbf{named}[\mathit{Exp}]}\ \textit{T-toast}$$

$$\frac{T[\delta, i : \tau] \in \Sigma}{\Gamma \vdash_\Sigma \mathbf{emetadata}[T] \rightsquigarrow \mathbf{imetadata}[T] \Rightarrow \tau}\ \textit{T-metadata}$$

$\boxed{\Gamma \vdash^T_\Sigma d \rightsquigarrow \dot{d} \Leftarrow \omega}$

$$\frac{}{\Gamma \vdash^T_\Sigma \emptyset \rightsquigarrow \emptyset \Leftarrow \emptyset}\ \textit{T-unit}$$

$$\frac{\Gamma \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau \quad \Gamma \vdash^T_\Sigma d \rightsquigarrow \dot{d} \Leftarrow \omega}{\Gamma \vdash^T_\Sigma \mathbf{eval}[\ell](e); d \rightsquigarrow \mathbf{ival}[\ell](i); \dot{d} \Leftarrow \mathbf{membdecl}[\ell, \tau]; \omega}\ \textit{T-val}$$

$$\frac{\Gamma, x : \mathbf{named}[T] \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau}{\Gamma \vdash^T_\Sigma \mathbf{edef}[\ell](x.e) \rightsquigarrow \mathbf{idef}[\ell](x.i) \Leftarrow \mathbf{membdecl}[\ell, \tau]}\ \textit{T-def}$$

$\boxed{\Gamma \vdash_\Sigma r \rightsquigarrow \dot{r} \Leftarrow \chi \Rightarrow \tau}$

$$\frac{}{\Gamma \vdash_\Sigma \emptyset \rightsquigarrow \emptyset \Leftarrow \emptyset \Rightarrow \tau}\ \textit{T-void}$$

$$\frac{\Gamma, x : \tau_1 \vdash_\Sigma e \rightsquigarrow i \Rightarrow \tau_2 \quad \Gamma \vdash_\Sigma r \rightsquigarrow \dot{r} \Leftarrow \chi \Rightarrow \tau_2}{\Gamma \vdash_\Sigma \mathbf{erule}[C](x.e); r \rightsquigarrow \mathbf{irule}[C](x.i); \dot{r} \Leftarrow \mathbf{casedecl}[C, \tau_1]; \chi \Rightarrow \tau_2}\ \textit{T-rule}$$

Fig. 10: Statics for external terms, e. The rule for literals is shown in Fig. 11.

$$\frac{\begin{array}{c}\Sigma_0 \subset \Sigma \quad T[\delta, i_m : HasTSL] \in \Sigma \quad \texttt{parsestream(}body\texttt{)=}i_{ps} \\ i_m.parser.parse(i_{ps}) \Downarrow_\Sigma i_{ast} \quad i_{ast} \uparrow \hat{\imath} \quad \Gamma; \emptyset \vdash_\Sigma \hat{\imath} \rightsquigarrow i \Leftarrow \mathbf{named}[T]\end{array}}{\Gamma \vdash_\Sigma \mathbf{lit}[body] \rightsquigarrow i \Leftarrow \mathbf{named}[T]}\ \textit{T-lit}$$

Fig. 11: Statics for external terms, e, continued. This is the key rule (see text).

$\boxed{\Gamma; \Gamma \vdash_\Sigma \hat{\imath} \rightsquigarrow i \Leftarrow \tau}$ $\boxed{\Gamma; \Gamma \vdash_\Sigma \hat{\imath} \rightsquigarrow i \Rightarrow \tau}$

$$\frac{x : \tau \in \Gamma}{\Gamma_o; \Gamma \vdash_\Sigma x \rightsquigarrow x \Rightarrow \tau}\ \textit{H-var} \qquad \frac{\Gamma_o; \Gamma, x : \tau_1 \vdash_\Sigma \hat{\imath} \rightsquigarrow i \Leftarrow \tau_2}{\Gamma_o; \Gamma \vdash_\Sigma \mathbf{hlam}(x.\hat{\imath}) \rightsquigarrow \mathbf{ilam}(x.i) \Leftarrow \mathbf{arrow}[\tau_1, \tau_2]}\ \textit{H-abs}$$

$$\ldots$$

$$\frac{\Gamma_o \vdash_\Sigma e \rightsquigarrow i \Leftarrow \tau}{\Gamma_o; \Gamma \vdash_\Sigma \mathbf{spliced}[e] \rightsquigarrow i \Leftarrow \tau}\ \textit{H-spl-A} \qquad \frac{\Gamma_o \vdash_\Sigma e \rightsquigarrow i \Rightarrow \tau}{\Gamma_o; \Gamma \vdash_\Sigma \mathbf{spliced}[e] \rightsquigarrow i \Rightarrow \tau}\ \textit{H-spl-S}$$

Fig. 12: Statics for translational internal terms, $\hat{\imath}$. Each rule in Fig. 10 corresponds to an analagous rule here by threading the outer context through opaquely (e.g. the rules for variables and functions, shown here). The outer context is only used by the rules for $\mathbf{spliced}[e]$, representing external terms spliced into TSL bodies. Only these terms can access outer variables, achieving hygiene (see text).

2. A well-typed, rewritten parser object is extracted from the type's metadata. This is the step where the parser generator rewrites a grammar to a parse method, recursively using the TSL mechanism itself.
3. A tokenstream, of type `TokenStream`, is generated from the body of the literal. This type is an object that that allows the reading of tokens, as well as an additional method discussed in the next section for parsing the stream as a Wyvern expression.
4. The `parse` method is called with this extracted tokenstream to produce a syntax tree and a remaining tokenstream.
5. The syntax tree, $\hat{e}'$ is *dereified* into its corresponding term, e (the hat is gone because the generated syntax tree might itself use TSLs). This is the only way terms of the form $\mathbf{fromTS}(e)$ can be generated (see below).
6. The dereified term is then recursively typechecked against the same type and rewritten, consistent with the semantics of TSLs as we have been describing them – they must produce a term of the type they are being checked against. It is checked under the empty local context to ensure hygiene (below).
7. The TSL must consume the entire token stream, so this is checked.

4.2 Hygiene

A concern with any term rewriting system is *hygiene* – how should variables in the generated AST be bound? In particular, if the rewriting system generates an *open term*, then it is making assumptions about the names of variables in scope at the site where the TSL is being used, which is incorrect - those variables should only be identifiable up to alpha renaming. Only the *user* of a TSL knows which variables are in scope. Strict hygiene would simply reject all open terms,

$\boxed{i \uparrow \hat{i}}$

$$\frac{i_{id} \uparrow x}{\mathbf{iinj}[Var](i_{id}) \uparrow x}\ \textit{U-Var}$$

$$\frac{i_1 \uparrow \tau \quad i_2 \uparrow \hat{i}}{\mathbf{iinj}[Asc]((i_1, i_2)) \uparrow \mathbf{hasc}[\tau](\hat{i})}\ \textit{U-Asc}$$

$$\frac{i_{id} \uparrow x \quad i \uparrow \hat{i}}{\mathbf{iinj}[Lam]((i_{id}, i)) \uparrow \mathbf{hlam}(x.\hat{i})}\ \textit{U-Lam}$$

$$\frac{i_1 \uparrow \hat{i}_1 \quad i_2 \uparrow \hat{i}_2}{\mathbf{iinj}[Ap]((i_1, i_2)) \uparrow \mathbf{hap}(\hat{i}_1, \hat{i}_2)}\ \textit{U-Ap}$$

$$\cdots$$

$$\frac{\texttt{body}(i_{ps})\texttt{=}body \quad \texttt{eparse}(body)\texttt{=}e}{\mathbf{iinj}[Spliced](i_{ps}) \uparrow \mathbf{spliced}[e]}\ \textit{U-Spliced}$$

$\boxed{i \uparrow \tau}$

$$\frac{i_{name} \uparrow T}{\mathbf{iinj}[Named](i_{name}) \uparrow \mathbf{named}[T]}\ \textit{U-T}$$

$$\frac{i_1 \uparrow \tau_1 \quad i_2 \uparrow \tau_2}{\mathbf{iinj}[Arrow]((i_1, i_2)) \uparrow \mathbf{arrow}[\tau_1, \tau_2]}\ \textit{U-A}$$

Fig. 13: Dereification rules, used by rule *T-lit* (above) to determine the translational internal term corresponding to the internal term of type **named**[*Exp*] returned by the *parse* method.

$\boxed{i \downarrow i}$

$$\frac{x \downarrow i_{id}}{x \downarrow \mathbf{iinj}[Var](i_{id})}\ \textit{R-Var}$$

$$\frac{\tau \downarrow i_1 \quad i \downarrow i_2}{\mathbf{iasc}[\tau](i) \downarrow \mathbf{iinj}[Asc]((i_1, i_2))}\ \textit{R-Asc}$$

$$\frac{x \downarrow i_{id} \quad i \downarrow i'}{\mathbf{ilam}(x.i) \downarrow \mathbf{iinj}[Lam]((i_{id}, i'))}\ \textit{R-Lam}$$

$$\frac{i_1 \downarrow i'_1 \quad i_2 \downarrow i'_2}{\mathbf{iap}(i_1; i_2) \downarrow \mathbf{iinj}[Ap]((i'_1, i_2))}\ \textit{R-Ap}$$

$$\cdots$$

$\boxed{\tau \downarrow i}$

$$\frac{T \downarrow i_{name}}{\mathbf{named}[T] \downarrow \mathbf{iinj}[Named](i_{name})}\ \textit{R-T}$$

$$\frac{\tau_1 \downarrow i_1 \quad \tau_2 \downarrow i_2}{\mathbf{arrow}[\tau_1, \tau_2] \downarrow \mathbf{iinj}[Arrow]((i_1, i_2))}\ \textit{R-A}$$

Fig. 14: Reification rules, used by the **itoast** ("to AST") operator (Fig. 16) to permit generating an internal term of type **named**[*Exp*] corresponding to the value of the argument (a form of serialization). This serves to make writing TSLs convenient.

but this would prevent even nested Wyvern expressions which the user provided from referring to local variables.

The solution to being able to capture variables in portions of the tokenstream that are parsed as Wyvern only is to add a new term to the abstract syntax that has no corresponding form in the concrete syntax: **fromTS**(e). This means: "this is a term that was parsed from the user's tokenstream". It can be generated by calling `ts.as_wyv_exp(tok)`, which returns a the remaining tokenstream as well as the value `Exp.FromTS(ts, tok)`. When we dereify this term, we turn this into the form **fromTS**(e), where e is the result of parsing the tokenstream as a Wyvern expression until an unexpected token `tok` appears.

When we attempt to typecheck this form, which will be starting from an empty local variable context by moving all the available variables into the surrounding variable context (in the *T-Lit* rule, above), we add in the bindings available in the surrounding variable context (to any that were introduced by the TSL, such as the TSL for `Parser` does with named non-terminals). In other words, variables in the surrounding variable context can only be used within a term of the form **fromTS**(e). These variables are precisely those that only the user can know exist, but not the extension.

$$\boxed{\Gamma \vdash_\Sigma i \Leftarrow \tau} \quad \boxed{\Gamma \vdash_\Sigma i \Rightarrow \tau} \quad \cdots \quad \frac{T[\omega, \mu] \in \Sigma \quad \Gamma \vdash^T_\Sigma \dot{d} \Leftarrow \omega}{\Gamma \vdash_\Sigma \mathbf{inew}\,\{\dot{d}\} \Leftarrow \mathbf{named}[T]}\ \textit{IT-obj-intro}$$

Fig. 15: Statics for internal terms, i. Each rule in Fig. 10 corresponds to an analogous rule here by removing the elaboration portion. Only the rule for object introduction differs, in that it does not restrict the introduction of parse streams.

$$\boxed{i \underset{\Sigma}{\longmapsto} i} \quad \cdots$$

$$\frac{T[\delta, i : \tau] \in \Sigma}{\mathbf{imetadata}[T] \underset{\Sigma}{\longmapsto} i}\ \textit{D1} \qquad \frac{i \underset{\Sigma}{\longmapsto} i'}{\mathbf{itoast}(i) \underset{\Sigma}{\longmapsto} \mathbf{itoast}(i')}\ \textit{D2} \qquad \frac{i\ \texttt{val} \quad i \downarrow i'}{\mathbf{itoast}(i) \underset{\Sigma}{\longmapsto} i'}\ \textit{D3}$$

Fig. 16: Dynamics for internal terms, i. Only internal terms have a dynamic semantics. Most constructs in TSL Wyvern are standard and omitted, as our focus in this paper is on the statics (see [15] Chs. XXXX). The rules for the operators for extracting metadata and extracting an AST from a value are shown.

For this mechanism to truly ensure hygiene, one must not be able to sidestep it by generating a tokenstream manually: expressions from a tokenstream must have actually come from the use site. This is ensured by preventing users from checking `new` against `TokenStream` in the statics.

A second facet of hygiene is being able to refer to local variables available within the parser itself, such as local helper functions, for convenience. This can be done using the primitive $\mathbf{valAST}(e)$. The semantics for this, shown in Fig. 13, first evaluate e to a value, then *reify* this value to an AST. This can be used to "bake in" a value known at compile time into the generated code safely. The rules for reification, used here, and dereification, used in the literal rule described above, are essentially dual, as seen in Figs. 11 and 12.

4.3 Safety

The semantics we have defined constitute a type safe language.

We begin with a lemma that shows that the statics for e and $\hat{e}$ are consistent. This makes us sure that the splitting of variable contexts to maintain hygiene was done correctly (because they can be brought back together at the end).

Lemma 1 (Forward Consistency).

1. *If* $\vdash \Delta$ *and* $\Delta \vdash \Gamma'$ *and* $\Delta \vdash \Gamma$ *and* $\Delta; \Gamma'; \Gamma \vdash e \Leftarrow \tau \rightsquigarrow \hat{e}$ *then* $\Delta; \Gamma', \Gamma \vdash \hat{e} : \tau$.
2. *If* $\vdash \Delta$ *and* $\Delta \vdash \Gamma'$ *and* $\Delta \vdash \Gamma$ *and* $\Delta; \Gamma'; \Gamma \vdash e \Rightarrow \tau \rightsquigarrow \hat{e}$ *then* $\Delta; \Gamma', \Gamma \vdash \hat{e} : \tau$.

Proof. Forward consistency is easily seen by observing that for each form shared by both e and $\hat{e}$, the bidirectional system simply rewrites to the corresponding form of $\hat{e}$ recursively. Thus, these cases are direct applications of the IH. For the literal form, we can apply the IH to arrive at the fact that $\Delta_0, \Delta; \Gamma', \Gamma, \emptyset \vdash \hat{e} : t$ which by congruence (removing the empty context at the end) is what we wish to show. Similarly, for the form $\mathbf{fromTS}(e)$ we have by the IH that $\Delta; \emptyset, \Gamma', \Gamma \vdash \hat{e} : \tau$ which again implies what we wish to show by simple congruence of contexts.

We then need to show type safety of $\hat{e}$. Because it doesn't contain any non-standard terms other than **valAST** and t.**metadata**, both of which have straightforward semantics (Fig. 13), this follows by the standard progress and preservation techniques. The only tricky case is *Dyn-valAST2*, which requires the following straightforward lemma about the reification rules in Fig. 12, as well as standard structural properties for the contexts (weakening; not shown).

Lemma 2 (Reification). *If* $\hat{e} \triangleright \hat{e}'$ *then* $\Delta_0; \emptyset \vdash \hat{e}' : Exp$.

Lemma 3 (Preservation). *If* $\vdash \Delta$ *and* $\Delta; \emptyset \vdash \hat{e} : \tau$ *and* $\hat{e} \underset{\Delta}{\longmapsto} \hat{e}'$ *then* $\Delta; \emptyset \vdash \hat{e}' : \tau$.

Lemma 4 (Progress). *If* $\vdash \Delta$ *and* $\Delta; \emptyset \vdash \hat{e} : \tau$ *then either* $\hat{e}$ `val` *or* $\hat{e} \underset{\Delta}{\longmapsto} \hat{e}'$.

These lemmas and associated judgements can be lifted to the level of programs by applying them to the top-level expression the program contains (simple, not shown). As a result, we have type safety: well-typed programs cannot "get stuck".

Theorem 1 (Type Safety). *If* $\Delta_0 \vdash \rho : \tau \rightsquigarrow \hat{\rho}$ *then* $\Delta_0 \vdash \hat{\rho} : \tau$ *and either* $\hat{\rho}$ `val` *or* $\hat{\rho} \underset{\Delta_0}{\longmapsto} \hat{\rho}'$ *such that* $\Delta_0 \vdash \hat{\rho}' : \tau$.

4.4 Decidability

Because we are executing user-defined parsers during typechecking, we do not have a straightforward statement of decidability (i.e. termination) of typechecking. The parser might not terminate, or it might generate a term that contains itself. Non-decidability is strictly due to user-defined parsing code. Typechecking of programs that do not contain literals is guaranteed to terminate, as is typechecking of $\hat{e}$ (which we do not actually need to do in practice by Lemma 1). Termination of parsers and parser generators has previously been studied (e.g. [18]) and the techniques can be applied to user-defined parsing code to increase confidence in termination. Few compilers, even those with high demands for correctness (e.g. CompCert [**?**]), have made it a priority to fully verify and prove termination of the parser. This is because it is perceived that most bugs in compilers arise due to incorrect optimization passes, not initial parsing and elaboration logic.

5 Corpus Analysis

An important question when introducing a new approach is how it would change the existing solutions and at what places it could be used. To answer these questions we performed a code analysis and tried to identify potential uses of the TSLs in the existing Java code. Our analysis is limited in scope and focused only on the class constructors. We are interested in class constructors because they are the programmatic constructs that potentially could be equipped with Wyvern types. For example, a Java class constructor such as

```
1  Path(String path) {...}
```

which takes in a single `String` and makes sure that it is of a specific format, could be equipped with a Wyvern type `Path` that would check for the format of the string. We examined this type of Java constructors in the first part of our analysis.

Further, having the pool of Java constructors we wanted to see in how many of them a TSL could be used. A place where we believe a TSL could be used is a `String` argument that has a specific format, for instance, constructors such as:

```
1  FileUpdatedEvent(Object source, String path) {...}
```

Here, the second argument `path`, which is of type `String`, could be represented using a Wyvern type `Path` that would guarantee that the passed in argument is of the required format.

Methodology To perform our analysis, we used a recent version (20130901r) of 107 Java projects in the Qualitas Corpus [30] and searched for the two types of constructors described above. We used command line tools, such as `grep` and `sed`, and editor features such as search and substitution. In a semi-manual procedure, we scanned though the Java code and picked out class constructors. After that, we chose constructors that take at least one `String` as an argument, and looking at the names of the constructors and their arguments, we inferred the intended use of the classes associated with them.

Constructors	**Number**	**% of Total**
Total analyzed	124,873	100
Have a String argument	30,161	24
Could be equipped with a TSL	603	0.5
Could use a TSL	19,288	15

Table 1: Summary of the Analyses Results

Type of String	**Number**	**% of Total**
Name	14,307	74.2
ID	1,335	6.9
Directory path	823	4.3
Pattern	495	2.6
URL/URI	396	2.0
Other (zip code, password, query, HTML/XML, IP address, version, etc.)	1,932	10.0
Total:	**19,288**	**100.0**

Table 2: Types of `String` Arguments in Java Constructors

Results Our findings are summarized in Figure 1 and 2. For the first part of our analysis, looking through the Java constructors, we found that there is 0.5% (603 out of 124,873 examined constructors) which could be equipped with a TSL. Those constructors were used for classes that represent URLs and URIs, identification numbers, versions, directory paths, and various types of names (e.g., user name, database name, column name, etc.).

For the second part of our analysis, we found that there is 15% (19,288 out of 124,873 constructors) of constructors that could use a TSL. More details on the kinds of `String` arguments that are passed into constructors can be found in Table 2. The "Name" category refers to the name of a file, a user, a class, etc. that do not have to be unique; the "ID" category comprises process IDs, user IDs, column or row IDs, etc. that must have the uniqueness property; the "Pattern" category includes regular expressions, prefixes and suffixes, delimiters, format templates, etc.; the "Other" category contains `String`s used for ZIP codes, passwords, queries, IP addresses, versions, HTML and XML code, etc.; and the "Directory path" and "URL/URI" categories are self-explanatory.

Hence, our empirical study has shown that there is significant portion of Java constructors that have a potential of taking advantage of TSLs. It is important to keep in mind that our analysis was narrow: it focused exclusively on the constructors and thus forwent many other types of programming constructs, such as methods, variable assignments, etc., that could possibly also benefit from our approach.

6 Implementation

As of this writing, our implementation of the techniques described herein is ongoing based on a fork of the public Wyvern implementation. The Wyvern implementation is written in Java, based around a custom recursive-descent parser, with self-hosting left as future work. Our continued use of the Wyvern custom parser preceded our awareness of Adams' formalism (which does not have a public implementation currently). Thus, it is implemented with a stateful lexer as in Python, producing `INDENT` and `DEDENT` tokens. The token stream produced by the lexer is then passed into the Wyvern parser. When a language transition occurs, the Wyvern core parser extracts a substream from the current token stream, using either `INDENT` and `DEDENT` or any of the TSL delimiters to indicate where the substream should begin or end. This substream is then passed to the extension parser as an argument. By subdividing the token stream, the parsers can avoid complicated issues with delegation of responsibility caused by a single shared stream. We anticipate shifting to a more elegant system based on what we have specified here by the time the paper is presented. Some extension parsers are added though the interpreter's Java interoperability support.

In order to invoke the correct extension, we combine the typechecking and parsing stages of the compiler, so that typechecking happens incrementally as semantically distinct portions of the source are parsed. Once the first stage of parsing is complete and all Wyvern expressions are known, the Wyvern constructs are typechecked. Then, types for TSL blocks are inferred from the local type context, and the associated parsers are invoked in the next stage on the

substreams inside the TSL blocks. This process then continued recursively, as TSL blocks can contain Wyvern code that contains TSLs and etc., until all expressions have been parsed and typechecked. This current implementation is as described in this paper.

7 Discussion

Safe TSL Composition Our primary contribution is a strategy, where nesting of TSLs occurs by briefly entering the host language, that ensures that ambiguities cannot occur. The host language ensures that TSLs are delimited unambiguously, and the TSL ensures that the host language is delimited unambiguously. The body of the TSL is interpreted by a fixed grammar – the one associated with its expected type. This avoids the kinds of conflicts a simple merger of the grammars would cause. Apart from the large number of TSLs that can be composed together in a short piece of code while producing meaningful results, we aim to provide a safe composability guarantee that other language extension solutions do not [13, 17].

Keyword-Directed Invocation In most domain-specific language frameworks, a switch to a different language is indicated by a keyword or function call naming the language to be used. Wyvern eliminates this overhead in many cases by determining the TSL based on the expected type of an expression. This lightweight mechanism is particularly useful for small languages. Keyword-directed invocation is simply a special case of our type-directed approach. In particular, a keyword macro can be defined as a function with a single argument of a type specific to that keyword. The type contains the implementation of the domain-specific syntax associated with that keyword. In the most general sense, it may simply allow the entire Wyvern grammar, manipulating it in later phases of compilation.

As an example, consider control flow operators like `if`. This can be defined as a polymorphic method of the `bool` type with signature $(\texttt{unit} \to \alpha, \texttt{unit} \to \alpha) \to \alpha$. That is, it takes the two branches as functions and chooses which to invoke based on the value of the boolean, using perhaps a more primitive control flow operator, like case analysis, or even a Church encoding of booleans as functions. In Wyvern, the branches could be packaged together into a type, `IfBranches`, with an associated grammar that accepts the two branches as unwrapped expressions. Thus, `if` could be defined entirely in a library and used as follows:

```
1 if(guard, ~)
2   then
3     <any Wyvern>
4   else
5     <any Wyvern>
```

For methods like `if` where constructing the argument explicitly will almost never be done, it may be useful to mark the method in a way that allows Wyvern to assume it is being called with a TSL argument immediately following its use. This would eliminate the need for the `(~)` portion, supporting even more conventional notation.

Interaction with Subtyping The mechanism described here does not consider the case where multiple subtypes of a base type define a grammar. This can be resolved in several ways. We could require that only the *declared* type's grammar is used (if a subtype's grammar is desired, an explicit type annotation on the tilde can be used). Alternatively, we could attempt to parse against all relevant subtypes, only requiring explicit disambiguation when ambiguities arise. Wyvern does not currently support subtyping, so we leave this as future work.

Custom Lexers Our existing lexing strategy may be too restrictive, requiring all DSLs to be hierarchical in nature. One potential expansion would be to enable DSLs to define their own lexers, still perhaps delimited by indentation or parentheses. Such an extension would sacrifice some readability.

We do not allow a replacement parser for infix operators as we considered it to unnecessarily complicate the current prefixed parsing approach. In the future, we plan to further support redefining operators.

8 Related Work

[TODO: http://confluence.jetbrains.com/display/Kotlin/Type-safe+Groovy-style+builders] [TODO: AOSD submission] [TODO: staging parsers] Language macros are the most explored way of extending programming languages, with Scheme and other Lisp-style languages' hygienic macros being the 'gold standard.' In those languages, macros are written in the language itself and benefit from the simple syntax – parentheses universally serve as expression delimiters (although proposals for whitespace as a substitute for parentheses have been made [21]). Our work is inspired by this flexibility, but aims to support richer syntax as well as static types. Wyvern's use of types to trigger parsing avoids the overhead of needing to invoke macros explicitly by name and makes it easier to compose TSLs declaratively.

Another way to approach language extensibility is to go a level of abstraction above parsing as is done via metaprogramming facilities. For instance, OJ (previously, OpenJava) [29] provides a macro system based on a meta-object protocol, and Backstage Java [25], Template Haskell [28] and others employ compile-time meta-programming. Each of these systems provide macro-style rewriting of source code, but they provide at most limited extension of language parsing.

Other systems aim at providing forms of syntax extension that change the host language, as opposed to our whitespace-delimited approach. For example, Camlp4 [9] is a preprocessor for OCaml that offers the developer the ability to extend the concrete syntax of the language via the use of parsers and extensible grammars. SugarJ [11] takes a library-centric approach which supports syntactic extension of the Java language by adding libraries. In Wyvern, the core language is not extended directly, so conflicts cannot arise at link-time.

Scoping TSLs to expressions of a single type comes at the expense of some flexibility, but we believe that many uses of domain-specific languages are of this form already. A previous approach has considered type-based disambiguation of parse forests for supporting quotation and anti-quotation of arbitrary object languages [6]. Our work is similar in spirit, but does not rely on gener-

ation of parse forests and associates grammars with types, rather than types with grammar productions. We believe that this is a more simple and flexible methodology. C# expression trees [1] are similar in that, when the type of a term is `Expression<T->T'>`, it is parsed as a quotation. However, like the work just mentioned, this is *specifically* to support quotations. Our work supports quotations in addition to a variety of other work.

Many approaches to syntax extension, such as XJ [7] are keyword-directed in some form. We believe that a type-directed approach is more seamless and general, sacrificing a small amount of identifiability in some cases.

In terms of work on safe language composition, Schwerdfeger and van Wyk [27] proposed a solution that make strong safety guarantees provided that the languages comply with certain grammar restrictions, concerning first and follow sets of the host language and the added new languages. It also relied on strongly named entry tokens, like keyword delimited approaches. Our approach does not impose any such restrictions while still making safety guarantees.

Domain-specific language frameworks and language workbenches, such as Spoofax [16], Ensō [8] and others [17, 31], also provide a possible solution for the language extension task. They provide support for generating new programming languages and tooling in a modular manner. The Marco language [19] similarly provides macro definition at a level of abstraction that is largely independent of the target language. In these approaches, each TSL is *external* relative to the host language; in contrast, Wyvern focuses on extensibility *internal* to the language, improving interoperability and composability.

In addition, there is an ongoing work on projectional editors (e.g., [2, 10]) that use special graphical user interface to allow the developer to implicitly mark where the extensions are placed in the code, essential specifying directly the underlying ASTs. This solution to the language extension problem poses several challenges such as defining and implementing the semantics for the composition of the languages and the channels for communication between them. In Wyvern, we do not encounter these problems as the semantic rules for a language composition are incorporated within the host language by design.

There is a relation between recent work on Active Code Completion and our approach in that the Active Code Completion work associates code completion palettes with types [24] as well. Such palettes could be used for defining a TSL syntax for types. However, that syntax is immediately translated to Java syntax at edit time, while this work integrates with the core parsing facilities of the language.

9 Conclusion

In this paper, we described how extensible parsing in Wyvern makes for a solid platform to support whitespace-delimited, type-directed embedded DSLs or *Type-Specific Languages (TSLs)* for short. In the future, we aim to implement a wide variety of TSLs in Wyvern tweaking our approach and implementation thereof to provide a comprehensive example of supporting multiple interacting TSLs in a safe and easy-to-use manner.

References

1. Expression Trees (C# and Visual Basic). http://msdn.microsoft.com/en-us/library/bb397951.aspx.
2. JetBrains MPS – Meta Programming System. http://www.jetbrains.com/mps/.
3. OWASP Top 10 2013. https://www.owasp.org/index.php/Top_10_2013-Top_10, 2013.
4. M. D. Adams. Principled parsing for indentation-sensitive languages: Revisiting landin's offside rule. In *Proceedings of the 40th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '13, pages 511–522, New York, NY, USA, 2013. ACM.
5. M. Bravenboer, E. Dolstra, and E. Visser. Preventing injection attacks with syntax embeddings. In *Proceedings of the 6th International Conference on Generative Programming and Component Engineering*, GPCE '07, pages 3–12, New York, NY, USA, 2007. ACM.
6. M. Bravenboer, R. Vermaas, J. Vinju, and E. Visser. Generalized type-based disambiguation of meta programs with concrete object syntax. In *Generative Programming and Component Engineering*, 2005.
7. T. Clark, P. Sammut, and J. S. Willans. Beyond annotations: A proposal for extensible java (XJ). In *Source Code Analysis and Manipulation*, 2008.
8. W. R. Cook, A. Loh, and T. van der Storm. Ensō: A self-describing DSL workbench. http://enso-lang.org/.
9. D. de Rauglaudre. *Camlp4 - Reference Manual*, 2003.
10. L. Diekmann and L. Tratt. Parsing composed grammars with language boxes. In *Workshop on Scalable Language Specification*, 2013.
11. S. Erdweg, T. Rendel, C. Kästner, and K. Ostermann. SugarJ: library-based language extensibility. In *Object-Oriented Programming Systems, Languages, and Applications*, 2011.
12. S. Erdweg, T. Rendel, C. Kästner, and K. Ostermann. Sugarj: Library-based syntactic language extensibility. *ACM SIGPLAN Notices*, 46(10):391–406, 2011.
13. S. Erdweg and F. Rieger. A framework for extensible languages. In *Proceedings of the 12th International Conference on Generative Programming: Concepts & Experiences*, GPCE '13, pages 3–12, New York, NY, USA, 2013. ACM.
14. T. Green and M. Petre. Usability analysis of visual programming environments: A 'cognitive dimensions' framework. *Journal of Visual Languages and Computing*, 7(2):131–174, 1996.
15. R. Harper. *Practical foundations for programming languages*. Cambridge University Press, 2012.
16. L. C. L. Kats and E. Visser. The Spoofax Language Workbench. Rules for Declarative Specification of Languages and IDEs. In *Object-Oriented Programming Systems, Languages, and Applications*, 2010.
17. H. Krahn, B. Rumpe, and S. Völkel. Monticore: Modular development of textual domain specific languages. In *Objects, Components, Models and Patterns*, 2008.
18. L. Krishnan and E. V. Wyk. Termination analysis for higher-order attribute grammars. In *SLE*, pages 44–63, 2012.
19. B. Lee, R. Grimm, M. Hirzel, and K. S. McKinley. Marco: Safe, expressive macros for any language. In *ECOOP*, volume LNCS 7313, pages 356–382. Springer, 2012.
20. W. Lovas and F. Pfenning. A bidirectional refinement type system for lf. In *Electronic Notes in Theoretical Computer Science, 196:113–128, January 2008. [NPP07] [Pfe92] [Pfe93] [Pfe01] Aleksandar Nanevski, Frank Pfenning, and Brigitte*, 2008.

21. E. Möller. SRFI-49: Indentation-sensitive syntax. `http://srfi.schemers.org/srfi-49/srfi-49.html`, 2005.
22. L. Nistor, D. Kurilova, S. Balzer, B. Chung, A. Potanin, and J. Aldrich. Wyvern: A simple, typed, and pure object-oriented language. In *Proceedings of the 5th Workshop on MechAnisms for SPEcialization, Generalization and inHerItance*, MASPEGHI '13, pages 9–16, New York, NY, USA, 2013. ACM.
23. C. Omar, B. Chung, D. Kurilova, A. Potanin, and J. Aldrich. Type-directed, whitespace-delimited parsing for embedded dsls. In *Proceedings of the First Workshop on the Globalization of Domain Specific Languages*, GlobalDSL '13, pages 8–11, New York, NY, USA, 2013. ACM.
24. C. Omar, Y. Yoon, T. D. LaToza, and B. A. Myers. Active code completion. In *International Conference on Software Engineering*, 2012.
25. Z. Palmer and S. F. Smith. Backstage Java: Making a Difference in Metaprogramming. In *Object-Oriented Programming Systems, Languages, and Applications*, 2011.
26. B. C. Pierce. *Types and Programming Languages*. MIT Press, 2002.
27. A. C. Schwerdfeger and E. R. Van Wyk. Verifiable composition of deterministic grammars. In *Proceedings of the 2009 ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '09, pages 199–210, New York, NY, USA, 2009. ACM.
28. T. Sheard and S. Jones. Template meta-programming for haskell. *ACM SIGPLAN Notices*, 37(12):60–75, 2002.
29. M. Tatsubori, S. Chiba, M.-O. Killijian, and K. Itano. OpenJava: A Class-based Macro System for Java. In *Reflection and Software Engineering*, 2000.
30. E. Tempero, C. Anslow, J. Dietrich, T. Han, J. Li, M. Lumpe, H. Melton, and J. Noble. Qualitas corpus: A curated collection of java code for empirical studies. In *Proc. 2010 Asia Pacific Software Engineering Conference (APSEC'10)*, pages 336–345, Dec. 2010.
31. M. G. J. van den Brand. *Pregmatic: A Generator for Incremental Programming Environments*. PhD thesis, Katholieke Universiteit Nijmegen, 1992.

A Appendix

$$\frac{}{\vdash \emptyset}\ \textit{T-D-E} \qquad \frac{t \notin \textit{dom}(\Delta) \quad \Delta, t : \{\chi, -\} \vdash \chi \quad \Delta, t : \{\chi, -\} \vdash \delta}{\vdash \Delta, t : \{\chi, \delta\}}\ \textit{T-D-C}$$

$$\frac{t \notin \mathrm{dom}(\Delta) \quad \Delta, t : \{\omega, -\} \vdash \omega \quad \Delta, t : \{\omega, -\} \vdash \delta}{\vdash \Delta, t : \{\omega, \delta\}}\ \textit{T-D-O}$$

$$\frac{}{\Delta \vdash -}\ \textit{T-d-e} \qquad \frac{\Delta, \emptyset \vdash \hat{e} : \tau}{\Delta \vdash \hat{e} : \tau}\ \textit{T-d-m} \qquad \frac{}{\Delta \vdash \emptyset}\ \textit{T-DC-e} \qquad \frac{\Delta \vdash \Gamma \quad \Delta \vdash \tau}{\Delta \vdash \Gamma, x : \tau}\ \textit{T-DC-t}$$

$$\frac{t \in \mathrm{dom}(\Delta)}{\Delta \vdash t}\ \textit{T-T-v} \qquad \frac{\Delta \vdash \tau_1 \quad \Delta \vdash \tau_2}{\Delta \vdash \tau_1 \to \tau_2}\ \textit{T-T-a} \qquad \frac{\Delta \vdash \tau_1 \quad \Delta \vdash \tau_2}{\Delta \vdash \tau_1 \times \tau_2}\ \textit{T-T-p}$$

Fig. 17: Context and type well-formedness rules