

Reasonably Programmable Literal Forms

ANONYMOUS AUTHOR(S)

General-purpose programming languages typically provide literal notation for only a small number of standard data structures, like lists. This language design practice is decidedly *ad hoc*: there are many more data structures for which literal notation might also be useful (consider, for example, maps, regexes, HTML data, SQL queries, syntax trees and chemical structures.) Availability of literal notation is far from a frivolous concern: studies suggest that literal notation can have a non-trivial impact on software quality by removing the temptation to use string encodings, which are vectors for injection attacks and cause other difficulties. This paper details a mechanism for decentralizing control over literal notation: *typed literal macros (TLMs)*. TLMs give library providers programmatic control over the parsing and expansion of expressions and patterns of *generalized literal form* at a specified type or parametric family of types. Compared to existing approaches, including syntax extension systems and macro-based quotation systems, TLMs are uniquely *reasonable*: TLM providers can reason modularly about syntactic determinism, and TLM clients can reason abstractly about the type and binding structure of the program. The system only needs to convey to clients (via secondary notation) the inferred *segmentation* of each literal, which gives the locations of spliced terms. Maintaining information about segmentation supports the novel TLM hygiene mechanism, i.e. the mechanism that ensures that clients can reason abstractly about binding (previously, only unhygienic macros could freely splice terms out of string literal bodies.) This paper establishes these abstract reasoning principles formally with a calculus of expressions, patterns and ML-style modules. This calculus is the first detailed type-theoretic account of a typed, hygienic term-rewriting macro system of any design for a language with these features of ML or Scala.

1 Introduction

When designing the surface syntax of a general-purpose programming language, it is common practice to build in a few *literal forms* that decrease the syntactic cost of constructing and pattern matching over values of certain standard types. For example, in Standard ML (SML) and its cousins, a list expression literal like `[x, y, z]` is equivalent by desugaring to `Cons(x, Cons(y, Cons(z, Nil)))` where `Nil` and `Cons` identify the constructors of the standard list datatype (Harper 1997; Milner et al. 1997).¹ Another example is found in Ur/Web’s surface syntax, which extends Ur’s surface syntax with expression and pattern literal forms for encodings of HTML data (Chlipala 2010, 2015). Fig. 1 shows two such HTML literals, one that “splices in” a string expression (Line 1) and the other an HTML expression (Line 2).

This language design practice is decidedly *ad hoc* – it is easy to come up with other examples of data structures that are not privileged with literal forms in these languages, but for which mathematicians, scientists or programmers have elsewhere invented specialized notation (Cajori 1928; Iverson 1980). Consider, for example, that 1) clients of a “collections” library might like not just list literals, but also matrix, set and map literals following standard mathematical conventions; 2) clients of a “web programming” library might like not just HTML literals but also CSS literals (which Ur/Web does not provide); 4) a compiler writer might like “quotation” literals for the terms of the object language and various intermediate languages of interest; and 5) clients of a “chemistry” library might like composite chemical structure literals based on the SMILES standard (Anderson et al. 1987).

¹In SML, list literals work even in contexts where the list expression constructors have been shadowed by other bindings, i.e. the meaning of literal forms is *context independent*. We will return to the concept of context independence again.

```

1 1 fun heading first_name = <xml><h1>Hello, {[first_name]}!</h1></xml>
2 2 val body = <xml><body>{heading "World"} ...</body></xml>

```

Fig. 1. HTML literals with support for splicing at two different types are built primitively into Ur/Web.

Although concerns about “missing” literal notation are easy to dismiss as superficial or frivolous, the reality is that literal forms, or the absence thereof, have had a substantial influence on the quality of software that is written today. For example, Bravenboer et al. (2007) finds that literal forms for structured encodings of queries, like the SQL-based query literals now found in many languages (?), reduce the temptation to use string encodings of queries and therefore reduce the prevalence of string injection attacks, which are amongst the most common and serious security vulnerabilities on the web today (OWASP 2017). More generally, evidence suggests that programmers frequently engage in “stringly-typed programming”, i.e. they choose strings instead of composite data structures largely for reasons of syntactic convenience. In particular, Omar et al. (2014) sampled strings from open source projects and found that about 15% of them were parseable according to some readily apparent type-specific grammar, e.g. for SQL queries, regexes, file system paths, URLs and many other data structures. Literal forms at these types, with support for splicing, would decrease the syntactic cost of composite encodings, which are more amenable to programmatic manipulation and compositional reasoning than flat string encodings.

Of course, it would not be sensible to ask general-purpose language designers to accomodate all known notations with built-in literal forms. Instead, there has been considerable interest in mechanisms that allow library providers to define new literal notation on their own. For example, general syntax extension mechanisms, including Camlp4 (Leroy et al. 2014), SugarJ/Sugar* (Erdweg et al. 2011; Erdweg and Rieger 2013) and others that we will return to in Sec. 7, can be used to define new literal forms. The problem that specifically motivates this paper is that these mechanisms are quite *unreasonable* when “programming in the large” (DeRemer and Kron 1976). Consider, for instance, the perspective of a programmer attempting to comprehend (i.e. reason about) the program text in Fig. 2, which is written in a dialect of ML’s (or Scala’s) syntax, constructed by one of these mechanisms, where a large number of new literal forms have been defined.

```

1 val w = compute_w()
2 val x = compute_x(w)
3 val y = {|(!R)@&{/x!/:2_!x}'!R|}|

```

Fig. 2. An example of unreasonable program text.

The literal form on Line 3 above constructs an encoding of a query in the stack-based database query language K, using its terse surface syntax (Whitney and Shasha 2001). Of course, the programmer reading this program text may well be unfamiliar with K’s query syntax. Unfortunately for the programmer, there are no simple, abstract protocols for answering questions like:

- (1) **(Responsibility)** Which syntax extension determined the desugaring of the literal form on Line 3? Might there be many possible desugarings (i.e. syntactic non-determinism)?
- (2) **(Segmentation)** Are the characters x and R on Line 3 taken as spliced expressions (meaning that they appear directly in the desugaring), or are they parsed in some other way peculiar to this literal form (e.g. as operators in the K language)?
- (3) **(Capture)** If x is in fact a spliced expression, does it refer to the binding of x on Line 2, or might it capture an unseen binding of the same identifier used in the desugaring of Line 3?
- (4) **(Context Dependence)** If w , on Lines 1-2, is renamed, could that possibly break Line 3 or change its meaning because the desugaring assumes that some variable identified as w is in scope? What other hidden assumptions might the desugaring be making?
- (5) **(Typing)** What type does y have?

```

1  1  val w = compute_w()
2  2  val x = compute_x(w)
3  3  val y = $kquery `(!R)@&{&/x!//:2_!x}'!R}`

```

Fig. 3. TLMs make examples like the one from Fig. 2 more reasonable.

Forcing the programmer to reason transparently, i.e. to examine the desugaring of the program text, or even the logic that produced that desugaring, to answer basic questions like these defeats the ultimate purpose of syntactic sugar: decreasing cognitive cost (Green 1989). Analogous problems do not arise when programming without syntax extensions in languages like ML and Scala – the binding discipline of the language ensures that programmers can reason lexically about binding, and types mediate abstraction over function and module implementations.

This paper formally details a more reasonable mechanism for defining new literal forms: *typed literal macros* (TLMs). TLMs give library providers full programmatic control over the parsing and expansion of *generalized literal forms* (Omar et al. 2014), which syntactically subsume other literal forms because their bodies are left initially unparsed according to the fixed context-free syntax of the language (i.e. they are initially parsed like raw string literals.) For example, Fig. 3 applies a TLM, `$kquery`, to a backtick-delimited generalized literal form to express the “unreasonable” K query example from Fig. 2 just discussed. The applied TLM parses the literal body and generates an expansion statically, during a semantic phase called *typed expansion*, which generalizes the usual typing phase. TLMs come equipped with uniquely powerful abstract reasoning principles. We will characterize these reasoning principles precisely as we proceed, but for now, let us develop some intuitions: the programmer can reason abstractly about Line 3 as follows:

- (1) **(Responsibility)** The applied TLM, `$kquery`, is solely responsible for parsing and expansion of the literal body. The context-free syntax of the language is fixed, so the TLM provider can modularly establish that the syntax this TLM implements is deterministic.
- (2) **(Segmentation)** The output that a TLM generates is structured so that the system can infer from it an accurate *segmentation* of the literal body that distinguishes spliced terms from segments that are parsed in some other way by the TLM. The segmentation can be communicated by a code editor using secondary notation, e.g. colors in this document. By examining Fig. 3, we know that the two instances of `x` on Line 3 are parsed as spliced expressions (because they are in black), whereas the `R`’s must be parsed in some other way, e.g. as operators in the K query language (because they are in green.)
- (3) **(Capture)** Splicing is capture-avoiding, so the spliced expression `x` must refer to the binding of `x` on Line 2. It cannot capture a coincidental binding of `x` in the expansion.
- (4) **(Context Dependence)** The system enforces context independence, so the expansion of Line 3 cannot rely on the fact that, e.g., `w` is in scope. The client is free to rename `w`.
- (5) **(Typing)** An explicit type annotation on the definition of `$kquery` determines the type that the expansion must have. We will see an example of a TLM definition in the next section. Moreover, each spliced segment in the segmentation also comes paired with the type it is expected to have. This information is usually not necessary to reason about typing, but it can be conveyed to the programmer upon request by the program editor if desired.

The remainder of this document is organized as follows. Sec. 2 defines another simple expression TLM – one that implements Ur/Web-style HTML literal syntax (Fig. 1) – and uses it as a running example to informally detail how TLMs work from start to finish. Sec. 3 briefly introduces pattern TLMs, i.e. TLMs that generate patterns. Sec. 4 defines a type-theoretic calculus of simple expression and pattern TLMs and formally establishes the reasoning principles implied above. Sec. 5 then adds type functions and an ML-style module system and introduces *parametric TLMs* (pTLMs), i.e. TLMs that take type and module parameters. Parameters serve two key purposes. First, they allow a TLM

to operate not just at a single type, but over a type- and module-parameterized family of types. We give as an example a parametric TLM for string-keyed dictionaries that operates uniformly over different implementations of the abstract dictionary signature, and also over different choices of the range type. Second, parameters give expansions access to terms other than those explicitly spliced in to the literal body by the client. Partial parameter application is possible. Together, this lowers the syntactic cost of the strict context independence discipline that TLMs enforce. Sec. 6 considers the topic of term evaluation during the typed expansion phase, i.e. *static evaluation*, in more detail, and also gives examples of TLMs that are useful for defining other TLMs, e.g. TLMs that implement parser generators and quasiquotation. Sec. 7 compares TLMs to related work, including recent related work by Omar et al. (2014, 2015), and the long history of related work on macro systems. Sec. 8 concludes with a summary of the specific contributions of this paper and a brief discussion of the limitations of TLMs, along with some suggestions for future work.

Due to space limitations, certain technical details of the calculi defined in this paper, as well as proofs, are relegated to the supplemental material. The examples in this paper are written in an emerging dialect of ML called VerseML that is focused on improving the ergonomics of ML. However, the essential ideas are not ML-specific – it should be possible to adapt TLMs to other languages that take a similar approach to types and binding (e.g. Haskell, Scala and others.) When a design choice is ML-specific, we will make note of it. Languages that have a disciplined binding structure but that lack a rich static type structure, e.g. Racket and other LISP derivatives, might also benefit from TLMs – they can apply the results of this paper by deploying the usual trick of reducing the language to one that is “untyped” (Harper 2016; Scott 1980).

2 Simple Expression TLMs (seTLMs)

This section informally details the TLM mechanism. It will suffice in this section to consider only *simple expression TLMs* (seTLMs), i.e. TLMs for abbreviating expressions of a single specified type. In particular, this section is organized around an seTLM named \$html that implements Ur/Web-style HTML literals, which abbreviate expressions of the type html defined as follows:

```
type html = BodyElement of list(html_attr) * list(html)
          | H1Element of list(html_attr) * list(html)
          | (* ... *) | TextNode of string
```

The type html_attr is straightforward and the type function list is standard, so neither are shown. In VerseML, strings are primitive (though this is not essential, see footnote 1 below.)

Using this TLM, we can express the Ur/Web program from Fig. 1 as shown in Fig. 4: on each line, we apply \$html to a *generalized literal form* delimited by backticks. TLM names are prefixed by \$ to clearly them from variable identifiers. The semantics delegates control over the parsing and expansion of each literal body to the applied TLM during a semantic phase called *typed expansion*, which generalizes the usual typing phase.

```
1 fun heading first_name = $html `<h1>Hello, {[first_name]}!</h1>`
2 val body = $html `<body>{heading "World!"} ...</body>`
```

Fig. 4. Two examples of a TLM being applied to a generalized literal form. We assume strings are primitive.

Generalized literal forms, which first arose in related work by Omar et al. (2014) that we will detail in Sec. 8, subsume a variety of common syntactic forms because the context-free syntax of the language only defines which outer delimiters are available (in this paper, we will mostly use backticks, but in general, there can be many choices.) *Literal bodies*, i.e. the characters between the delimiters, are otherwise syntactically unconstrained. Choices regarding internal delimitation of spliced terms is not imposed by the language definition. Notice, however, that Figure 4 reveals the locations of the spliced expressions by coloring them black. We have designed our system so that a

figure like this, which presents a *segmentation* of each literal body into spliced terms (in black) and characters parsed in some other way by the applied TLM (in green)², can always be generated.

2.1 TLM Definitions

The definition of the seTLM named \$html that was shown being applied in Fig. 4 takes the following form:

```

1  syntax $html at html by
2    static fn(b : body) -> parse_result(proto_expr) =>
3      (* Ur/Web-style HTML syntax parser here *)
4  in ... end

```

Every seTLM definition consists of a TLM name, here \$html, a *type annotation*, here **at** html, and a *parse function* between **by** and **in**. The TLM definition is in scope between **in** and **end**. In practice, if an **in** clause is not provided, the definition is in scope until the end of the enclosing declaration (e.g. the enclosing function or module.) We will consider how TLM definitions are exported from libraries in Sec 6.

The parse function is a *static function* delegated responsibility over parsing the literal bodies of the literal forms to which the TLM is applied. Static functions, marked by the **static** keyword, are applied during the typed expansion process, so they cannot refer to the surrounding variable bindings (because those variables stand for dynamic values.) For now, we will simply assume that static functions are closed, and that they do not themselves make use of TLMs (we will eliminate these impractical limitations in Sec. 6.)

The parse function must have type `body -> parse_result(proto_expr)`. The input type, `body`, classifies encodings of literal bodies. In VerseML, literal bodies are sequences of characters, so it suffices to define `body` as an abbreviation for the string type, as shown in Figure 5. The return type is a labeled sum type, defined by applying the type function `parse_result` defined in Figure 5, that distinguishes between parse errors and successful parses. Let us consider these two possibilities in turn.

Parse Errors. If the parse function determines that the literal body is not well-formed (according to whatever syntax definition that it implements), it returns:

```
ParseError {msg=emsg, loc=eloc}
```

where e_{msg} is an error message and e_{loc} is a value of type `segment`, defined in Figure 5, that designates a segment of the literal body as the location of the error (van Deursen et al. 1993). This information is for use by VerseML compilers when reporting the error to the programmer (but it otherwise has no semantic significance.) We might alternatively have used an exception to signal a parse error, but sum types are easier to formalize.

Successful Parses. If parsing succeeds, the parse function returns

```
Success eproto
```

where e_{proto} is called the *encoding of the proto-expansion*. For expression TLMs, proto-expansions are *proto-expressions*, which are encoded as values of the recursive sum type `proto_expr` outlined in Figure 5. Most of the variants of `proto_expr` are individually uninteresting – they encode VerseML’s various expression forms (just as in a self-hosted compiler, e.g. SML/NJ’s Visible Compiler library.) Expressions can mention types, so we also need the type `proto_typ` also outlined in Figure 5. More

²In general, spliced expressions might themselves apply TLMs, in which case the convention is to use a distinct color for unspliced segments at each depth. For example, if strings were not primitive, we might write `<body>{heading ($str "World!")} ...</body>`.

```

1  type body = string
2  type segment = {startIdx : nat, endIdx : nat} (* inclusive *)
3  type parse_result('a) = ParseError of { msg : string, loc : segment }
4                        | Success of 'a
5  type var_t = string (* ... or some other encoding of variables ... *)
6  type proto_typ = TyVar of var_t
7                | Arrow of proto_typ * proto_typ
8                | (* ... *)
9                | SplicedT of segment
10 type proto_expr = Var of var_t
11                | Fn of var_t * proto_typ * proto_expr
12                | Ap of proto_expr * proto_expr
13                | (* ... *)
14                | SplicedE of segment * proto_typ

```

Fig. 5. Definitions of various types used by TLM definitions. We assume that these are available ambiently.

sophisticated encoding techniques for terms would cause no fundamental problems. It is only the SplicedE and SplicedT variants of these types that are interesting.

2.2 Splicing

When a parse function determines that a segment of the literal body should be taken as a spliced expression, it does not directly insert the syntax tree of that expression into the encoding of the expansion. Instead, the TLM must refer to the spliced expression *by its relative location* using the SplicedE variant of proto_expr, which takes a value of type segment that indicates the zero-indexed location of the spliced expression relative to the start of the provided literal body. The SplicedE variant also requires a value of type proto_typ, which indicates the type that the spliced expression is expected to have. Types can be spliced out by using the SplicedT variant of proto_typ analogously.

For example, consider again the TLM application on Line 1 of Fig. 4, reproduced below:

```
$html `<h1>Hello, {[first_name]}</h1>`
```

Let us assume that html abbreviates a recursive labeled sum type having at least the cases

```
H1Element of list(html_attr) * list(html) | TextNode of string | ...
```

The proto-expansion generated by \$html for the supplied literal body, if written in a hypothetical textual syntax where references to spliced expressions are written **spliced**<startIdx; endIndex; ty>, is:

```
H1Element(Nil, Cons(TextNode "Hello, ", Cons(
    TextNode spliced<13; 22; string>, Nil)))
```

Here, **spliced**<13; 22; string> refers to the spliced expression first_name by location, rather than directly, and indicates that it must be of type string.

Requiring that TLMs refer to spliced expressions indirectly in this manner prevents them from “forging” spliced terms (i.e. claiming that an term is a spliced term when it does not in fact appear textually in the literal body.)

The *splice summary* of a proto-expression is the finite set of references to spliced terms contained within. For example, the summary of the proto-expression above is the finite set containing only **spliced**<13; 22; string>.

The *segmentation* of a proto-expression is the finite set of segments extracted from the splice summary, where a segment is simply a pair of natural numbers. For example, the segmentation

of the proto-expansion above is $\{(13, 22)\}$. The semantics checks that all of the segments in the segmentation are 1) in bounds relative to the literal body; and 2) non-overlapping. This resolves the problem of **Segmentation** described in Sec. 1, i.e. every literal body in a well-typed program has a well-defined segmentation.

A program editor or pretty-printer can communicate the segmentation information to the programmer, e.g. by coloring non-spliced segments green as is our convention in this document. A program editor or pretty-printer can also communicate the type of each spliced term, as indicated in the splice summary, to the programmer upon request (for example, the Emacs and Vim packages for working with OCaml defer to the Merlin tool when the programmer requests the type of an expression.)

2.3 Proto-Expansion Validation

Three concerns described in Sec. 1 remain: those related to reasoning abstractly about **Capture**, **Context Dependence** and **Typing**. Addressing these concerns is the purpose of the *proto-expansion validation* process. The result of proto-expansion validation is the *final expansion*, which is simply the proto-expansion with references to spliced terms replaced with their corresponding final expansions.

2.3.1 Capture. Proto-expansion validation ensures that spliced terms have access *only* to the bindings that appear at the application site – spliced terms cannot capture the bindings that appear in the proto-expansion. For example, consider the following application site:

```
let tmp = (* ... application site temporary ... *) in
$html /<h1>{f(tmp)}</h1>/
```

Now consider the scenario where the proto-expansion generated by `$html` has the following form:

```
let tmp = (* ... expansion-internal temporary ... *) in
H1Element(tmp, spliced<5; 10; html>)
```

Naïvely, the binding of the variable `tmp` in the proto-expansion could shadow the application-site binding of `tmp` in the final expansion.

To address this problem, proto-expansion validation is capture-avoiding. In other words, it discharges the requirement that capture not occur by implicitly alpha-varying the bindings in the proto-expansion as needed. For example, the final expansion of the example above might take the following form:

```
let tmp = (* ... application site temporary ... *) in
let tmp' = (* ... expansion-internal temporary ... *) in
H1Element(tmp', f(tmp))
```

Notice that the expansion-internal binding of `tmp` has been alpha-varied to `tmp'` to avoid shadowing the application site binding of `tmp`. As such, the reference to `tmp` in the spliced expression refers, as intended, to the application site binding of `tmp`.

For TLM providers, the benefit of this mechanism is that they can name the variables used internally within expansions freely, without worrying about whether their chosen identifiers might shadow those that a client might have used at the application site. There is no need for a mechanism that generates “fresh variables”.

TLM clients can, in turn, reliably reason about binding within every spliced expression without examining the expansion that the spliced expression appears within.

The trade-off is that this prevents library providers from intentionally introducing bindings into spliced terms. For example, Haskell’s derived form for monadic commands (i.e. **do**-notation) supports binding the result of executing a command to a variable that is then available in the

subsequent commands in a command sequence. In VerseML, this cannot be expressed in the same way. Values can be communicated from the expansion to a spliced expression only via function arguments. We will return to this issue in Sec. 8.

2.3.2 Context Dependence. The proto-expansion validation process also ensures that variables that appear in the proto-expansion do not refer to bindings that appear either at the TLM definition or the application site. In other words, expansions must be completely *context independent* – they can make no assumptions about the surrounding context.

A minimal example of a “broken” TLM that does not generate context-independent proto-expansions is below:

```
syntax $broken at rx by
  static fn(_) => Success (Var "x")
end
```

The proto-expansion that this TLM generates (for any given literal body) refers to a variable x that it does not itself bind. If proto-expansion validation permitted such a proto-expansion, it would be well-typed only under those application site typing contexts where x is bound. This “hidden assumption” makes reasoning about binding and renaming difficult, so this proto-expansion is deemed invalid (even when $\$broken$ is applied where x is coincidentally bound.)

Of course, this prohibition does not extend into the spliced terms in a proto-expansion – spliced terms appear at the application site, so they can justifiably refer to application site bindings. Indeed, we saw examples of spliced terms that referred to variables bound at the application site above. Because proto-expansions refer to spliced terms indirectly, enforcing context independence is straightforward – we need only that the proto-expansion itself be closed, without considering the spliced terms.

One subtlety is that we assumed in the examples above that constructors like `H1Element` were available to the proto-expansion. In most dialects of ML, where datatype constructors are injected into the context as variables when the datatype is defined, this would violate context independence. As such, we will have to assume for now (and in our formalism in Sec. 4) that explicit injections into labeled sum types are available. Once we introduce module parameters in Sec. 5, we will be able to pass in a module exporting the datatype constructors as a parameter and partially apply that parameter to hide this detail from clients.

2.3.3 Typing. Finally, proto-expansion validation maintains a reasonable *typing discipline* by:

- (1) checking to ensure that the generated expansion is of the type specified in the TLM’s type annotation;
- (2) checking that the each spliced type is valid at the application site;
- (3) checking that the type annotation on each spliced expression is valid at the application site; and
- (4) checking each spliced expression against the provided type annotation.

3 Simple Pattern TLMs (spTLMs)

Let us now briefly consider the topic of TLMs that expand to *patterns*, rather than expressions. For example, we can pattern match on a value x : `html` by applying a pattern TLM `$html` as follows:

```
match x with
| $html `<h1>{y}</h1>` -> y (* y : list(html) *)
| _ -> (* ... *)
end
```

Pattern TLM definitions look much like expression TLM definitions:


```

1  type proto_pat = (* no variable pattern form! *)
2                    Wild
3                    | (* ... other standard pattern forms ... *)
4                    | SplicedP of segment * proto_typ

```

Fig. 6. Abbreviated definition of proto_pat.

```

8  1  syntax $html at html for patterns by
9  2      static fn(b : body) : parse_result(proto_pat) =>
10 3      (* HTML pattern parser here *)
11 4  in (* ... *) end

```

The *sort qualifier*, **for patterns**, indicates that this is a pattern TLM definition. Expression TLM definitions can also include a sort qualifier, **for expressions**, but this is the default if omitted. Note that we used the same name, `$html`, for this pattern TLM as for the expression TLM defined in the previous section. This is possible because patterns and expressions are distinct sorts of terms. It does not make sense to apply an expression TLM in pattern position, and *vice versa*.³ The return type of the parse function is `parse_result(proto_pat)`, rather than `parse_result(proto_expr)`. The type `proto_pat`, defined in part in Figure 6, classifies encodings of *proto-patterns*.

3.1 Splicing

The constructor `SplicedP` serves much like `SplicedE` to allow a proto-pattern to refer indirectly to spliced pattern segments by their location within the literal body. For example, the proto-pattern generated for the example at the top of this section would be written concretely as follows:

```
H1Element (_, spliced<5; 5; list(html)>)
```

3.2 Proto-Pattern Validation

Proto-pattern validation serves, like proto-expression validation, to maintain the ability to reason abstractly about binding and typing.

3.2.1 Binding. To maintain a reasonable abstract binding discipline, i.e. to allow clients to reason about variable binding without examining pattern TLM expansions directly, variable patterns can appear only within spliced patterns. Enforcing this restriction is straightforward: we simply have not defined a variant of the `proto_pat` type that encodes variable patterns (though wildcards are allowed.) This prohibition ensures that no variables other than those visible to the client in a spliced pattern are captured by the corresponding branch of the match expression.

In our simple language, patterns have no way to refer to surrounding bindings (they only induce bindings). However, the type annotations on references to spliced patterns could refer to type variables, so we also need to enforce context independence in the manner discussed in the previous discussion.

Patterns cannot bind variables for use in sub-patterns, so we do not face the problem of capture.

3.2.2 Typing. To maintain a reasonable abstract typing discipline, proto-pattern validation checks:

- (1) that the final expansion is a pattern that matches values of the type specified in the TLM type annotation;
- (2) that each spliced pattern matches values of the type indicated in the splice summary; and

³The fact that some patterns look like expressions in the textual syntax is immaterial to this fundamental semantic distinction. Additionally, many expression-level constructs, e.g. lambdas, do not correspond even syntactically to patterns.

```

1  UType  $\hat{\tau} ::= \hat{t} \mid \hat{\tau} \rightarrow \hat{\tau} \mid \forall \hat{t}. \hat{\tau} \mid \mu \hat{t}. \hat{\tau} \mid \langle \{i \hookrightarrow \hat{\tau}_i\}_{i \in L} \rangle \mid [\{i \hookrightarrow \hat{\tau}_i\}_{i \in L}]$ 
2  UExp  $\hat{e} ::= \hat{x} \mid \lambda \hat{x}. \hat{t}. \hat{e} \mid \hat{e}(\hat{e}) \mid \Lambda \hat{t}. \hat{e} \mid \hat{e}[\hat{t}] \mid \text{fold}(\hat{e}) \mid \langle \{i \hookrightarrow \hat{e}_i\}_{i \in L} \rangle \mid \text{inj}[\ell](\hat{e}) \mid \text{match } \hat{e} \{ \hat{r}_i \}_{1 \leq i \leq n}$ 
3       $\mid \text{syntax } \hat{a} \text{ at } \hat{\tau} \text{ for expressions by static } e \text{ in } \hat{e} \mid \hat{a} 'b'$ 
4       $\mid \text{syntax } \hat{a} \text{ at } \hat{\tau} \text{ for patterns by static } e \text{ in } \hat{e}$ 
5  URule  $\hat{r} ::= \hat{p} \Rightarrow \hat{e}$ 
6  UPat  $\hat{p} ::= \hat{x} \mid \_ \mid \text{fold}(\hat{p}) \mid \langle \{i \hookrightarrow \hat{p}_i\}_{i \in L} \rangle \mid \text{inj}[\ell](\hat{p}) \mid \hat{a} 'b'$ 

```

Fig. 7. Syntax of the miniVerse_S unexpanded language (UL). Metavariable \hat{t} ranges over type identifiers, \hat{x} over expression identifiers, ℓ over labels, L over finite sets of labels, \hat{a} over TLM names and b over literal bodies. We write $\{i \hookrightarrow \hat{\tau}_i\}_{i \in L}$ for a finite mapping of each label i in L to some unexpanded type $\hat{\tau}_i$, and similarly for other sorts. We write $\{\hat{r}_i\}_{1 \leq i \leq n}$ for a finite sequence of n unexpanded rules.

```

11 Typ  $\tau ::= t \mid \text{parr}(\tau; \tau) \mid \text{all}(t. \tau) \mid \text{rec}(t. \tau) \mid \text{prod}[L](\{i \hookrightarrow \tau_i\}_{i \in L}) \mid \text{sum}[L](\{i \hookrightarrow \tau_i\}_{i \in L})$ 
12 Exp  $e ::= x \mid \text{lam}\{\tau\}(x. e) \mid \text{ap}(e; e) \mid \text{tlam}(t. e) \mid \text{tap}\{\tau\}(e) \mid \text{fold}(e) \mid \text{tpl}[L](\{i \hookrightarrow e_i\}_{i \in L}) \mid \text{inj}[\ell](e)$ 
13       $\mid \text{match}[n](e; \{r_i\}_{1 \leq i \leq n})$ 
14 Rule  $r ::= \text{rule}(p. e)$ 
15 Pat  $p ::= x \mid \text{wildp} \mid \text{foldp}(p) \mid \text{tplp}[L](\{i \hookrightarrow p_i\}_{i \in L}) \mid \text{injp}[\ell](p)$ 

```

Fig. 8. Syntax of the miniVerse_S (XL). XL terms are *abstract binding trees* (ABTs) identified up to alpha-equivalence, so we follow the syntactic conventions of Harper (2012). Metavariable x ranges over variables and t over type variables.

(3) that each of these types are themselves well-formed types at the application site.

4 Simple TLMs, Formally

Before continuing on to consider parametric TLMs, let us develop a calculus of simple expression and pattern TLMs called miniVerse_S. It is organized as an *unexpanded language*, or *UL*, defined by typed expansion to an *expanded language*, or *XL*. Figures 7 and 8 summarize the syntax of the UL and the XL, respectively. Programs are written as unexpanded expressions but evaluate as well-typed expanded expressions. We will start with a brief overview of our XL before turning in the remainder of the section on the UL and the typed expansion process.

4.1 Expanded Language (XL)

The XL of miniVerse_S forms a standard pure functional language with partial function types, quantification over types, recursive types, labeled product types and labeled sum types. The reader is directed to *PFPL* (Harper 2012) for a detailed introductory account of these standard constructs. We will only tersely summarize the statics and dynamics of the XL below. The particularities of the XL are not critical to the ideas we will introduce below.

4.1.1 Statics of the Expanded Language. The *statics of the XL* is defined by hypothetical judgements of the following form:

$\Delta \vdash \tau \text{ type}$	τ is a well-formed type
$\Delta \Gamma \vdash e : \tau$	e is assigned type τ
$\Delta \Gamma \vdash r : \tau \Rightarrow \tau'$	r takes values of type τ to values of type τ'
$\Delta \vdash p : \tau \dashv \Gamma$	p matches values of type τ and generates hypotheses Γ

Type formation contexts, Δ , are finite sets of hypotheses of the form $t \text{ type}$. *Typing contexts*, Γ , are finite functions that map each variable $x \in \text{dom}(\Gamma)$, where $\text{dom}(\Gamma)$ is a finite set of variables, to the hypothesis $x : \tau$, for some τ . The judgements above are inductively defined in the supplemental material and validate standard lemmas, also given in the supplement: Weakening, Substitution and Decomposition.

4.1.2 *Structural Dynamics*. The *structural dynamics*, a.k.a. the *structural operational semantics* (Plotkin 2004), of miniVerse_S is defined as a transition system by judgements of the following form:

$$\begin{array}{ll} e \mapsto e' & e \text{ transitions to } e' \\ e \text{ val} & e \text{ is a value} \end{array}$$

We also define auxiliary judgements for *iterated transition*, $e \mapsto^* e'$, and *evaluation* to a value, $e \Downarrow e'$. Additional auxiliary judgements, not shown, are needed to define the dynamics of pattern matching, but they do not appear directly in our subsequent developments, so we omit them. We assume an eager dynamics for simplicity and again assume that standard lemmas, stated in the supplement, hold: Canonical Forms, Progress and Preservation.

4.2 Syntax of the Unexpanded Language

Unexpanded types and expressions are simple inductive structures. They are **not** abstract binding trees – we do **not** define notions of renaming, alpha-equivalence or substitution for unexpanded terms. This is because unexpanded expressions remain “partially parsed” due to the presence of literal bodies, b , from which spliced terms might be extracted during typed expansion. In fact, unexpanded types and expressions do not involve variables at all, but rather *type identifiers*, \hat{t} , and *expression identifiers*, \hat{x} . Identifiers are given meaning by expansion to variables during typed expansion, as we will see below. This distinction between identifiers and variables is technically crucial.

There is also a corresponding context-free textual syntax for the UL. Giving a complete definition of the context-free textual syntax as, e.g., a context-free grammar, is not critical to our purposes here. Instead, we need only posit the existence of partial metafunctions $\text{parseUTyp}(b)$, $\text{parseUExp}(b)$ and $\text{parseUPat}(b)$ that go from character sequences, b , to unexpanded types, expressions and patterns, respectively.

CONDITION 4.1 (TEXTUAL REPRESENTABILITY).

- (1) For each \hat{t} , there exists b such that $\text{parseUTyp}(b) = \hat{t}$.
- (2) For each \hat{e} , there exists b such that $\text{parseUExp}(b) = \hat{e}$.
- (3) For each \hat{p} , there exists b such that $\text{parseUPat}(b) = \hat{p}$.

4.3 Typed Expansion

Unexpanded terms are checked and expanded simultaneously according to the *typed expansion judgements*:

$$\begin{array}{ll} \hat{\Delta} \vdash \hat{t} \rightsquigarrow \tau \text{ type} & \hat{t} \text{ has well-formed expansion } \tau \\ \hat{\Delta} \hat{\Gamma} \vdash_{\Psi; \hat{\Phi}} \hat{e} \rightsquigarrow e : \tau & \hat{e} \text{ has expansion } e \text{ of type } \tau \\ \hat{\Delta} \hat{\Gamma} \vdash_{\Psi; \hat{\Phi}} \hat{r} \rightsquigarrow r : \tau \Rightarrow \tau' & \hat{r} \text{ has expansion } r \text{ taking values of type } \tau \text{ to values of type } \tau' \\ \hat{\Delta} \vdash_{\hat{\Phi}} \hat{p} \rightsquigarrow p : \tau \dashv \hat{\Gamma} & \hat{p} \text{ has expansion } p \text{ matching against } \tau \text{ generating hypotheses } \hat{\Gamma} \end{array}$$

These judgements are inductively defined in the supplement. We will reproduce the interesting rules below.

Common Forms. Most of the unexpanded forms in Figure 7 mirror the expanded forms. We refer to these as the *common forms*. The mapping from expanded forms to common unexpanded forms is defined explicitly in the supplement. The typed expansion rules that handle these common forms mirror the corresponding typing rules. For example, the rules for variables and lambdas are

reproduced below:

$$\frac{\text{EE-ID}}{\hat{\Delta} \hat{\Gamma}, \hat{x} \rightsquigarrow x : \tau \vdash_{\hat{\Psi}, \hat{\Phi}} \hat{x} \rightsquigarrow x : \tau} \quad \frac{\text{EE-LAM} \quad \hat{\Delta} \vdash \hat{\tau} \rightsquigarrow \tau \text{ type} \quad \hat{\Delta} \hat{\Gamma}, \hat{x} \rightsquigarrow x : \tau \vdash_{\hat{\Psi}, \hat{\Phi}} \hat{e} \rightsquigarrow e : \tau'}{\hat{\Delta} \hat{\Gamma} \vdash_{\hat{\Psi}, \hat{\Phi}} \lambda \hat{x} : \hat{\tau}. \hat{e} \rightsquigarrow \text{lam}\{\tau\}(x.e) : \text{parr}(\tau; \tau')}$$

The only subtlety has to do with the relationship between identifiers in the UL and variables in the XL. To understand this, we must first describe in detail how unexpanded contexts work.

Unexpanded typing contexts, $\hat{\Gamma}$, are pairs of the form $\langle \mathcal{G}; \Gamma \rangle$, where \mathcal{G} is an *expression identifier expansion context*, and Γ is a standard typing context. An expression identifier expansion context, \mathcal{G} , is a finite function that maps each expression identifier $\hat{x} \in \text{dom}(\mathcal{G})$ to the hypothesis $\hat{x} \rightsquigarrow x$, for some expression variable, x , called its expansion. We write $\mathcal{G} \uplus \hat{x} \rightsquigarrow x$ for the expression identifier expansion context that maps \hat{x} to $\hat{x} \rightsquigarrow x$ and defers to \mathcal{G} for all other expression identifiers (i.e. the previous mapping is **updated**.) Note the distinction between update and extension (which requires that the new identifier is not already in the domain.) We define $\hat{\Gamma}, \hat{x} \rightsquigarrow x : \tau$ when $\hat{\Gamma} = \langle \mathcal{G}; \Gamma \rangle$ as an abbreviation of $\langle \mathcal{G} \uplus \hat{x} \rightsquigarrow x; \Gamma, x : \tau \rangle$.

To develop an intuition for why the update operation is necessary, it is instructive to inspect the derivation of the expansion of the unexpanded expression $\lambda \hat{x} : \hat{\tau}. \lambda \hat{x} : \hat{\tau}. \hat{x} \rightsquigarrow \text{lam}\{\tau\}(x_1. \text{lam}\{\tau\}(x_2. x_2))$ assuming $\hat{\Delta} \vdash \hat{\tau} \rightsquigarrow \tau \text{ type}$:

$$\frac{\frac{\frac{\hat{\Delta} \vdash \hat{\tau} \rightsquigarrow \tau \text{ type}}{\hat{\Delta} \langle \hat{x} \rightsquigarrow x_2; x_1 : \tau, x_2 : \tau \rangle \vdash_{\hat{\Psi}, \hat{\Phi}} \hat{x} \rightsquigarrow x_2 : \tau} \text{EE-ID}}{\hat{\Delta} \langle \hat{x} \rightsquigarrow x_1; x_1 : \tau \rangle \vdash_{\hat{\Psi}, \hat{\Phi}} \lambda \hat{x} : \hat{\tau}. \hat{x} \rightsquigarrow \text{lam}\{\tau\}(x_2. x_2) : \text{parr}(\tau; \tau)} \text{EE-LAM}}{\hat{\Delta} \langle \emptyset; \emptyset \rangle \vdash_{\hat{\Psi}, \hat{\Phi}} \lambda \hat{x} : \hat{\tau}. \lambda \hat{x} : \hat{\tau}. \hat{x} \rightsquigarrow \text{lam}\{\tau\}(x_1. \text{lam}\{\tau\}(x_2. x_2)) : \text{parr}(\tau; \text{parr}(\tau; \tau))} \text{EE-LAM}$$

Notice that when Rule EE-LAM is applied, the type identifier expansion context is updated but the typing context is extended with a (necessarily fresh) variable, first x_1 then x_2 . Without this mechanism, expansions for unexpanded terms with shadowing, like this minimal example, would not exist, because we cannot implicitly alpha-vary the unexpanded term to sidestep this problem in the usual manner.

Unexpanded type formation contexts, $\hat{\Delta}$, consist of a *type identifier expansion context*, \mathcal{D} , paired with a standard type formation context, Δ , and operate analogously (see supplement.)

Well-Typed Expansion. Before we continue, let us state an important invariant: that typed expansion produces a well-typed expression.

THEOREM 4.2 (TYPED EXPRESSION EXPANSION). *If $\langle \mathcal{D}; \Delta \rangle \langle \mathcal{G}; \Gamma \rangle \vdash_{\hat{\Psi}, \hat{\Phi}} \hat{e} \rightsquigarrow e : \tau$ then $\Delta \Gamma \vdash e : \tau$.*

For the typed expansion rules governing common forms, like the two example rules above, the typed expansion rules mirror the corresponding typing rules so it is easy to see that this invariant holds. The details are in the supplement. The rules of particular interest are the rules governing TLM definitions and TLM application, which are covered in the next two subsections.

4.4 TLM Definitions

An unexpanded expression of seTLM definition form, syntax \hat{a} at $\hat{\tau}$ for expressions by static e_{parse} in \hat{e} , defines an seTLM identified as \hat{a} with *unexpanded type annotation* $\hat{\tau}$ and *parse function* e_{parse} for

use within \hat{e} . The rule below defines typed expansion of this form:

$$\text{EE-DEF-SETSM} \quad \frac{\begin{array}{c} \hat{\Delta} \vdash \hat{\tau} \rightsquigarrow \tau \text{ type} \quad \emptyset \emptyset \vdash e_{\text{parse}} : \text{parr}(\text{Body}; \text{ParseResultSE}) \\ e_{\text{parse}} \Downarrow e'_{\text{parse}} \quad \hat{\Delta} \hat{\Gamma} \vdash_{\hat{\Psi}, \hat{a} \rightsquigarrow a \hookrightarrow \text{seTLM}(\tau; e'_{\text{parse}}); \hat{\Phi}} \hat{e} \rightsquigarrow e : \tau' \end{array}}{\hat{\Delta} \hat{\Gamma} \vdash_{\hat{\Psi}, \hat{\Phi}} \text{syntax } \hat{a} \text{ at } \hat{\tau} \text{ for expressions by static } e_{\text{parse}} \text{ in } \hat{e} \rightsquigarrow e : \tau'}$$

The premises of this rule can be understood as follows, in order:

- (1) The first premise expands the unexpanded type annotation.
- (2) The second premise checks that e_{parse} is a closed expanded function⁴ of type $\text{parr}(\text{Body}; \text{ParseResultSE})$.

The type abbreviated *Body* classifies encodings of literal bodies, b . The mapping from literal bodies to values of type *Body* is defined by the *body encoding judgement* $b \downarrow_{\text{Body}} e_{\text{body}}$. An inverse mapping is defined by the *body decoding judgement* $e_{\text{body}} \uparrow_{\text{Body}} b$. Rather than defining *Body* explicitly, and these judgements inductively against that definition (which would be tedious and uninteresting), it suffices to take as a condition that there is an isomorphism between literal bodies and values of type *Body* mediated by these judgements (see supplement.)

The return type of the parse function, *ParseResultSE*, abbreviates a labeled sum type that distinguishes parse errors from successful parses:

$$L_{\text{SE}} \stackrel{\text{def}}{=} \text{ParseError}, \text{SuccessE}$$

$$\text{ParseResultSE} \stackrel{\text{def}}{=} \text{sum}[L_{\text{SE}}](\text{ParseError} \hookrightarrow \langle \rangle, \text{SuccessE} \hookrightarrow \text{PrExpr})$$

The type abbreviated *PrExpr* classifies encodings of *proto-expressions*, \hat{e} (pronounced “grave e ”). The syntax of proto-expressions, defined in Figure 9, will be described when we describe proto-expansion validation in Sec. 4.6. The mapping from proto-expressions to values of type *PrExpr* is defined by the *proto-expression encoding judgement*, $\hat{e} \downarrow_{\text{PrExpr}} e$. An inverse mapping is defined by the *proto-expression decoding judgement*, $e \uparrow_{\text{PrExpr}} \hat{e}$. Again, rather than picking a particular definition of *PrExpr* and defining the judgements above inductively against it, we take as a condition that there is an isomorphism between values of type *PrExpr* and proto-expressions mediated by these judgements (see supplement.)

- (3) The third premise of Rule *EE-DEF-SETSM* evaluates the parse function to a value. This is not semantically necessary, but it is the choice one would expect to make assuming an eagerly evaluated language.
- (4) The final premise of Rule *EE-DEF-SETSM* extends the *seTLM* context, $\hat{\Psi}$, with the newly determined *seTLM* definition, and proceeds to assign a type, τ' , and expansion, e , to \hat{e} . The conclusion of the rule then assigns this type and expansion to the *seTLM* definition as a whole.

seTLM contexts, $\hat{\Psi}$, are of the form $\langle \mathcal{A}; \Psi \rangle$, where \mathcal{A} is a *TLM identifier expansion context* and Ψ is an *seTLM definition context*.

A *TLM identifier expansion context*, \mathcal{A} , is a finite function mapping each *TLM identifier* $\hat{a} \in \text{dom}(\mathcal{A})$ to the *TLM identifier expansion*, $\hat{a} \rightsquigarrow a$, for some *TLM name*, a . We distinguish *TLM identifiers*, \hat{a} , from *TLM names*, a , for much the same reason that we distinguish type and expression identifiers from type and expression variables: in order to allow a *TLM* definition to shadow a previously defined *TLM* definition without relying on an implicit identification convention.

⁴In Sec. 6, we add the machinery necessary for parse functions that are neither closed nor yet expanded.

An seTLM definition context, Ψ , is a finite function mapping each TLM name $a \in \text{dom}(\Psi)$ to an *expanded seTLM definition*, $a \mapsto \text{setlm}(\tau; e_{\text{parse}})$, where τ is the seTLM's type annotation, and e_{parse} is its parse function. We define $\hat{\Psi}, \hat{a} \rightsquigarrow a \mapsto \text{setlm}(\tau; e_{\text{parse}})$, when $\hat{\Psi} = \langle \mathcal{A}; \Psi \rangle$, as an abbreviation of $\langle \mathcal{A} \uplus \hat{a} \rightsquigarrow a; \Psi, a \mapsto \text{setlm}(\tau; e_{\text{parse}}) \rangle$.

The spTLM definition form, syntax \hat{a} at $\hat{\tau}$ for patterns by static e_{parse} in \hat{e} , operates analogously. The rule is reproduced below, and the details, which mirror those for seTLMs, are in the supplement. Notice that the spTLM context, $\hat{\Phi}$, rather than the $\hat{\Psi}$ is updated. This allows expression and pattern TLMs to share an identifier.

$$\begin{array}{c} \text{EE-DEF-SPTSM} \\ \frac{\begin{array}{c} \hat{\Delta} \vdash \hat{\tau} \rightsquigarrow \tau \text{ type} \quad \emptyset \emptyset \vdash e_{\text{parse}} : \text{parr}(\text{Body}; \text{ParseResultSP}) \\ e_{\text{parse}} \Downarrow e'_{\text{parse}} \quad \hat{\Delta} \hat{\Gamma} \vdash_{\hat{\Psi}; \hat{\Phi}} \hat{a} \rightsquigarrow a \mapsto \text{sptlm}(\tau; e'_{\text{parse}}) \quad \hat{e} \rightsquigarrow e : \tau' \end{array}}{\hat{\Delta} \hat{\Gamma} \vdash_{\hat{\Psi}; \hat{\Phi}} \text{syntax } \hat{a} \text{ at } \hat{\tau} \text{ for patterns by static } e_{\text{parse}} \text{ in } \hat{e} \rightsquigarrow e : \tau'} \end{array}$$

4.5 TLM Application

The unexpanded expression form for applying an seTLM named \hat{a} to a literal form with literal body b is $\hat{a} \text{ ' } b \text{ '}$. The typed expansion rule governing seTLM application is below:

$$\begin{array}{c} \text{EE-AP-SETSM} \\ \frac{\begin{array}{c} \hat{\Psi} = \hat{\Psi}', \hat{a} \rightsquigarrow a \mapsto \text{setlm}(\tau; e_{\text{parse}}) \\ b \downarrow_{\text{Body}} e_{\text{body}} \quad e_{\text{parse}}(e_{\text{body}}) \Downarrow \text{inj}[\text{SuccessE}](e_{\text{proto}}) \quad e_{\text{proto}} \uparrow_{\text{PrExpr}} \hat{e} \\ \text{seg}(\hat{e}) \text{ segments } b \quad \emptyset \emptyset \vdash_{\hat{\Delta}; \hat{\Gamma}; \hat{\Psi}; \hat{\Phi}} \hat{a} \text{ ' } b \text{ ' } \hat{e} \rightsquigarrow e : \tau \end{array}}{\hat{\Delta} \hat{\Gamma} \vdash_{\hat{\Psi}; \hat{\Phi}} \hat{a} \text{ ' } b \text{ ' } \rightsquigarrow e : \tau} \end{array}$$

The premises of this rule can be understood as follows, in order:

- (1) The first premise ensures that \hat{a} has been defined and extracts the type annotation and parse function.
- (2) The second premise determines the encoding of the literal body, e_{body} , which is a value of type *Body*.
- (3) The third premise applies the parse function e_{parse} to the encoding of the literal body. If parsing succeeds, i.e. a value of the form $\text{inj}[\text{SuccessE}](e_{\text{proto}})$ results from evaluation, then e_{proto} will be a value of type *PrExpr* (assuming a well-formed seTLM context, by application of the Preservation assumption.) We call e_{proto} the *encoding of the proto-expansion*. If the parse function produces a value labeled *ParseError*, then typed expansion fails. Formally, no rule is necessary to handle this case.
- (4) The fourth premise decodes the encoding of the proto-expansion to produce the proto-expansion itself.
- (5) The fifth premise determines the segmentation, $\text{seg}(\hat{e})$, and ensures that it is valid with respect to b via the predicate $\psi \text{ segments } b$, which checks that each segment in the finite set of segments ψ has non-negative length and is within bounds of b , and that the segments in ψ do not overlap.
- (6) The final premise *validates* the proto-expansion and simultaneously generates the *final expansion*, e , which appears in the conclusion of the rule. The proto-expression validation judgement is discussed next.

The typed pattern expansion rule governing spTLM application is analogous:

$$\begin{array}{c}
 \text{PE-AP-SPTSM} \\
 \frac{
 \begin{array}{c}
 \hat{\Phi} = \hat{\Phi}', \hat{a} \rightsquigarrow a \hookrightarrow \text{sptlm}(\tau; e_{\text{parse}}) \\
 b \downarrow_{\text{Body}} e_{\text{body}} \quad e_{\text{parse}}(e_{\text{body}}) \Downarrow \text{inj}[\text{SuccessP}](e_{\text{proto}}) \quad e_{\text{proto}} \uparrow_{\text{PrPat}} \hat{p} \\
 \text{seg}(\hat{p}) \text{ segments } b \quad \hat{p} \rightsquigarrow p : \tau \dashv \hat{\Delta}; \hat{\Phi}; b \hat{\Gamma}
 \end{array}
 }{
 \hat{\Delta} \vdash_{\hat{\Phi}} \hat{a} \text{ ' } b \text{ ' } \rightsquigarrow p : \tau \dashv \hat{\Gamma}
 }
 \end{array}$$

4.6 Proto-Expansion Validation

$\text{PrTyp } \hat{\tau} ::= t \mid \text{prparr}(\hat{\tau}; \hat{\tau}) \mid \text{prall}(t, \hat{\tau}) \mid \text{prrec}(t, \hat{\tau}) \mid \text{prprod}[L](\{i \hookrightarrow \hat{\tau}_i\}_{i \in L}) \mid \text{prsum}[L](\{i \hookrightarrow \hat{\tau}_i\}_{i \in L})$
 $\mid \text{splicedt}[m; n]$
 $\text{PrExp } \hat{e} ::= x \mid \text{prlam}\{\hat{\tau}\}(x, \hat{e}) \mid \text{prap}(\hat{e}; \hat{e}) \mid \text{prtlam}(t, \hat{e}) \mid \text{prtap}\{\hat{\tau}\}(\hat{e}) \mid \text{prfold}(\hat{e})$
 $\mid \text{prtpl}[L](\{i \hookrightarrow \hat{e}_i\}_{i \in L}) \mid \text{prinj}[\ell](\hat{e}) \mid \text{prmatch}[n](\hat{e}; \{\hat{r}_i\}_{1 \leq i \leq n}) \mid \text{splicede}[m; n; \hat{\tau}]$
 $\text{PrRule } \hat{r} ::= \text{prerule}(p, \hat{e})$
 $\text{PrPat } \hat{p} ::= \text{prwildp} \mid \text{prfoldp}(p) \mid \text{prtplp}[L](\{i \hookrightarrow \hat{p}_i\}_{i \in L}) \mid \text{prinjp}[\ell](\hat{p}) \mid \text{splicedp}[m; n; \hat{\tau}]$

Fig. 9. Syntax of miniVerses proto-expansions. Proto-expansion terms are ABTs identified up to alpha-equivalence.

The *proto-expansion validation judgements* validate proto-expansions and simultaneously generate their final expansions:

$$\begin{array}{ll}
 \Delta \vdash^{\mathbb{T}} \hat{\tau} \rightsquigarrow \tau \text{ type} & \hat{\tau} \text{ has well-formed expansion } \tau \\
 \Delta \Gamma \vdash^{\mathbb{E}} \hat{e} \rightsquigarrow e : \tau & \hat{e} \text{ has expansion } e \text{ of type } \tau \\
 \Delta \Gamma \vdash^{\mathbb{R}} \hat{r} \rightsquigarrow r : \tau \Rightarrow \tau' & \hat{r} \text{ has expansion } r \text{ taking values of type } \tau \text{ to values of type } \tau' \\
 \hat{p} \rightsquigarrow p : \tau \dashv^{\mathbb{P}} \hat{\Gamma} & \hat{p} \text{ has expansion } p \text{ matching against } \tau \text{ generating assumptions } \hat{\Gamma}
 \end{array}$$

The purpose of the *splicing scenes* \mathbb{T} , \mathbb{E} and \mathbb{P} is to “remember”, during proto-expansion validation, the contexts and the literal body from the TLM application site (cf. Rules EE-AP-SETSM and PE-AP-SPTSM) for when validation encounters spliced terms. *Type splicing scenes*, \mathbb{T} , are of the form $\hat{\Delta}; b$. *Expression splicing scenes*, \mathbb{E} , are of the form $\hat{\Delta}; \hat{\Gamma}; \hat{\Psi}; \hat{\Phi}; b$. *Pattern splicing scenes*, \mathbb{P} , are of the form $\hat{\Delta}; \hat{\Phi}; b$.

4.6.1 Common Forms. Most of the proto-expansion forms mirror corresponding expanded forms. The rules governing proto-expansion validation for these common forms correspondingly mirror the typing rules. They are given in the supplement. Splicing scenes pass opaquely through these rules, i.e. none of these rules can access the application site contexts. This maintains context independence.

Notice that proto-rules, \hat{r} , involve expanded patterns, p , not proto-patterns, \hat{p} . The reason is that proto-rules appear in proto-expressions, which are generated by expression TLMs. Proto-patterns, in contrast, arise only from pattern TLMs. There a variable proto-pattern form.

4.6.2 References to Spliced Terms. The only interesting forms are the references to spliced unexpanded types, expressions and patterns. Let us first consider the rule for references to spliced unexpanded types:

$$\begin{array}{c}
 \text{PTV-SPLICED} \\
 \frac{
 \text{parseUTyp}(\text{subseq}(b; m; n)) = \hat{\tau} \quad \langle \mathcal{D}; \Delta_{\text{app}} \rangle \vdash \hat{\tau} \rightsquigarrow \tau \text{ type} \quad \Delta \cap \Delta_{\text{app}} = \emptyset
 }{
 \Delta \vdash \langle \mathcal{D}; \Delta_{\text{app}} \rangle; b \text{ splicedt}[m; n] \rightsquigarrow \tau \text{ type}
 }
 \end{array}$$

This first premise of this rule parses out the requested segment of the literal body, b , to produce an unexpanded type, $\hat{\tau}$. It then invokes type expansion to ensure that this unexpanded type is well-formed *under the application site context*, $\langle \mathcal{D}; \Delta_{\text{app}} \rangle$, but *not* the expansion-local type formation

context, Δ . The final premise requires that the application site type formation context is disjoint from the expansion-local type formation context. Because proto-expansions are identified up to alpha-equivalence, we can always discharge the final premise by alpha-varying the proto-expansion. Collectively, this ensures that type variable capture does not occur (we will formally state this property in the next section.)

The rule for references to spliced unexpanded expressions is fundamentally analogous:

PEV-SPLICED

$$\frac{\text{parseUExp}(\text{subseq}(b; m; n)) = \hat{e} \quad \emptyset \vdash^{\hat{\Delta}; b} \dot{\tau} \rightsquigarrow \tau \text{ type} \quad \langle \mathcal{D}; \Delta_{\text{app}} \rangle \langle \mathcal{G}; \Gamma_{\text{app}} \rangle \vdash_{\hat{\Psi}, \hat{\Phi}} \hat{e} \rightsquigarrow e : \tau \quad \Delta \cap \Delta_{\text{app}} = \emptyset \quad \text{dom}(\Gamma) \cap \text{dom}(\Gamma_{\text{app}}) = \emptyset}{\Delta \Gamma \vdash \langle \mathcal{D}; \Delta_{\text{app}} \rangle; \langle \mathcal{G}; \Gamma_{\text{app}} \rangle; \hat{\Psi}; \hat{\Phi}; b \text{ spliced}[m; n; \dot{\tau}] \rightsquigarrow e : \tau}$$

We first splice out the requested segment. The second premise expands the type annotation under an empty context, because the type annotation must be meaningful at the application site (so, independent of Δ and Γ) and not itself make any assumptions about the application site context. Spliced types can appear in the annotation. The third premise performs typed expansion of the spliced unexpanded expression under the application site contexts, but not the expansion-local contexts. The final two premises ensure that these contexts are disjoint, again to force capture avoidance.

The rule for references to spliced unexpanded patterns is entirely analogous:

PPV-SPLICED

$$\frac{\text{parseUPat}(\text{subseq}(b; m; n)) = \hat{p} \quad \emptyset \vdash^{\hat{\Delta}; b} \dot{\tau} \rightsquigarrow \tau \text{ type} \quad \hat{\Delta} \vdash_{\hat{\Phi}} \hat{p} \rightsquigarrow p : \tau \dashv \hat{\Gamma}}{\text{splicedp}[m; n; \dot{\tau}] \rightsquigarrow p : \tau \dashv^{\hat{\Delta}; \hat{\Phi}; b} \hat{\Gamma}}$$

4.7 Metatheory

4.7.1 Typed Expression Expansion. Let us now return to Theorem 4.2, the typed expression expansion theorem which was mentioned at the end of Sec. 4.3. As it turns out, in order to prove this theorem, we must prove the following stronger theorem, because the proto-expression validation judgement is defined mutually inductively with the typed expansion judgement (due to the three rules just described.)

THEOREM 4.3 (TYPED EXPRESSION EXPANSION (STRONG)).

- (1) If $\langle \mathcal{D}; \Delta \rangle \langle \mathcal{G}; \Gamma \rangle \vdash_{\langle \mathcal{A}; \Psi \rangle} \hat{e} \rightsquigarrow e : \tau$ then $\Delta \Gamma \vdash e : \tau$.
- (2) If $\Delta \Gamma \vdash \langle \mathcal{D}; \Delta_{\text{app}} \rangle; \langle \mathcal{G}; \Gamma_{\text{app}} \rangle; \langle \mathcal{A}; \Psi \rangle; b \dot{e} \rightsquigarrow e : \tau$ and $\Delta \cap \Delta_{\text{app}} = \emptyset$ and $\text{dom}(\Gamma) \cap \text{dom}(\Gamma_{\text{app}}) = \emptyset$ then $\Delta \cup \Delta_{\text{app}} \Gamma \cup \Gamma_{\text{app}} \vdash e : \tau$.

The additional second clause simply states that the final expansion produced by proto-expression validation is well-typed under the combined application site and expansion-internal context (spliced terms are distinguished only in the proto-expansion.) Such combined contexts can only be formed if the constituents are disjoint.

The proof proceeds by mutual rule induction and appeal to simple lemmas about type expansion and proto-type validation (see supplement). The proof is straightforward but for one issue: it is not immediately clear that the mutual induction is well-founded, because the case in the proof of part 2 for Rule PEV-SPLICED invokes part 1 of the induction hypothesis on a term that is not a sub-term of the conclusion, but rather parsed out of the literal body, b . To establish that the mutual induction is well-founded, then, we need to explicitly establish a decreasing metric. The intuition is that parsing a term out of a literal body cannot produce a bigger term than the term that contained that very literal body. More specifically, the sum of the lengths of the literal bodies that appear in the term strictly decreases each time you perform a nested TLM application because some portion

of the term has to be consumed by the TLM name and the delimiters. The details are given in the supplemental material.

4.7.2 Typed Pattern Expansion. A similar argument is needed to prove Typed Pattern Expansion:

THEOREM 4.4 (TYPED PATTERN EXPANSION (STRONG)).

- (1) If $\langle \mathcal{D}; \Delta \rangle \vdash_{\langle \mathcal{A}; \Phi \rangle} \hat{p} \rightsquigarrow p : \tau \dashv \langle \mathcal{G}; \Gamma \rangle$ then $\Delta \vdash p : \tau \dashv \Gamma$.
- (2) If $\hat{p} \rightsquigarrow p : \tau \dashv \langle \mathcal{D}; \Delta \rangle; \langle \mathcal{A}; \Phi \rangle; b \langle \mathcal{G}; \Gamma \rangle$ then $\Delta \vdash p : \tau \dashv \Gamma$.

4.7.3 seTLM Reasoning Principles. The following theorem summarizes the abstract reasoning principles that programmers can rely on when applying an seTLM. Informal descriptions of important clauses are given inline.

THEOREM 4.5 (seTLM ABSTRACT REASONING PRINCIPLES). If $\langle \mathcal{D}; \Delta \rangle \langle \mathcal{G}; \Gamma \rangle \vdash_{\hat{\Psi}} \hat{a} \text{ ‘} b \text{’} \rightsquigarrow e : \tau$ then:

- (1) (**Typing 1**) $\hat{\Psi} = \hat{\Psi}'$, $\hat{a} \rightsquigarrow a \hookrightarrow \text{setlm}(\tau; e_{\text{parse}})$ and $\Delta \Gamma \vdash e : \tau$
The type of the expansion is consistent with the type annotation on the applied seTLM definition.
- (2) $b \downarrow_{\text{Body}} e_{\text{body}}$
- (3) $e_{\text{parse}}(e_{\text{body}}) \Downarrow \text{inj}[\text{SuccessE}](e_{\text{proto}})$
- (4) $e_{\text{proto}} \Uparrow_{\text{PrExpr}} \hat{e}$
- (5) (**Segmentation**) $\text{seg}(\hat{e})$ segments b
The segmentation determined by the proto-expansion actually segments the literal body (i.e. each segment is in-bounds and the segments are non-overlapping.)
- (6) $\text{seg}(\hat{e}) = \{\text{splicedt}[m'_i; n'_i]\}_{0 \leq i < n_{\text{ty}}} \cup \{\text{splicede}[m_i; n_i; \hat{\tau}_i]\}_{0 \leq i < n_{\text{exp}}}$
- (7) (**Typing 2**) $\{\langle \mathcal{D}; \Delta \rangle \vdash \text{parseUTyp}(\text{subseq}(b; m'_i; n'_i)) \rightsquigarrow \tau'_i \text{ type}\}_{0 \leq i < n_{\text{ty}}}$ and $\{\Delta \vdash \tau'_i \text{ type}\}_{0 \leq i < n_{\text{ty}}}$
Each spliced type has a well-formed expansion at the application site.
- (8) (**Typing 3**) $\{\emptyset \vdash \langle \mathcal{D}; \Delta \rangle; b \hat{\tau}_i \rightsquigarrow \tau_i \text{ type}\}_{0 \leq i < n_{\text{exp}}}$ and $\{\Delta \vdash \tau_i \text{ type}\}_{0 \leq i < n_{\text{exp}}}$
Each type annotation on a reference to a spliced expression has a well-formed expansion at the application site.
- (9) (**Typing 4**) $\{\langle \mathcal{D}; \Delta \rangle \langle \mathcal{G}; \Gamma \rangle \vdash_{\hat{\Psi}} \text{parseUExp}(\text{subseq}(b; m_i; n_i)) \rightsquigarrow e_i : \tau_i\}_{0 \leq i < n_{\text{exp}}}$ and $\{\Delta \Gamma \vdash e_i : \tau_i\}_{0 \leq i < n_{\text{exp}}}$
Each spliced expression has a well-typed expansion consistent with its type annotation.
- (10) (**Capture Avoidance**) $e = [\{\tau'_i / t_i\}_{0 \leq i < n_{\text{ty}}}, \{e_i / x_i\}_{0 \leq i < n_{\text{exp}}}]e'$ for some $\{t_i\}_{0 \leq i < n_{\text{ty}}}$ and $\{x_i\}_{0 \leq i < n_{\text{exp}}}$ and e'
The final expansion can be decomposed into a term with variables in place of each spliced type or expression. The expansions of these spliced types and expressions can be substituted into this term in the standard capture avoiding manner.
- (11) (**Context Independence**) $\text{fv}(e') \subset \{t_i\}_{0 \leq i < n_{\text{ty}}} \cup \{x_i\}_{0 \leq i < n_{\text{exp}}}$
The aforementioned decomposed term makes no mention of bindings in the application site context, i.e. the only free variables are those corresponding to the spliced terms.

The proof, which involves auxiliary lemmas about the decomposition of proto-types and proto-expressions, is given in the supplement.

4.7.4 spTLM Reasoning Principles. Finally, the following theorem summarizes the abstract reasoning principles available to programmers when applying an spTLM. Most of the labeled clauses are analogous to those described above, so we omit their descriptions.

THEOREM 4.6 (SPTLM ABSTRACT REASONING PRINCIPLES). If $\hat{\Delta} \vdash_{\hat{\Phi}} \hat{a} \text{ 'b' } \rightsquigarrow p : \tau \Vdash \hat{\Gamma}$ where $\hat{\Delta} = \langle \mathcal{D}; \Delta \rangle$ and $\hat{\Gamma} = \langle \mathcal{G}; \Gamma \rangle$ then all of the following hold:

- (1) (**Typing 1**) $\hat{\Phi} = \hat{\Phi}', \hat{a} \rightsquigarrow a \hookrightarrow \text{sptlm}(\tau; e_{\text{parse}})$ and $\Delta \vdash p : \tau \Vdash \Gamma$
- (2) $b \downarrow_{\text{Body}} e_{\text{body}}$
- (3) $e_{\text{parse}}(e_{\text{body}}) \Downarrow \text{inj}[\text{SuccessP}](e_{\text{proto}})$
- (4) $e_{\text{proto}} \uparrow_{\text{PrPat}} \hat{p}$
- (5) (**Segmentation**) $\text{seg}(\hat{p})$ segments b
- (6) $\text{seg}(\hat{p}) = \{\text{splicedt}[n'_i; m'_i]\}_{0 \leq i < n_{\text{ty}}} \cup \{\text{splicedp}[m_i; n_i; \hat{\tau}_i]\}_{0 \leq i < n_{\text{pat}}}$
- (7) (**Typing 2**) $\{\hat{\Delta} \vdash \text{parseUTyp}(\text{subseq}(b; m'_i; n'_i)) \rightsquigarrow \tau'_i \text{ type}\}_{0 \leq i < n_{\text{ty}}}$ and $\{\Delta \vdash \tau'_i \text{ type}\}_{0 \leq i < n_{\text{ty}}}$
- (8) (**Typing 3**) $\{\emptyset \vdash_{\hat{\Delta}; b} \hat{\tau}_i \rightsquigarrow \tau_i \text{ type}\}_{0 \leq i < n_{\text{pat}}}$ and $\{\Delta \vdash \tau_i \text{ type}\}_{0 \leq i < n_{\text{pat}}}$
- (9) (**Typing 4**) $\{\hat{\Delta} \vdash_{\hat{\Phi}} \text{parseUPat}(\text{subseq}(b; m_i; n_i)) \rightsquigarrow p_i : \tau_i \Vdash \langle \mathcal{G}_i; \Gamma_i \rangle\}_{0 \leq i < n_{\text{pat}}}$ and $\{\Delta \vdash p_i : \tau_i \Vdash \Gamma_i\}_{0 \leq i < n_{\text{pat}}}$
- (10) (**No Hidden Bindings**) $\mathcal{G} = \biguplus_{0 \leq i < n_{\text{pat}}} \mathcal{G}_i$ and $\Gamma = \bigcup_{0 \leq i < n_{\text{pat}}} \Gamma_i$

The hypotheses generated by the TLM application are exactly those generated by the spliced patterns.

The proof, in the supplement, relies on an auxiliary lemma about decomposing proto-patterns.

5 Parametric TLMs (pTLMs)

Simple TLMs operate at a single specified type and the proto-expansions they generate must be closed. This simplifies matters formally, but it is not convenient in practice. This section introduces *parametric TLMs*. Parametric TLMs can be defined over a type- and module-parameterized family of types, and the proto-expansions they generate can refer to supplied type and module parameters.

5.1 Parametric TLMs By Example

Consider the following ML signature specifying an abstract data type (Harper 1997; Liskov and Zilles 1974) of string-keyed polymorphic dictionaries:

```
signature DICT = sig
  type t('a)
  val empty : t('a)
  val extend : t('a) -> string * 'a -> t('a)
  (* ... *)
end
```

Let us now define a TLM that is parametric over all implementations of this signature, $D : \text{DICT}$, and also over all choices of type 'a:

```
syntax $dict (D : DICT) (type 'a) at D.t('a) by static
  fn(b : body) -> parse_result(proto_expr) => (* ... *)
end
```

For example, assuming some implementation $\text{Dict} : \text{DICT}$ and some values $v1 : \text{int}$ and $v2 : \text{int}$ (and assuming that VerseML offers a generalized literal form delimited by $\{|$ and $| \}$), we can apply $\$dict$ as follows:

```
$dict Dict int {| "key1" => v1; "key2" => v2 |}
```

Notice that the segmentation immediately reveals which punctuation is particular to this TLM and where the spliced key and value expressions appear. Because the context-free syntax of unexpanded terms is never modified, it is possible to reason modularly about syntactic determinism (i.e. we can reason above that \Rightarrow does not appear in the follow set of unexpanded terms, so there can never be an ambiguity about where a key expression ends.)

If we will use the Dict implementation ubiquitously, we can abbreviate the partial application of \$dict to Dict, resulting in a TLM that is parametric over only the type 'a, as follows:

```
let syntax $Dict = $dict Dict in
$Dict int { | "key1" => v1; "key2" => v2 | }
```

The proto-expansion generated for the TLM application above might be:

```
D.extend (D.extend D.empty (spliced<1; 6; string>, spliced<11; 12; 'a>))
(spliced<15; 20; string>, spliced<25; 26; 'a>)
```

The generated proto-expansion must truly be parametric, i.e. it must be valid for all modules $D : \text{Dict}$ and types 'a. It is only after proto-expansion validation that we substitute the actual parameters, here Dict for D and int for 'a, into the final expansion. Additionally, substitution occurs on the type annotations on references to spliced terms before recursively expanding those terms (otherwise, the expressions v1 and v2 above would need to be well-typed for all types, 'a.)

Module parameters are also useful for giving the expansion access to helper functions in a context-independent manner. For example, in Sec. 2.3.2 we discussed the problem of applying datatype constructors like H1Element, because they are introduced into the context as variables in most ML-like languages. With module parameters, it is possible to sidestep this problem by passing in a module containing the datatype constructors. In practice, the language might implicitly construct and open such a module within the expansion when the type annotation on the TLM is such a datatype.

It is important to note that module parameters are not accessible by parse functions directly, because the applied module parameters will not have been dynamically instantiated when typed expansion occurs. Parameters are accessible only from within the expansion that the parse function generates. We will discuss a distinct mechanism for providing helper functions to parse functions in Sec. 6.

5.2 Parametric TLMs, Formally

We will now outline miniVerse_P, a calculus that extends miniVerse_S with support for parametric TLMs. This calculus is organized, like miniVerse_S, as an unexpanded language (UL) defined by typed expansion to an expanded language (XL). There is not enough space to describe miniVerse_P with the same level of detail as in Sec. 4, so we highlight only the most important concepts below.

The XL consists of 1) module expressions, M , classified by signatures, σ ; 2) constructions, c , classified by kinds, κ ; and 3) expressions classified by types, which are constructions of kind Type (we use metavariables τ instead of c for types.) Metavariables X ranges over module variables and u or t over construction variables. The module and construction languages are based closely on those defined by Harper (2012), which in turn are based on early work by MacQueen (1984, 1986), subsequent work on the phase splitting interpretation of modules (Harper et al. 1989) and on using dependent singleton kinds to track type identity (Crary 2009; Stone and Harper 2006), and finally on formal developments by Dreyer (2005) and Lee et al. (2007). A complete account of these developments is unfortunately beyond the scope of this paper. The expression language extends the language of miniVerse_S only to allow projection out of modules.

The main conceptual difference between miniVerse_S and miniVerse_P is that miniVerse_P introduces the notion of unexpanded and expanded TLM expressions and types, as shown in Figure 10.

$$\begin{array}{ll} \text{UMType } \hat{\rho} ::= \hat{\tau} \mid \forall \hat{X} : \hat{\sigma}. \hat{\rho} & \text{MType } \rho ::= \text{type}(\tau) \mid \text{allmod}\{\sigma\}(X.\rho) \\ \text{UMExp } \hat{\epsilon} ::= \hat{a} \mid \wedge \hat{X} : \hat{\sigma}. \hat{\epsilon} \mid \hat{\epsilon}(\hat{X}) & \text{MExp } \epsilon ::= \text{defref}[a] \mid \text{absmod}\{\sigma\}(X.\epsilon) \mid \text{apmod}\{X\}(\epsilon) \end{array}$$

Fig. 10. Syntax of unexpanded and expanded TLM types and expressions in miniVerse_P

This is necessary to express abstraction over parameters and application of parameters. For simplicity, we formalize only module parameters. Type parameters can be realized as module parameters having exactly one abstract type member.

The rule governing expression TLM application, reproduced below, touches all of the main ideas in miniVersep.

$$\begin{array}{c}
 \text{EE-AP-PETSM} \\
 \frac{
 \begin{array}{l}
 \hat{\Omega} = \langle \mathcal{M}; \mathcal{D}; \mathcal{G}; \Omega_{\text{app}} \rangle \quad \hat{\Psi} = \langle \mathcal{A}; \Psi \rangle \\
 \hat{\Omega} \vdash_{\Psi}^{\text{Exp}} \hat{e} \rightsquigarrow \epsilon @ \text{type}(\tau_{\text{final}}) \quad \Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \epsilon \Downarrow \epsilon_{\text{normal}} \\
 \text{tlmdef}(\epsilon_{\text{normal}}) = a \quad \Psi = \Psi', a \hookrightarrow \text{petlm}(\rho; e_{\text{parse}}) \\
 b \downarrow_{\text{Body}} e_{\text{body}} \quad e_{\text{parse}}(e_{\text{body}}) \Downarrow \text{inj}[\text{SuccessE}](e_{\text{pproto}}) \quad e_{\text{pproto}} \uparrow_{\text{PPrExpr}} \dot{e} \\
 \Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \dot{e} \hookrightarrow_{\epsilon_{\text{normal}}} \dot{e} ? \text{type}(\tau_{\text{proto}}) \dashv \omega : \Omega_{\text{params}} \\
 \text{seg}(\dot{e}) \text{ segments } b \quad \Omega_{\text{params}} \vdash^{\omega: \Omega_{\text{params}}; \hat{\Omega}; \hat{\Psi}; \hat{\Phi}; b} \dot{e} \rightsquigarrow e : \tau_{\text{proto}}
 \end{array}
 }{
 \hat{\Omega} \vdash_{\hat{\Psi}; \hat{\Phi}} \hat{e} 'b' \rightsquigarrow [\omega]e : [\omega]\tau_{\text{proto}}
 }
 \end{array}$$

The first two premises simply deconstruct the (unified) unexpanded context $\hat{\Omega}$ (which tracks the expansion of expression, constructor and module identifiers, as $\hat{\Lambda}$ and $\hat{\Gamma}$ did in miniVersep) and peTLM context, $\hat{\Psi}$. Next, we expand \hat{e} according to straightforward unexpanded peTLM expression expansion rules. The resulting TLM expression, ϵ , must be defined at a type (i.e. no quantification must remain.)

The fourth premise performs *peTLM expression normalization*. Normalization, $\Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \epsilon \Downarrow \epsilon_{\text{normal}}$, is defined in terms of a simple structural dynamics with two stepping rules:

$$\begin{array}{c}
 \text{EPS-DYN-APMOD-SUBST-E} \\
 \frac{
 \Omega \vdash_{\Psi}^{\text{Exp}} \text{apmod}\{X\}(\text{absmod}\{\sigma\}(X'.\epsilon)) \mapsto [X/X']\epsilon \quad \Omega \vdash_{\Psi}^{\text{Exp}} \text{apmod}\{X\}(\epsilon) \mapsto \text{apmod}\{X\}(\epsilon')
 }{
 \Omega \vdash_{\Psi}^{\text{Exp}} \epsilon \mapsto \epsilon'
 }
 \end{array}$$

Normalization eliminates parameters introduced in higher-order abbreviations, leaving only those parameter applications specified by the original TLM definition. Normal forms and progress and preservation theorems are established in the supplement.

The third row of premises looks up the applied TLM's definition by first invoking a simple metafunction to extract the name, then looking up this name within the peTLM definition context, Ψ .

The fourth row of premises 1) encodes the body as a value of the type *Body*; 2) applies the parse function; and 3) decodes the result, producing a *parameterized proto-expression*, \dot{e} . Parameterized proto-expressions, \dot{e} , are ABTs that serve simply to introduce the parameter bindings into an underlying proto-expression, \dot{e} . The syntax of parameterized proto-expressions is given below.

$$\text{PPrExpr } \dot{e} ::= \text{prexp}(\dot{e}) \mid \text{prbindmod}(X.\dot{e})$$

There must be one binder in \dot{e} for each TLM parameter specified by $\text{tlmdef}(\epsilon_{\text{normal}})$. (VersepML can insert these binders automatically as a convenience, but we consider only the underlying mechanism in this core calculus.)

The judgement on the fifth row of Rule EE-AP-PETSM then *deparameterizes* \dot{e} by peeling away these binders to produce 1) the underlying proto-expression, \dot{e} , with the variables that stand for the parameters free; 2) a corresponding deparameterized type, τ_{proto} , that uses the same free variables to stand for the parameters; 3) a *substitution*, ω , that pairs the applied parameters from ϵ_{normal} with the corresponding variables generated when peeling away the binders in \dot{e} ; and 4) a corresponding *parameter context*, Ω_{params} , that tracks the signatures of these variables. The two rules governing

the proto-expression deparameterization judgement are below:

$$\frac{\Omega_{\text{app}} \vdash_{\Psi, a \hookrightarrow \text{petlm}(\rho; e_{\text{parse}})}^{\text{Exp}} \text{prexp}(\hat{e}) \rightsquigarrow_{\text{defref}[a]} \hat{e} ? \rho \dashv \emptyset : \emptyset}{\Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \hat{e} \rightsquigarrow_{\epsilon} \hat{e} ? \text{allmods}\{\sigma\}(X.\rho) \dashv \omega : \Omega \quad X \notin \text{dom}(\Omega_{\text{app}})}{\Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \text{prbindmod}(X.\hat{e}) \rightsquigarrow_{\text{apmod}\{X'\}(\epsilon)} \hat{e} ? \rho \dashv (\omega, X'/X) : (\Omega, X : \sigma)}$$

This judgement can be pronounced “when applying peTLM ϵ , \hat{e} has deparameterization \hat{e} leaving ρ with parameter substitution ω ”. Notice based on the second rule that every module binding in \hat{e} must pair with a corresponding module parameter application. Moreover, the variables standing for parameters must not appear in Ω_{app} , i.e. $\text{dom}(\Omega_{\text{params}})$ must be disjoint from $\text{dom}(\Omega_{\text{app}})$ (this requirement can always be discharged by alpha-variation.)

The final row of premises checks that the segmentation of \hat{e} is valid and performs proto-expansion validation under the parameter context, Ω_{param} (rather than the empty context, as was the case in miniVerseq.) The conclusion of the rule applies the parameter substitution, ω , to the resulting expression and the deparameterized type.

Proto-expansion validation operates conceptually as in miniVerseq. The only subtlety has to do with the type annotations on references to spliced terms. As described at the end of Sec. 5.1, these annotations might refer to the parameters, so the parameter substitution, ω , which is tracked by the splicing scene, must be applied to the type annotation before proceeding recursively to expand the referenced unexpanded term. However, the spliced term itself must treat parameters parametrically, so the substitution is not applied in the conclusion of the rule:

$$\frac{\text{P-PEV-SPLICED} \quad \text{parseUExp}(\text{subseq}(b; m; n)) = \hat{e} \quad \Omega_{\text{params}} \vdash^{\omega; \Omega_{\text{params}}; \hat{\Omega}; b} \hat{\tau} \rightsquigarrow \tau :: \text{Type} \quad \hat{\Omega} \vdash_{\hat{\Psi}; \hat{\Phi}} \hat{e} \rightsquigarrow e : [\omega]\tau \quad \hat{\Omega} = \langle \mathcal{M}; \mathcal{D}; \mathcal{G}; \Omega_{\text{app}} \rangle \quad \text{dom}(\Omega) \cap \text{dom}(\Omega_{\text{app}}) = \emptyset}{\Omega \vdash^{\omega; \Omega_{\text{params}}; \hat{\Omega}; \hat{\Psi}; \hat{\Phi}; b} \text{splicede}[m; n; \hat{\tau}] \rightsquigarrow e : \tau}$$

(This is only sensible because we maintain the invariant that Ω is always an extension of Ω_{params} .)

The calculus enjoys metatheoretic properties analogous to those described in Sec. 4.7, modified to account for the presence of modules, kinds and parameterization. The following theorem establishes the abstract reasoning principles available when applying a parametric expression TLM.

THEOREM 5.1 (PETLM REASONING PRINCIPLES). *If $\hat{\Omega} \vdash_{\hat{\Psi}; \hat{\Phi}} \hat{e} \rightsquigarrow b' \rightsquigarrow e : \tau$ then:*

- (1) $\hat{\Omega} = \langle \mathcal{M}; \mathcal{D}; \mathcal{G}; \Omega_{\text{app}} \rangle$
- (2) $\hat{\Psi} = \langle \mathcal{A}; \Psi \rangle$
- (3) (**Typing 1**) $\hat{\Omega} \vdash_{\hat{\Psi}}^{\text{Exp}} \hat{e} \rightsquigarrow \epsilon @ \text{type}(\tau)$ and $\Omega_{\text{app}} \vdash e : \tau$
- (4) $\Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \epsilon \Downarrow \epsilon_{\text{normal}}$
- (5) $\text{tlmdef}(\epsilon_{\text{normal}}) = a$
- (6) $\Psi = \Psi', a \hookrightarrow \text{petlm}(\rho; e_{\text{parse}})$
- (7) $b \downarrow_{\text{Body}} e_{\text{body}}$
- (8) $e_{\text{parse}}(e_{\text{body}}) \Downarrow \text{inj}[\text{SuccessE}](e_{\text{pproto}})$
- (9) $e_{\text{pproto}} \uparrow_{\text{PPrExpr}} \hat{e}$
- (10) $\Omega_{\text{app}} \vdash_{\Psi}^{\text{Exp}} \hat{e} \rightsquigarrow_{\epsilon_{\text{normal}}} \hat{e} ? \text{type}(\tau_{\text{proto}}) \dashv \omega : \Omega_{\text{params}}$
- (11) (**Segmentation**) $\text{seg}(\hat{e})$ segments b
- (12) $\Omega_{\text{params}} \vdash^{\omega; \Omega_{\text{params}}; \hat{\Omega}; \hat{\Psi}; \hat{\Phi}; b} \hat{e} \rightsquigarrow e' : \tau_{\text{proto}}$
- (13) $e = [\omega]e'$
- (14) $\tau = [\omega]\tau_{\text{proto}}$

- (15) $\text{seg}(\dot{e}) = \{\text{splicedk}[m_i; n_i]\}_{0 \leq i < n_{\text{kind}}} \cup \{\text{splicedc}[m'_i; n'_i; \dot{\kappa}'_i]\}_{0 \leq i < n_{\text{con}}} \cup \{\text{splicedc}[m''_i; n''_i; \dot{\tau}_i]\}_{0 \leq i < n_{\text{exp}}}$
- (16) (**Kinding 1**) $\{\hat{\Omega} \vdash \text{parseUKind}(\text{subseq}(b; m_i; n_i)) \rightsquigarrow \kappa_i \text{ kind}\}_{0 \leq i < n_{\text{kind}}} \text{ and } \{\Omega_{\text{app}} \vdash \kappa_i \text{ kind}\}_{0 \leq i < n_{\text{kind}}}$
- (17) (**Kinding 2**) $\{\Omega_{\text{params}} \vdash \omega; \Omega_{\text{params}}; \hat{\Omega}; b \ \dot{\kappa}'_i \rightsquigarrow \kappa'_i \text{ kind}\}_{0 \leq i < n_{\text{con}}} \text{ and } \{\Omega_{\text{app}} \vdash [\omega] \kappa'_i \text{ kind}\}_{0 \leq i < n_{\text{con}}}$
- (18) (**Kinding 3**) $\{\hat{\Omega} \vdash \text{parseUCon}(\text{subseq}(b; m'_i; n'_i)) \rightsquigarrow c_i :: [\omega] \kappa'_i\}_{0 \leq i < n_{\text{con}}} \text{ and } \{\Omega_{\text{app}} \vdash c_i :: [\omega] \kappa'_i\}_{0 \leq i < n_{\text{con}}}$
- (19) (**Kinding 4**) $\{\Omega_{\text{params}} \vdash \omega; \Omega_{\text{params}}; \hat{\Omega}; b \ \dot{\tau}_i \rightsquigarrow \tau_i :: \text{Type}\}_{0 \leq i < n_{\text{exp}}} \text{ and } \{\Omega_{\text{app}} \vdash [\omega] \tau_i :: \text{Type}\}_{0 \leq i < n_{\text{exp}}}$
- (20) (**Typing 2**) $\{\hat{\Omega} \vdash_{\hat{\Psi}; \hat{\Phi}} \text{parseUExp}(\text{subseq}(b; m''_i; n''_i)) \rightsquigarrow e_i : [\omega] \tau_i\}_{0 \leq i < n_{\text{exp}}} \text{ and } \{\Omega_{\text{app}} \vdash e_i : [\omega] \tau_i\}_{0 \leq i < n_{\text{exp}}}$
- (21) (**Capture Avoidance**) $e = [\{\kappa_i / k_i\}_{0 \leq i < n_{\text{kind}}}, \{c_i / u_i\}_{0 \leq i < n_{\text{con}}}, \{e_i / x_i\}_{0 \leq i < n_{\text{exp}}}, \omega] e''$ for some e'' and fresh $\{k_i\}_{0 \leq i < n_{\text{kind}}}$ and fresh $\{u_i\}_{0 \leq i < n_{\text{con}}}$ and fresh $\{x_i\}_{0 \leq i < n_{\text{exp}}}$
- (22) (**Context Independence**) $\text{fv}(e'') \subset \{k_i\}_{0 \leq i < n_{\text{kind}}} \cup \{u_i\}_{0 \leq i < n_{\text{con}}} \cup \{x_i\}_{0 \leq i < n_{\text{exp}}} \cup \text{dom}(\Omega_{\text{params}})$

We leave out the details of parametric pattern TLMs (ppTLMs), but they should again follow analogously.

6 Static Evaluation

Thus far, we have assumed that the parse functions in TLM definitions are closed expanded expressions. This assumption simplifies matters formally, but in practice it is quite difficult, because this means that TLM definitions cannot themselves access any libraries nor use TLMs internally. To address this problem, we need to introduce the concept of a *static environment*.

6.1 The Static Environment

Figure 11 shows an example of a module, `ParserCombos`, that defines a number of parser combinators following Hutton (1992). The **static** qualifier indicates that this module is bound for use only within similarly qualified values, including in particular the parse functions of the subsequent TLM definitions.

```

1  static module ParserCombos = struct
2    type parser('c, 't) = list('c) -> list('t * list('c))
3    val alt : parser('c, 't) -> parser('c, 't) -> parser('c, 't)
4    (* ... *)
5  end
6  syntax $a at t by static fn(b) => (* OK *) ParserCombos.alt (* ... *) end
7  static val y = (* OK *) ParserCombos.alt (* ... *)
8  val z = (* neither ParserCombos nor y are available here *)

```

Fig. 11. Binding a static module for use within parse functions.

The values that arise during the static evaluation (in particular, of parse functions) do not themselves persist from “compile-time” to “run-time”, so we do not need a full staged computation system, e.g. as described by Taha (1999). Instead, a sequence of static bindings operates like a read-evaluate-print loop (REPL) scoped according to the program structure, in that each static expression is evaluated immediately and the evaluated values are tracked by a *static environment*. The static environment is discarded at the end of typed expansion.

A language designer might choose to restrict the external effects available to static terms in some way, e.g. to ensure deterministic builds. It might also be helpful to restrict mutable state shared between TLMs to prevent undesirable TLM application order dependencies. On the other hand, these features, if used for the purposes of caching in a disciplined manner, might speed up typed expansion. These are orthogonal design decisions.

6.2 Applying TLMs Within TLM Definitions

TLMs and TLM abbreviations can also be qualified with the **static** keyword, which marks them for use within subsequent static expressions and patterns. Let us consider some examples of relevance to TLM providers.

6.2.1 Quasiquotation. TLMs must construct values of type `proto_expr` or `proto_pat`. Constructing values of these types explicitly can have high syntactic cost. To decrease this cost, we can define TLMs that provide support for *quasiquotation syntax* similar to that built in to languages like Lisp (Bawden 1999) and Scala (Shabalin et al. 2013). The following TLM defines quasi-quotation for encodings of proto-expressions:

```
static syntax $proto_expr at proto_expr by static fn(b) => (* ... *) end
```

For example, `$proto_expr `g x `x`` might have expansion `App(App(Var "g", Var "x"), x)`. Notice that prefixing a variable (or parenthesized expression) with ``` serves to splice in its value, which here must be of type `proto_expr` (though in other syntactic positions, it might be `proto_typ`.) This is also known as *anti-quotation*.

6.2.2 Parser Generators. Abstractly, a grammar-based parser generator is a module matching the signature `PARSEGEN` defined below:

```
signature PARSEGEN = sig
  type grammar('a)
  (* ... operations on grammars ... *)
  val generate : grammar('a) -> (body -> parse_result('a))
end
```

Rather than constructing a grammar (equipped with semantic actions) using the associated operations (whose specifications are elided in `PARSEGEN`), we wish to use a syntax for context-free grammars that follows standard conventions. We can do so by defining a static parametric TLM `$grammar`:

```
static syntax $grammar (P : PARSEGEN) 'a at P.grammar('a) by (* ... *) end
```

To support splicing, we need non-terminals that recognize unexpanded terms and produce the corresponding splice reference, rather than the AST itself. This would require that the parser generator keep track of location information (as most production-grade parser generators already do for error reporting.) For spliced expressions, this non-terminal would need to be a family of non-terminals indexed by a value of type `proto_typ`.

A grammar containing such non-terminals can serve as a *summary specification* – a human can simply take this grammar as a specification of what the splice summary will be for every recognized string, rather than relying on an editor to communicate this information. The associated semantic actions can be held abstract as long as the system performs a simple check to ensure that the generated proto-expansion does in fact mention each spliced expression from the corresponding production.

6.3 Static Evaluation, Formally

It is not difficult to extend `miniVersep` to account for static evaluation. Static environments, Σ , take the form $\omega : \hat{\Omega}; \hat{\Psi}; \hat{\Phi}$, where ω is a substitution. Each binding form is annotated with a *phase*, ϕ , either static or standard. The rules for binding forms annotated with standard are essentially unchanged, differing only in that Σ passes through opaquely. The rules for binding forms annotated with static are based on the corresponding standard phase rules, differing only in that 1) they operate on Σ and 2) evaluation occurs immediately. Finally, the forms for TLM definition are

modified so that the parse function is now an unexpanded, rather than an expanded expression. The substitution ω is applied to the parse function after it is expanded. The full details are defined as a minor patch of `miniVerseP` called `miniVersePH` in the supplement.

6.4 Library Management

In the examples above, and in our formal treatment, we explicitly qualified various definitions with the **static** keyword to make them available within static values. In practice, we would like to be able to use libraries within both static values and standard values as needed without duplicating code. This can be accomplished either by the package manager (e.g. SML/NJ's CM (Blume 2002), extended with phase annotations) or by allowing one to explicitly lower a new instance of a module imported into the standard phase for use also in the static phase, as in a recent proposal for modular staging macros in OCaml (Yallop and White 2015). We will leave these details as design decisions for the reader to consider.

TLMs definitions can be exported by packages, but they cannot be exported from within modules because that would require that they also appear in signatures, and that, in turn, would complicate reasoning about signature equivalence, since TLM definitions contain arbitrary parse functions. It would also bring in confusion about whether the generated expansions can use knowledge about type identity available within the module enclosing the TLM definition. That said, it should be possible to export TLM *abbreviations* from modules, since they refer to TLM definitions only through symbolic names. We have not yet formalized this intuition.

7 Related Work

7.1 Syntax Definition Systems

One approach available to library providers seeking to introduce new literal forms is to use a syntax definition system to construct a library-specific *syntax dialect*: a new syntax definition that extends the syntax definition given in the language definition with new forms, including literal forms. For example, Ur/Web's syntax (Fig. 1) is a library-specific dialect of Ur's syntax (Chlipala 2010, 2015).

There are hundreds of syntax definition systems of various design. Notable examples include grammar-oriented systems like `Camlp4` (Leroy et al. 2014), `Copper` (Schwerdfeger and Van Wyk 2009; Van Wyk and Schwerdfeger 2007), `SugarJ/Sugar*/SoundExt` (Erdweg et al. 2011; Erdweg and Rieger 2013; Lorenzen and Erdweg 2013, 2016), as well as various parser combinator systems (Hutton 1992). The parsers generated by these systems can be invoked to preprocess program text in various ways, e.g. by invoking them from within a build script, by using a preprocessor-aware build system (e.g. `ocamlbuild`), or via language-integrated preprocessing directives (e.g. Racket's `#lang` directive or its reader macros (Flatt 2012), or the import mechanism of `SugarJ/Sugar*/SoundExt`.)

The first problem with using a syntax definition system to directly extend the context-free syntax of a language is that clients cannot always deterministically combine the resulting extended syntax dialects when they want to use the new forms that they define together, i.e. there can be syntactic conflicts. This is a significant problem because large programs use many separately developed libraries (Lämmel et al. 2011) and often do not fall cleanly into a single "problem domain".

One possible solution is found in `Copper`, which integrates a modular grammar analysis that guarantees that determinism is conserved when syntax dialects of a certain restricted class are combined (Schwerdfeger and Van Wyk 2009). The caveat is that the constituent dialects must prefix all newly introduced forms with marking tokens drawn from disjoint sets. To be confident that the marking tokens used are disjoint, providers must base them on the domain name system or some other coordinating entity. Because the mechanism operates at the level of the context-free grammar, it is difficult for the client to define scoped abbreviations for these verbose marking tokens.

TLMs sidestep these complexities entirely because the context-free syntax of the language is fixed, and composition is via splicing rather than direct combination. The TLM provider can therefore modularly establish that the syntax definition that the TLM implements (possibly using a syntax definition system internally, as discussed in Sec. 6) is deterministic. There is no need for verbose marking tokens, and programmers can define scoped TLM abbreviations as discussed in Sec. 5.

Even putting aside the problem of syntactic conflict, there are questions about just how reasonable sprinkling many possibly unfamiliar library-specific literal forms throughout a program may be. The intuition was given by example in Sec. 1. To reiterate, more generally:

- (1) **Responsibility:** Clients using a combined syntax dialect cannot easily determine which constituent extension is responsible for a given form, whereas TLM clients can follow the binding structure of the language in the usual manner.
- (2) **Segmentation:** Clients of a syntax dialect cannot accurately determine which segments of the program text appear directly in the desugaring. In contrast, TLM clients can inspect the inferred segmentation (communicate via, e.g., secondary notation.)
- (3) **Binding:** Clients of a context-free syntax dialect cannot be sure that the desugaring is context-independent and that spliced terms are capture avoiding, as TLM clients can.
- (4) **Typing:** Clients of a syntax dialect cannot reason abstractly about types. In contrast, TLM clients can determine the type of any expansion by referring to the parameter and type declarations on the TLM definition, and nothing else. The inferred segmentation also gives types for each spliced expression or pattern. This information can be communicated upon request by the type inspection service of a program editor.

The work of Lorenzen and Erdweg (2013, 2016) introduces SoundExt, a grammar-based syntax extension system where extension providers can equip their new forms with derived typing rules. The system then attempts to automatically verify that the expansion logic (expressed using a rewrite system, rather than an arbitrary function) is sound relative to these derived rules.

TLMs differ in several ways. First, as already discussed, we leave the context-free syntax fixed, so different TLMs cannot conflict. Second, SoundExt does not enforce hygiene, i.e. expansions might depend on the context and intentionally induce capture. Similarly, there is no abstract segmentation discipline. While the client can indirectly reason about binding (but not segmentation) by inspecting the derived typing rules, this is weaker than the strict hygiene and segmentation discipline that TLMs enforce. Another distinction is that TLMs do not support type-dependent expansions. The trade-off is that this allows expansions, and therefore segmentations, to be generated even when the program is ill-typed, as long as the static values are well-typed. Another important distinction is that TLMs rely on proto-expansion validation, rather than verification as in SoundExt. The trade-off is that TLMs do not require that the expansion logic be written using a restricted rewriting system, nor does the system require a fully mechanized language definition and heavyweight theorem proving machinery. Finally, it would be difficult to define syntax extensions in SoundExt that operate over module-parameterized families of types. Parameters can be approximated using splicing, but there is no clear notion of “partial application”.

7.2 Term Rewriting Systems

Another approach – and the approach that TLMs are rooted in – is to leave the context-free syntax of the language fixed, and instead contextually rewrite existing literal forms.

For example, OCaml’s textual syntax now includes *preprocessor extension (ppx) points* used to identify terms that some term rewriting preprocessor must rewrite (Leroy et al. 2014). We could mark a string literal containing Ur/Web-style HTML syntax with a ppx annotation, `xml`, as follows:

```
[%xml "<h1>Hello, {[first_name]}!</h1>"]
```

elaborat

There are problems reasoning modularly and abstractly about such constructions. More than one applied preprocessor might recognize this annotation (there are, in practice, many XML/HTML libraries), so the problems of **Responsibility** come up. It is also impossible to reason abstractly about **Segmentation**, **Capture**, **Context Dependence** and **Typing** because the code that the preprocessor generates is unconstrained.

Term-rewriting macro systems are language-integrated local term rewriting systems that require that the client explicitly apply the intended rewriting, implemented by a macro, to the term that is to be rewritten. This addresses the issue of **Responsibility**. However, unhygienic, untyped macro systems, like the earliest variants of the Lisp macro system (Hart 1963), Template Haskell (Sheard and Peyton Jones 2002) and GHC’s quasiquotation system (Mainland 2007) (which is based on Template Haskell), do not allow clients to reason abstractly about the remaining issues, again because the expansion that they produce is unconstrained. (It is not enough that with Template Haskell / GHC quasiquotation, the generated expansion is typechecked – to satisfy the **Typing** criterion, it must be possible to reason abstractly about *what the type of the generated expansion is*.)

Hygienic macro systems prevent, or abstractly account for (Herman 2010; Herman and Wand 2008), **Capture**, and they enforce application-site **Context Independence** (Adams 2015; Clinger and Rees 1991; Dybvig et al. 1992; Kohlbecker et al. 1986). However, this turns out to make it impossible to repurpose string literal forms to introduce compositional literal forms at other types. Consider again our running XHTML example, which we might try to realize by applying a hygienic macro, `xml!`, to a string literal that the macro parses:

```
(xml! " <h1>Hello, {[first_name]}!</h1>" )
```

The expansion of this macro will either fail the check for context independence because `first_name` will appear as a free variable, no different from any other, or else it will have incorrectly captured an expansion-local binding of the same identifier. The hygiene mechanism for TLMs addresses this problem by explicitly distinguishing spliced segments of the literal body in the proto-expansion. This also addresses the problem of **Segmentation** – the segmentation abstractly communicates the fact that (only) `first_name` is a spliced expression of type string.

Much of the research on macro systems has been for languages in the LISP tradition (McCarthy 1978) that do not have rich static type structure. The formal macro calculus studied by Herman and Wand (2008) uses types to encode the binding structure of the generated expansion (in particular, to indicate which identifiers passed as arguments can be captured by which other arguments), but the expanded language itself does not have rich type structure (nor is this calculus is not capable of expressing new literal forms, for the reasons already discussed.) Research on typed *staging macro systems* like MetaML (Sheard 1999) and MacroML (Ganz et al. 2001) is also not directly applicable to the problem of defining new literal forms – the syntax tree of the arguments cannot be inspected at all (staging macros are useful mainly for performance reasons.) The Scala macro system is a hygienic macro system which supports “black-box” term rewriting macros. These support reasoning abstractly about **Typing** because type annotations constrain the macro arguments and the generated expansions, though the precise reasoning principles available are unclear because the Scala macro system has not been formally specified. In any case, black-box macros also cannot be used to repurpose string literal forms for the reasons above.

Some languages, including Scala (Odersky et al. 2008), build in *string splicing* (a.k.a. *string interpolation*) forms, or similar but more general *fragmentary quotation forms* (Slind 1991), e.g. the SML/NJ dialect of ML. These designate a particular delimiter to escape out into the expression language. The problem with using these (dynamically, or in conjunction with term rewriting macros) as vehicles to introduce literal forms at various other types is 1) there is no “one-size-fits-all” escape delimiter, and 2) typing is problematic because every escaped term is checked against the same

type. In the XHTML example, we have splicing at two different types using two different delimiters. These forms are also not available in pattern position.

This brings us to the most closely related work, that of Omar et al. (2014) on *type-specific languages* (TSLs). Like simple expression TLMs (Sec. 2), TSLs allow library providers to programmatically control the parsing of expressions of generalized literal form. Parse functions are associated directly with nominal types and invoked according to a bidirectionally typed protocol. In contrast, TLMs are separately defined and explicitly applied. Accordingly, different TLMs can therefore be defined at the same type without conflict, and can operate at any type, not just at nominal (i.e. abstract) types. In a subsequent short paper, Omar et al. (2015) proposed explicit application of simple expression TLMs in a bidirectional typed setting (Pierce and Turner 2000), but this proposal did not give any formal details. With TLMs, it is not necessary for the language to be bidirectionally typed – context independence implies that ML-style type inference can be performed using only the segmentation (because the remainder of an expansion cannot mention the very variables whose types are being inferred.) Another critical distinction is that the metatheory presented by Omar et al. (2014) establishes only that generated expansions are well-typed. It does not establish the remaining abstract reasoning principles that have been a major focus of this paper. In particular, there is no formal notion of hygiene (though it is discussed informally), and the mechanism does not guarantee that a valid segmentation will exist, nor associate types with segments. Finally, this prior work did not consider pattern matching, type functions, ML-style modules, parameters or static evaluation. All of these are important for integration into an ML- or Scala-like language.

8 Discussion

The importance of specialized notation as a “tool for thought” has long been recognized (Iverson 1980). According to Whitehead, a good notation “relieves the brain of [...] unnecessary work” and “sets it free to concentrate on more advanced problems” (Cajori 1928). And indeed, advances in mathematics, science and programming have often been accompanied by new notation.

Of course, this desire to “relieve the brain of unnecessary work” has motivated not only the syntax but also the semantics of languages like ML and Scala – these languages maintain a strong type and binding discipline so that programmers, and their tools, can hold certain implementation details abstract when reasoning about program behavior. In the words of Reynolds (1983), “type structure is a syntactic discipline for enforcing levels of abstraction.”

Previously, these two relief mechanisms were in tension – mechanisms that allowed programmers to express new notation would obscure the type and binding structure of program text, as discussed in Sec. 1 and detailed in Sec. 7. TLMs resolve this tension for the broad class of literal forms that generalized literal forms subsume. This class includes all of the examples enumerated in Sec. 1 (up to the choice of outermost delimiter), the additional examples described in Sec. 6, the more detailed example in the supplement, and the many examples in the prior papers by Omar et al. (2014, 2015).

Of course, not all possible literal notation will prove to be in good taste, and we make no empirical claims about particular notation in this paper. TLMs leave the burden of establishing the value of any particular literal notation to individual library providers, rather than to the language designer. The reasoning principles that TLMs provide, which are the specific contributions of this paper, allow clients to “reason around” poor literal designs, using principles analagous to those already familiar to programmers in languages like these. The technical content of this paper, which formally establishes these reasoning principles, constitutes the first detailed type-theoretic account of a typed, hygienic macro system for an ML-like language, i.e. one with a rich static type system, support for pattern matching and an ML-like module system with type functions and abstract type members. Our own language integration efforts have been focused on an emerging ML dialect called

do this

VerseML, but Scala supports analagous features (Amin et al. 2014) and would also be a suitable host for the TLM mechanism. We also plan to implement TLMs into *typy*, a typed functional language embedded into Python as a library (Omar and Aldrich 2016).

Haskell’s *do*-notation at types with monadic structure cannot be expressed directly using TLMs because it needs to purposefully induce capture in spliced sub-terms, and TLMs prevent capture. This limitation could be resolved by allowing identifiers that appear in literal bodies to be explicitly designated as bound within spliced expressions in the segmentation. The prior work by Herman and Wand (2008) considered a similar approach in a setting where the locations of sub-terms is known ahead of time (building on the *tree locations* of Gorn (1965), which are conceptually similar to our spliced segment locations.) We leave as future work the details of this proposal and a consideration of whether this additional expressive power would justify the increased reasoning complexity for clients (it is presently the author’s opinion that it would not.)

A correct parse function never returns an encoding of a proto-expansion that fails validation. This invariant cannot be enforced by the simple ML-style type system we have been working with. Using a proof assistant, it would be possible to verify that a parse function generates only encodings of valid proto-expansions. Alternatively, in a dependently typed setting, the type of the parse function itself could be enriched so as to enforce this invariant intrinsically. We leave the details of this approach as future work.

Another important direction has to do with automated refactoring. The unexpanded language does not come equipped with standard notions of renaming and substitution. Given a splice summary, it should be possible to “backpatch” refactorings into literal bodies, but the details are left as future work.

At several points in the paper, we allude to editor integration. However, some important questions having to do with TLM-specific syntax highlighting, incremental parsing and error recovery remain to be considered. Another interesting direction would be to generalize TLMs to support non-textual notation in the setting of a structure editor. The semantic foundations for structure editors proposed recently by Omar et al. (2017) may guide such an effort.

Other directions for future work are given in the supplement.

REFERENCES

- Michael D. Adams. 2015. Towards the Essence of Hygiene. In *POPL*. <http://doi.acm.org/10.1145/2676726.2677013>
- Nada Amin, Tiark Rompf, and Martin Odersky. 2014. Foundations of path-dependent types. In *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications, OOPSLA 2014, part of SPLASH 2014, Portland, OR, USA, October 20-24, 2014*, Andrew P. Black and Todd D. Millstein (Eds.). ACM, 233–249. <http://dl.acm.org/citation.cfm?id=2660193>
- Eric Anderson, Gilman D Veith, and David Weininger. 1987. *SMILES, a line notation and computerized interpreter for chemical structures*. US Environmental Protection Agency, Environmental Research Laboratory.
- Alan Bawden. 1999. Quasiquotation in Lisp. In *Partial Evaluation and Semantic-Based Program Manipulation*. <http://repository.readscheme.org/ftp/papers/pepm99/bawden.pdf>
- Matthias Blume. 2002. *The SML/NJ Compilation and Library Manager*. Available from <http://www.smlnj.org/doc/CM/index.html>.
- Martin Bravenboer, Eelco Dolstra, and Eelco Visser. 2007. Preventing Injection Attacks with Syntax Embeddings. In *GPCE*. <http://doi.acm.org/10.1145/1289971.1289975>
- Florian Cajori. 1928. *A history of mathematical notations*. Vol. 1. Courier Corporation.
- Adam Chlipala. 2010. Ur: statically-typed metaprogramming with type-level record computation. In *PLDI*. <http://doi.acm.org/10.1145/1806596.1806612>
- Adam Chlipala. 2015. Ur/Web: A Simple Model for Programming the Web. In *POPL*. <http://dl.acm.org/citation.cfm?id=2676726>
- William D. Clinger and Jonathan Rees. 1991. Macros That Work. In *POPL*. <http://doi.acm.org/10.1145/99583.99607>
- Karl Cray. 2009. A syntactic account of singleton types via hereditary substitution. In *Fourth International Workshop on Logical Frameworks and Meta-Languages: Theory and Practice (LFMTP)*. <http://doi.acm.org/10.1145/1577824.1577829>

- 1 F. DeRemer and H. Kron. 1976. Programming-in-the-Large versus Programming-in-the-Small. *IEEE Transactions on Software*
- 2 *Engineering* 2 (1976), 80–86.
- 3 Derek Dreyer. 2005. *Understanding and evolving the ML module system*. Ph.D. Dissertation. Carnegie Mellon University.
- 4 R. Kent Dybvig, Robert Hieb, and Carl Bruggeman. 1992. Syntactic Abstraction in Scheme. *Lisp and Symbolic Computation*
- 5 5, 4 (1992), 295–326.
- 6 Sebastian Erdweg, Tillmann Rendel, Christian Kastner, and Klaus Ostermann. 2011. SugarJ: Library-based syntactic language
- 7 extensibility. In *OOPSLA*.
- 8 Sebastian Erdweg and Felix Rieger. 2013. A framework for extensible languages. In *GPCE*.
- 9 Matthew Flatt. 2012. Creating Languages in Racket. *Commun. ACM* 55, 1 (Jan. 2012), 48–56. [http://doi.acm.org/10.1145/](http://doi.acm.org/10.1145/2063176.2063195)
- 10 [2063176.2063195](http://doi.acm.org/10.1145/2063176.2063195)
- 11 Steven Ganz, Amr Sabry, and Walid Taha. 2001. Macros as Multi-Stage Computations: Type-Safe, Generative, Binding
- 12 Macros in MacroML. In *ICFP*.
- 13 Saul Gorn. 1965. Explicit definitions and linguistic dominoes. In *Systems and Computer Science, Proceedings of the Conference*
- 14 *held at Univ. of Western Ontario*. 77–115.
- 15 T. R. G. Green. 1989. Cognitive Dimensions of Notations. In *Proceedings of the HCI'89 Conference on People and Computers V*
- 16 *(Cognitive Ergonomics)*. 443–460.
- 17 Robert Harper. 1997. Programming in Standard ML. <http://www.cs.cmu.edu/~rwh/smlbook/book.pdf>. (1997). Working
- 18 draft, retrieved June 21, 2015.
- 19 Robert Harper. 2012. *Practical Foundations for Programming Languages*. Cambridge University Press.
- 20 Robert Harper. 2016. *Practical Foundations for Programming Languages* (2nd ed.). Cambridge University Press. [https://](https://www.cs.cmu.edu/~rwh/plbook/2nded.pdf)
- 21 www.cs.cmu.edu/~rwh/plbook/2nded.pdf
- 22 Robert Harper, John C Mitchell, and Eugenio Moggi. 1989. Higher-order modules and the phase distinction. In *POPL*.
- 23 T. P. Hart. 1963. *MACRO Definitions for LISP*. Report A. I. MEMO 57. Massachusetts Institute of Technology, A.I. Lab.,
- 24 Cambridge, Massachusetts.
- 25 David Herman. 2010. *A Theory of Typed Hygienic Macros*. Ph.D. Dissertation. Northeastern University, Boston, MA.
- 26 David Herman and Mitchell Wand. 2008. A Theory of Hygienic Macros. In *ESOP*. [http://dx.doi.org/10.1007/978-3-540-78739-6_](http://dx.doi.org/10.1007/978-3-540-78739-6_4)
- 27 [4](http://dx.doi.org/10.1007/978-3-540-78739-6_4)
- 28 Graham Hutton. 1992. Higher-order functions for parsing. *Journal of Functional Programming* 2, 3 (July 1992), 323–343.
- 29 <http://www.cs.nott.ac.uk/Department/Staff/gmh/parsing.ps>
- 30 Kenneth E. Iverson. 1980. Notation as a Tool of Thought. *Commun. ACM* 23, 8 (1980), 444–465. DOI : [http://dx.doi.org/10.](http://dx.doi.org/10.1145/358896.358899)
- 31 [1145/358896.358899](http://dx.doi.org/10.1145/358896.358899)
- 32 Eugene E. Kohlbecker, Daniel P. Friedman, Matthias Felleisen, and Bruce Duba. 1986. Hygienic Macro Expansion. In
- 33 *Symposium on LISP and Functional Programming*. 151–161.
- 34 Ralf Lämmel, Ekaterina Pek, and Jürgen Starek. 2011. Large-scale, AST-based API-usage analysis of open-source Java
- 35 projects. In *ACM Symposium on Applied Computing (SAC)*. <http://doi.acm.org/10.1145/1982185.1982471>
- 36 Daniel K. Lee, Karl Crary, and Robert Harper. 2007. Towards a mechanized metatheory of Standard ML. In *POPL*. [http://](http://dl.acm.org/citation.cfm?id=1190216)
- 37 dl.acm.org/citation.cfm?id=1190216
- 38 Xavier Leroy, Damien Doligez, Alain Frisch, Jacques Garrigue, Didier Rémy, and Jérôme Vouillon. 2014. *The OCaml system*
- 39 *release 4.02 Documentation and user's manual*. Institut National de Recherche en Informatique et en Automatique.
- 40 Barbara Liskov and Stephen Zilles. 1974. Programming with abstract data types. In *ACM SIGPLAN Notices*, Vol. 9. ACM,
- 41 50–59.
- 42 Florian Lorenzen and Sebastian Erdweg. 2013. Modular and automated type-soundness verification for language extensions.
- 43 In *ICFP*. 331–342. <http://dl.acm.org/citation.cfm?id=2500365>
- 44 Florian Lorenzen and Sebastian Erdweg. 2016. Sound type-dependent syntactic language extension. In *POPL*. [http://](http://dl.acm.org/citation.cfm?id=2837614)
- 45 dl.acm.org/citation.cfm?id=2837614
- 46 David MacQueen. 1984. Modules for Standard ML. In *Symposium on LISP and Functional Programming*. [http://doi.acm.org/](http://doi.acm.org/10.1145/800055.802036)
- 47 [10.1145/800055.802036](http://doi.acm.org/10.1145/800055.802036)
- 48 David B. MacQueen. 1986. Using Dependent Types to Express Modular Structure. In *Conference Record of the Thirteenth*
- 49 *Annual ACM Symposium on Principles of Programming Languages, St. Petersburg Beach, Florida, USA, January 1986*.
- 50 277–286. DOI : <http://dx.doi.org/10.1145/512644.512670>
- 51 Geoffrey Mainland. 2007. Why it's nice to be quoted: quasiquoting for Haskell. In *Haskell Workshop*.
- 52 J. McCarthy. 1978. History of LISP. In *History of programming languages I*. ACM, 173–185.
- 53 Robin Milner, Mads Tofte, Robert Harper, and David MacQueen. 1997. *The Definition of Standard ML (Revised)*. The MIT
- 54 Press.
- 55 Martin Odersky, Lex Spoon, and Bill Venners. 2008. *Programming in Scala*. Artima Inc.
- 56 Cyrus Omar and Jonathan Aldrich. 2016. Programmable Semantic Fragments. In *GPCE*.

- 1 Cyrus Omar, Darya Kurilova, Ligia Nistor, Benjamin Chung, Alex Potanin, and Jonathan Aldrich. 2014. Safely Composable
2 Type-Specific Languages. In *ECOOP*.
- 3 Cyrus Omar, Ian Voysey, Michael Hilton, Jonathan Aldrich, and Matthew A. Hammer. 2017. Hazelnut: a bidirectionally
4 typed structure editor calculus. In *POPL*. <http://dl.acm.org/citation.cfm?id=3009900>
- 5 Cyrus Omar, Chenglong Wang, and Jonathan Aldrich. 2015. Composable and Hygienic Typed Syntax Macros. In *ACM
6 Symposium on Applied Computing (SAC)*.
- 7 OWASP. 2017. OWASP Top 10 2013. https://www.owasp.org/index.php/Top_10_2017-Top_10. Retrieved May 28, 2017.
(2017).
- 8 Benjamin C. Pierce and David N. Turner. 2000. Local Type Inference. *ACM Trans. Program. Lang. Syst.* 22, 1 (Jan. 2000), 1–44.
9 <http://doi.acm.org/10.1145/345099.345100>
- 10 Gordon D. Plotkin. 2004. A structural approach to operational semantics. *J. Log. Algebr. Program.* 60–61 (2004), 17–139.
- 11 J C Reynolds. 1983. Types, abstraction and parametric polymorphism. In *IFIP Congress*.
- 12 August Schwerdfeger and Eric Van Wyk. 2009. Verifiable composition of deterministic grammars. In *PLDI*. <http://doi.acm.org/10.1145/1542476.1542499>
- 13 Dana Scott. 1980. Lambda calculus: some models, some philosophy. *Studies in Logic and the Foundations of Mathematics* 101
(1980), 223–265.
- 14 Denys Shabalin, Eugene Burmako, and Martin Odersky. 2013. *Quasiquotes for Scala*. Technical Report EPFL-REPORT-185242.
- 15 Tim Sheard. 1999. Using MetaML: A Staged Programming Language. *Lecture Notes in Computer Science* 1608 (1999).
- 16 Tim Sheard and Simon Peyton Jones. 2002. Template meta-programming for Haskell. In *Haskell Workshop*.
- 17 Konrad Slind. 1991. Object language embedding in Standard ML of New-Jersey. In *ICFP*.
- 18 Christopher A Stone and Robert Harper. 2006. Extensional equivalence and singleton types. *ACM Transactions on
19 Computational Logic (TOCL)* 7, 4 (2006), 676–722.
- 20 Walid Taha. 1999. *Multi-Stage Programming: Its Theory and Applications*. Ph.D. Dissertation. Oregon Graduate Institute of
Science and Technology.
- 21 Arie van Deursen, Paul Klint, and Frank Tip. 1993. Origin Tracking. *J. Symb. Comput.* (1993), 523–545. [http://dx.doi.org/10.
22 1016/S0747-7171\(06\)80004-0](http://dx.doi.org/10.1016/S0747-7171(06)80004-0)
- 23 Eric Van Wyk and August Schwerdfeger. 2007. Context-aware scanning for parsing extensible languages. In *GPCE*.
24 <http://doi.acm.org/10.1145/1289971.1289983>
- 25 Arthur Whitney and Dennis Shasha. 2001. Lots O’Ticks: Real Time High Performance Time Series Queries on Billions of
Trades and Quotes. *SIGMOD Rec.* 30, 2 (May 2001), 617–617. DOI : <http://dx.doi.org/10.1145/376284.375783>
- 26 Jeremy Yallop and Leo White. 2015. Modular macros (extended abstract). In *OCaml Users and Developers Workshop*.