FUNCTIONAL PEARL

Typed quote/antiquote or:
Compile-time parsing

RALF HINZE

Department of Computer Science, University of Oxford, Wolfson Building, Parks Road,
Oxford OX1 3QD, England, UK
(e-mail: ralf.hinze@cs.ox.ac.uk)

1 Introduction

Haskell (Peyton Jones, 2003) is often used as a host language for embedding other
languages. Typically, the abstract syntax of the guest language is defined by a
collection of datatype declarations; parsers and pretty-printers convert between
the concrete syntax and its abstract representation. A quote/antiquote mechanism
permits a tighter integration of the guest language into the host language by allowing
one to use phrases in the guest language's concrete syntax.

For a simple example, assume that the abstract syntax of the guest language is
given by the following datatype of binary trees.

```
data Tree = Leaf | Fork Tree Tree
```

To dispense with the need for parentheses, we choose prefix notation for the concrete
syntax. The following interactive session illustrates the use of quotations.

```
Main> «fork fork leaf leaf leaf»
Fork (Fork Leaf Leaf) Leaf
Main> size («fork fork leaf leaf leaf») + 1
4
```

A quotation is delimited by guillemets (« and ») and consists of a phrase in concrete
syntax, in our case, a prefix expression. The concrete syntax is a sequence of terminal
symbols, written in typewriter font. A quotation evaluates to abstract syntax and
can be freely mixed with ordinary Haskell expressions. In our example, a quotation
yields a value of type Tree and may therefore serve as an argument to size, which
computes the size of a tree.

Perhaps surprisingly, our quote mechanism guarantees that the guest-language
phrase is well formed: the malformed quotations «fork» and «leaf leaf» are
both rejected by Haskell’s type-checker. This is a big advantage over the use of
strings, which often serve as an approximation to quotations.

The relationship between host and guest language also suggests a notion of
antiquotation: the ability to splice a host-language expression into the middle of
a guest-language phrase. Continuing the example above, here is a session that
demonstrates the use of antiquotations:

\[
\begin{align*}
&\text{Main} \triangleright \langle \text{fork \{full 2\} leaf} \rangle \\
&\text{Fork (Fork (Fork Leaf Leaf) (Fork Leaf Leaf)) Leaf} \\
&\text{Main} \triangleright \text{let foo } t = \langle \text{fork } t \text{ leaf} \rangle \\
&\text{Main} \triangleright \text{foo } (\langle \text{fork leaf fork leaf leaf} \rangle)
\end{align*}
\]

An antiquotation is written as a back-quote (') followed by an atomic Haskell expression, for instance, an identifier or a parenthesised expression. The Haskell expression typically generates a piece of abstract syntax, for instance, in the first expression above, a fully balanced binary tree of depth 2.

A quote/antiquote mechanism usually requires an extension of the host language. The purpose of this pearl is to show that one can program such a mechanism within Haskell itself. The technique is based on Okasaki’s flattening combinators (Okasaki, 2002, 2003), which we shall review in the next section. To make the idea fly, I assume that we can use an arbitrary terminal symbol in typewriter font as a Haskell identifier. If you think that this assumption undermines the argument, then you should read the pearl as an exercise in compile-time parsing.

2 Background: the other continuation-passing style monad

To illustrate the basic idea consider a very simple example, which implements concrete syntax for the naturals.

\[
\begin{align*}
&\text{Main} \triangleright (\langle ||| \rangle, \langle ||||| \rangle + 7) \\
&(3, 12)
\end{align*}
\]

We have only one terminal symbol, the vertical bar, where a sequence of \(n\) bars represents the number \(n\).

The succession of symbols \(\langle ||| \rangle\) looks like a sequence of terminals enclosed in guillemets. But, of course, this is an illusion; the sequence is, in fact, a nested application of functions. If we take “\(\langle\)”, “\(\rangle\)” and “\(\mid\)” as aliases for \texttt{quote}, \texttt{endquote}, and \texttt{tick}, then \(\langle ||| \rangle\) abbreviates the fully parenthesised expression \(((\text{quote tick} ) \text{ tick} ) \text{ tick} \text{ endquote}\). In what follows, we shall use “\(\langle\)” and \texttt{quote}, “\(\rangle\)” and \texttt{endquote}, “\(\mid\)” and \texttt{antiquote} interchangeably.

Now, if Haskell used postfix function application, then we could simply define \texttt{quote} = 0, \texttt{tick} = \texttt{succ}, \texttt{endquote} = \texttt{id} and we would be done. For Haskell’s prefix function application, we must additionally arrange that functions and arguments are swapped:

\[
\begin{align*}
&\texttt{quote} \quad f = f 0 \\
&\texttt{tick} \quad n f = f (\texttt{succ} n) \\
&\texttt{endquote} n = n
\end{align*}
\]

The stepwise evaluation of \(\langle ||| \rangle\) shows that \texttt{tick} increments the counter, initialised to 0 by \texttt{quote}, and then passes control to the next function, which is either
another tick or endquote.

\[
\text{quote tick tick tick endquote} = \text{tick 0 tick tick endquote} = \text{tick 1 tick endquote} = \text{tick 2 endquote} = \text{endquote 3} = 3
\]

The evaluation is solely driven by the terminal symbols, which is why we call them active terminals. This technique of passing control to a function argument is reminiscent of continuation-passing style (CPS). And indeed, if we call the CPS Monad to mind\(^1\)

\[
\text{type CPS } \alpha = \forall \text{ans} . (\alpha \to \text{ans}) \to \text{ans}
\]

\[
\text{instance Monad CPS where}
\]

\[
\text{return a } = \lambda \kappa \to \kappa \ a
\]

\[
\text{m } \gg\gg \ k = \lambda \kappa \to m (\lambda a \to k \ a \ \kappa)
\]

we can identify quote as return 0 and tick as lift succ where lift turns a pure function into a monadic one:

\[
\text{type } \alpha \to \beta = \alpha \to CPS \ \beta
\]

\[
\text{lift } :: (\alpha \to \beta) \to (\alpha \to \beta)
\]

\[
\text{lift } f \ a = \text{return } (f \ a)
\]

However, the bind of the monad, “\(\gg\gg\)”, seems unrelated: in the CPS monad the continuation represents the “rest of the computation” whereas in our example the continuation only stands for the next parsing step.

It may come as a surprise that the instance declaration above is not the only possibility for turning CPS into a monad. Here is a second instance introducing the monad of partial continuations.

\[
\text{instance Monad CPS where}
\]

\[
\text{return a } = \lambda \kappa \to \kappa \ a
\]

\[
\text{m } \gg\gg \ k = m \ k
\]

The definition of return is unchanged; “\(\gg\gg\)” is now a type-restricted instance of function application. Actually, it is a combination of type application—the universally quantified variable ans in the type of m is instantiated to CPS \(\beta\)—and function application, but this is not visible in Haskell. Since “\(\gg\gg\)” is postfix application of “effectful” functions, this CPS monad implements postfix function application! Consequently, the quotation « || || » can be seen as a monadic computation in disguise:

\[
(((\text{quote tick) tick) tick}) \text{ endquote} = \text{run (quote } \gg\gg \text{ tick } \gg\gg \text{ tick } \gg\gg \text{ tick})
\]

\(^1\) The instance declaration, which is not legal Haskell, serves only illustrative purposes. We shall only need return and only at this particular type.
where \( \text{run} \) encapsulates a CPS computation:

\[
\text{run} :: \text{CPS } \alpha \rightarrow \alpha
\]

\[\text{run } m = m \text{id} \]

Generalising from the example, quotations are based on the identity

\[
\text{quote } \text{act}_1 \ldots \text{act}_n \text{ endquote} = \text{run} (\text{quote} \gg= \text{act}_1 \gg= \cdots \gg= \text{act}_n)
\]

where \( \text{quote} :: \text{CPS } \tau_1, \text{act}_i :: \tau_i \rightarrow \tau_{i+1}, \) and \( \text{endquote} = \text{id}. \) It is useful to think of the \( \tau_i \) as state types and the \( \text{act}_i \) as transitions: \( \text{quote} \) initialises the state; each active terminal \( \text{act}_i \) transforms the state. In our example, we had a single state type but this need not be the case in general. In fact, choosing precise state types is the key to “typed quotes/antiquotes.”

Just in case you were wondering, none of the two CPS monads is a very exciting one in terms of expressiveness: they are both isomorphic to the identity monad with \( \text{return} \) and \( \text{run} \) converting between them. In other words, CPS offers no effects.

Without loss of generality, we may therefore assume that \( \text{quote} \) and \( \text{act}_i \) are liftings: \( \text{quote} = \text{return } a \) and \( \text{act}_i = \text{lift } f_i \) for some \( a \) and suitable \( f_i \). The following calculation summarises our findings:

\[
\begin{align*}
\text{quote } \text{act}_1 \ldots \text{act}_n \text{ endquote} &= \{ \text{definition of } \gg= \text{ and } \text{run, and } \text{endquote} = \text{id } \} \\
\text{run} (\text{quote} \gg= \text{act}_1 \gg= \cdots \gg= \text{act}_n) &= \{ \text{CPS is a pure monad: } \text{quote} = \text{return } a \text{ and } \text{act}_i = \text{lift } f_i \} \\
\text{run} (\text{return } a \gg= \text{lift } f_1 \gg= \cdots \gg= \text{lift } f_n) &= \{ \text{monad laws } \} \\
\text{run} (\text{return} (f_n (\cdots (f_1 a) \cdots))) &= \{ \text{run } \cdot \text{return } = \text{id } \} \\
f_n (\cdots (f_1 a) \cdots) &=
\end{align*}
\]

In the toy example and in the formal development above, \( \text{endquote} \) was always the identity. This is, however, not quite adequate, as the desired value of a quotation is not necessarily identical to the last state. Fortunately, \( \text{endquote} \) can be any function since we can fuse a post-processing step with the final continuation: \( \text{post} (\text{run } m) = m \text{ post}. \) This is an immediate consequence of the free theorem for the type \( \text{CPS } \alpha \) (Wadler, 1989).

To summarise, a quotation of type \( \tau \) is of the form

\[
\text{quote } \text{act}_1 \ldots \text{act}_n \text{ endquote}
\]

where \( \text{quote} :: \text{CPS } \tau_1, \text{act}_i :: \tau_i \rightarrow \tau_{i+1} \) and \( \text{endquote} :: \tau_{n+1} \rightarrow \tau. \)

Since the evaluation of a quotation is driven by the terminal symbols, the implementation of a quote/antiquote mechanism for a particular guest language goes hand in hand with the development of a parser for the concrete syntax. The
following sections are ordered by the underlying parser’s level of sophistication: Section 3 shows how to implement simple postfix and prefix parsers, Section 4 deals with predictive top-down parsers, and finally, Section 5 introduces quotations that are based on bottom-up parsers.

3 Parsing datatypes

Continuing the example from the introduction, we show how to parse elements of datatypes in postfix and in prefix notation. Section 3.1 is an excerpt of Okasaki’s extensive treatment of postfix languages (see Okasaki, 2002).

3.1 Postfix notation

In postfix notation, also known as reverse Polish notation, functions follow their arguments. Postfix notation dispenses with the need for parentheses, if the arity of functions is statically known. This is generally not the case in higher order typed languages, but it is true of data constructors (ignoring the fact that they are curried in Haskell).

Evaluation of postfix expressions is naturally stack-based: a function pops its arguments from the stack and pushes the result back onto it. To parse datatypes in postfix notation, we introduce for each data constructor \( C :: \tau_1 \to \cdots \to \tau_n \to \tau \) a postfix constructor:

\[
C :: \tau_1 \to \cdots \to \tau_n \\
C ((\tau_1, \ldots, \tau_n)) = (\tau, C t_1 \cdots t_n)
\]

The stack, represented by a nested pair, grows from left to right. The modification of the stack is precisely captured in the type: \( c \) is only applicable if the stack contains at least \( n \) arguments and the topmost \( n \) have the correct types. For the \( \text{Tree} \) type, this specialises to

\[
\begin{align*}
\text{leaf} & :: \tau \to (\tau, \text{Tree}) \\
\text{leaf }\, \tau & = (\tau, \text{Leaf}) \\
\text{fork} & :: ((\tau, \text{Tree}), \text{Tree}) \to (\tau, \text{Tree}) \\
\text{fork }\, ((\tau, l), r) & = (\tau, \text{Fork }\, l \, r)
\end{align*}
\]

Given these prerequisites, we can instantiate the framework of Section 2.

\[
\begin{align*}
\text{quote} & :: CPS () \\
\text{quote} & = \text{return }() \\
\text{leaf} & :: \tau \to (\tau, \text{Tree}) \\
\text{leaf} & = \text{lift }\text{leaf} \\
\text{fork} & :: ((\tau, \text{Tree}), \text{Tree}) \to (\tau, \text{Tree}) \\
\text{fork} & = \text{lift }\text{fork} \\
\text{endquote} & :: ((), \text{Tree}) \to \text{Tree} \\
\text{endquote }(((), t) & = t
\end{align*}
\]
The function \textit{quote} initialises the state to the empty stack; \textit{endquote} extracts the quoted tree from a singleton stack.

It is instructive to step through the static and dynamic elaboration of a quotation. Type checking statically guarantees that a quotation constitutes a well-formed postfix expression.

\begin{verbatim}
quote :: CPS ()
quote leaf :: CPS (((), Tree)
quote leaf leaf :: CPS ((((), Tree), Tree)
quote leaf leaf fork :: CPS (((), Tree)
quote leaf leaf fork leaf :: CPS ((((), Tree), Tree)
quote leaf leaf fork leaf fork :: CPS (((), Tree)
quote leaf leaf fork leaf fork endquote :: Tree
\end{verbatim}

In each step, the state type precisely mirrors the stack layout. Consequently, pushing too few or too many or the wrong types of arguments results in a static type-error. The dynamic evaluation shows how the state evolves.

\begin{verbatim}
quote leaf leaf fork leaf fork endquote
= leaf () leaf fork leaf fork endquote
= leaf (((), Leaf)) fork leaf fork endquote
= fork ((((), Leaf), Leaf)) leaf fork endquote
= leaf (((), Fork Leaf Leaf)) fork endquote
= fork ((((), Fork Leaf Leaf), Leaf)) endquote
= endquote (((), Fork (Fork Leaf Leaf) Leaf)
= Fork (Fork Leaf Leaf) Leaf
\end{verbatim}

The state is always passed as the first argument. This is something to bear in mind when implementing additional functionality, such as an antiquote mechanism.

\begin{verbatim}
antiquote :: st → Tree → (st, Tree)
antiquote st t = return (st,t)
\end{verbatim}

The tree is spliced into the current position simply by pushing it onto the stack.

\section*{3.2 Prefix notation}

Postfix notation was easy; its dual, prefix notation, is slightly harder. Prefix notation was invented in 1920 by Jan Łukasiewicz, a Polish logician, mathematician, and philosopher. Because of its origin, prefix notation is also known as Polish notation.

In postfix notation, a function follows its arguments; so a stack of arguments is a natural choice for the state. In prefix notation, a function precedes its arguments. Consequently, the state becomes a stack of \textit{pending} arguments. For each data constructor $C :: \tau_1 \to \cdots \to \tau_n \to \tau$, we introduce a \textit{prefix constructor}:

\begin{verbatim}
c_0 :: ((\tau, st) → a) → ((\tau_1, \ldots, (\tau_n, st))) → a
c_0 ctx = \lambda(t_1, \ldots, (t_n, st)) → ctx (C t_1 \ldots t_n, st)
\end{verbatim}
The stack, again represented by a nested pair, now grows from right to left. The first argument of $c^o$ can be seen as a request for a value of type $\tau$ (a request is also known as a context or as an expression with a hole). The prefix constructor can satisfy this request but, in turn, generates requests for its arguments. For the $\text{Tree}$ type, we obtain

\[
\begin{align*}
\text{leaf}^o & :: ((\text{Tree}, \text{st}) \to \alpha) \to (\text{st} \to \alpha) \\
\text{leaf}^o \text{ ctx} & = \lambda \text{st} \to \text{ctx} (\text{Leaf}, \text{st}) \\
\text{fork}^o & :: ((\text{Tree}, \text{st}) \to \alpha) \to ((\text{Tree}, (\text{Tree}, \text{st})) \to \alpha) \\
\text{fork}^o \text{ ctx} & = \lambda (t, (u, \text{st})) \to \text{ctx} (\text{Fork t u}, \text{st})
\end{align*}
\]

The implementation of quotations and antiquotations is again a straightforward application of the general framework:

\[
\begin{align*}
\text{quote} & :: \text{CPS} ((\text{Tree}, (\text{})) \to \text{Tree}) \\
\text{quote} & = \text{return } (\lambda (t, (\text{})) \to t) \\
\text{leaf} & :: ((\text{Tree}, \text{st}) \to \alpha) \to (\text{st} \to \alpha) \\
\text{leaf} & = \text{lift leaf}^o \\
\text{fork} & :: ((\text{Tree}, \text{st}) \to \alpha) \to ((\text{Tree}, (\text{Tree}, \text{st})) \to \alpha) \\
\text{fork} & = \text{lift fork}^o \\
\text{endquote} & :: ((\text{}) \to \text{Tree}) \to \text{Tree} \\
\text{endquote ctx} & = \text{ctx } () \\
\text{antiquote} & :: ((\text{Tree}, \text{st}) \to \alpha) \to \text{Tree} \to (\text{st} \to \alpha) \\
\text{antiquote ctx t} & = \text{return } (\lambda \text{st} \to \text{ctx} (t, \text{st}))
\end{align*}
\]

The stack is initialised to one pending argument; we are done if there are no pending arguments left. Let us again step through an example.

\[
\begin{align*}
\text{quote} & :: \text{CPS} ((\text{Tree}, (\text{})) \to \text{Tree}) \\
\text{quote fork} & :: \text{CPS} ((\text{Tree}, (\text{Tree}, (\text{})))) \to \text{Tree} \\
\text{quote fork fork} & :: \text{CPS} ((\text{Tree}, (\text{Tree}, (\text{Tree}, (\text{}))))) \to \text{Tree} \\
\text{quote fork fork leaf} & :: \text{CPS} ((\text{Tree}, (\text{Tree}, (\text{))})) \to \text{Tree} \\
\text{quote fork fork leaf leaf} & :: \text{CPS} (((\text{Tree}, (\text{))})) \to \text{Tree} \\
\text{quote fork fork leaf leaf leaf} & :: \text{CPS} ((\text{})) \to \text{Tree} \\
\text{quote fork fork leaf leaf leaf endquote} & :: \text{CPS} \text{ Tree}
\end{align*}
\]

The types show how the stack of pending arguments grows and shrinks. For instance, when the first two forks have been processed, three further subtrees are required: the left and the right subtree of the second fork and the right subtree of the first fork. The stepwise evaluation makes this explicit:

\[
\begin{align*}
\text{quote} & :: \text{CPS} ((\text{Tree}, (\text{})) \to \text{Tree}) \\
\text{quote fork} & :: \text{CPS} ((\text{Tree}, (\text{Tree}, (\text{})))) \to \text{Tree} \\
\text{quote fork fork} & :: \text{CPS} ((\text{Tree}, (\text{Tree}, (\text{Tree}, (\text{}))))) \to \text{Tree} \\
\text{quote fork fork leaf} & :: \text{CPS} ((\text{Tree}, (\text{Tree}, (\text{))})) \to \text{Tree} \\
\text{quote fork fork leaf leaf} & :: \text{CPS} (((\text{Tree}, (\text{))})) \to \text{Tree} \\
\text{quote fork fork leaf leaf leaf} & :: \text{CPS} ((\text{})) \to \text{Tree} \\
\text{quote fork fork leaf leaf leaf endquote} & :: \text{CPS} \text{ Tree}
\end{align*}
\]

\[\text{The type variable } \alpha \text{ that appears in the type signature of } c^o \text{ corresponds to the type of the entire quotation and can be safely instantiated to } \text{Tree}. \text{ The polymorphic type is only vital if } c^o \text{ is used in quotations of different types.}\]
Again, if we pass too few or too many or the wrong types of arguments, then we get a static type-error.

Remark 1
The deeply nested pairs can be avoided if we curry the prefix constructors:

\[ c^\circ : (\tau \to \alpha) \to (\tau_1 \to \cdots \to \tau_n \to \alpha) \]

\[ c^\circ \text{ctx} = \lambda t_1 \to \cdots \to \lambda t_n \to \text{ctx} (C t_1 \ldots t_n) \]

Additionally, we have generalised the result type of requests from \( st \to \alpha \) to \( \alpha \). The adaptation of the remaining code is left as an exercise to the reader.

\[ \square \]

4 Top-down parsing

The main reason for treating prefix parsers is that they pave the way for the more expressive class of LL(1) parsers. The basic setup remains unchanged; we need only one additional programming technique. To keep the learning curve smooth, however, we shall go through one intermediate step and treat grammars in Greibach normal form (GNF) first.

4.1 Greibach normal form

A context-free grammar is in GNF iff all productions are of the form \( A \to a\omega \), where \( a \) is a terminal symbol and \( \omega \) is a possibly empty sequence of nonterminal symbols. A grammar in GNF is (syntactically) unambiguous iff each pair of productions \( A_1 \to a\omega_1 \) and \( A_2 \to a\omega_2 \) satisfies \( A_1 = A_2 \implies \omega_1 = \omega_2 \). Unambiguous grammars in GNF generalise datatype declarations, as a terminal (data constructor) may appear in different productions (datatype declarations).

Here is an example grammar for a simple imperative language and an equivalent grammar in GNF.

\[
\begin{align*}
S & \to \text{id} := E \\
& | \text{if } E S S \\
& | \text{while } E S \\
& | \text{begin } B \text{ end} \\
E & \to \text{id} \\
B & \to S \mid S ; B \\
C & \to := \\
E & \to \text{id} \\
R & \to \text{end} \mid ; S R
\end{align*}
\]
Remark 2
If we add the production $S \rightarrow \text{if } E S$, then the grammar becomes ambiguous, an instance of the (in-) famous dangling-else problem.

The abstract syntax of the imperative language is given by

\[
\text{type } \text{Var} = \text{String} \\
\text{data } \text{Stat} = \text{Set } \text{Var} \text{ Var} \mid \text{If } \text{Var} \text{ Stat } \text{Stat} \mid \text{While } \text{Var} \text{ Stat} \mid \text{Begin } [\text{Stat}]
\]

As an example, the quotation

\[
\text{« begin} \\
\text{\hspace{0.5em} x := y ;} \\
\text{\hspace{0.5em} if } x \\
\text{\hspace{1.5em} y := z} \\
\text{\hspace{1.5em} z := y} \\
\text{\hspace{0.5em} end »}
\]

evaluates to $\text{Begin } [\text{Set } "x" \ "y", \text{If } "x" (\text{Set } "y" \ "z") (\text{Set } "z" \ "y")]$. The definition of the variables $x$, $y$, and $z$ poses a minor problem, which we shall discuss later.

A parser for a grammar in GNF is very similar to a prefix parser: the state is a stack of pending nonterminal symbols. An active terminal selects a production by looking at the topmost symbol on the stack. If the grammar is unambiguous, then there is at most one suitable production. The nonterminal is then replaced by the right-hand side of the production (omitting the leading terminal).

As before, we want to guarantee statically that a quotation is well formed, so that its parse does not fail. To this end, we represent nonterminals by types:

\[
\begin{align*}
\text{newtype } S \alpha &= S (\text{Stat} \rightarrow \alpha) \\
\text{newtype } C \alpha &= C (\alpha) \\
\text{newtype } E \alpha &= E (\text{Var} \rightarrow \alpha) \\
\text{newtype } R \alpha &= R ([\text{Stat}] \rightarrow \alpha)
\end{align*}
\]

The declarations also list the types of the semantic values that are attached to the nonterminals. Using these type-level nonterminals, we can program the type-checker to parse quotations: each production $A \rightarrow aB_1 \ldots B_n$ is mapped to a function $a$, the active terminal, of type $A \alpha \rightarrow B_1 (\cdots (B_n \alpha) \cdots)$, that implements the expansion of $A$. There is one hitch, however: the terminal $a$ may appear in different productions, so it cannot possibly translate to a single function. Rather, an active terminal stands for a family of functions, represented in Haskell by a multiple-parameter type-class.

We introduce one class for each terminal symbol that appears more than once. In our case, the only such terminal is $\text{id}$, so we need just one class.

\[
\text{class } \text{Id } lhs \rightarrow rhs \mid lhs \rightarrow rhs \text{ where} \\
\text{id :: } \text{String} \rightarrow (lhs \rightarrow rhs)
\]

The functional dependency $lhs \rightarrow rhs$ avoids ambiguities during type-inference making use of the fact that the underlying grammar in GNF is unambiguous. Each production is translated into an equation.
instance Id (S z) (C (E z)) where
  id l = lift (λ(S ctx) → C (E (λr → ctx (Set l r))))
if   = lift (λ(S ctx) → E (λc → S (λt → S (λe → ctx (If c t e))))))
while = lift (λ(S ctx) → E (λc → S (λs → ctx (While c s))))
begin = lift (λ(S ctx) → S (λs → R (λr → ctx (Begin s : r))))
:=   = lift (λ(C ctx) → ctx)

instance Id (E z) a where
  id i = lift (λ(E ctx) → ctx i)
end   = lift (λ(R ctx) → ctx [])
;     = lift (λ(R ctx) → S (λs → R (λr → ctx (s : r))))
quote  = return (S (λs → s))
endquote s = s

The quote function pushes the start symbol on the stack; endquote simply returns the final value of type Stat.

The terminal symbol id is a bit unusual in that it takes an additional argument, the string representation of the identifier. This has the unfortunate consequence that identifiers must be enclosed in parentheses as in « (id "x") := (id "y") ». We can mitigate the unwelcome effect by introducing shortcuts

\[
\begin{align*}
x, y :: (Id lhs rhs) & \Rightarrow lhs \rightarrow rhs \\
x &= \text{id "x"} \\
y &= \text{id "y"}
\end{align*}
\]

so that the quotation becomes « x := y ». Alternatively, we may swap the two arguments of id. In this case, the parentheses can, in fact, must be dropped, so that the quotation is written « id "x" := id "y" ».

It is instructive to walk through a derivation.

\[
\begin{align*}
\text{« while } x \ y \ := \ z \text{ »} \\
&= \text{while } (S (\lambda s \rightarrow s)) \ x \ y \ := \ z \text{ »} \\
&= x (E (\lambda c \rightarrow S (\lambda s \rightarrow While c s))) \ y \ := \ z \text{ »} \\
&= y (S (\lambda s \rightarrow While "x" s)) \ := \ z \text{ »} \\
&= :\ = (C (E (\lambda r \rightarrow While "x" (Set "y" r)))) \ z \text{ »} \\
&= z (E (\lambda r \rightarrow While "x" (Set "y" r))) \text{ »} \\
&= \text{» (While "x" (Set "y" "z"))} \\
&= While "x" (Set "y" "z")
\end{align*}
\]

To summarise, for each nonterminal symbol A, we define a type: newtype A α = A (Val → α), where Val is the type of the semantic values attached to A. For each terminal symbol a, we introduce a class: class a lhs rhs | lhs → rhs where a :: lhs → rhs. Finally, each production A → aB₁ ... Bₙ gives rise to an instance declaration:

instance a (A α) (B₁ (⋯ (Bₙ α) ⋯)) where
  a = lift (λ(A ctx) → B₁ (λv₁ → ⋯ → Bₙ (λvₙ → ctx (f v₁ ... vₙ))))

where f is the semantic action associated with the production. Of course, if a terminal appears only once, then there is no need for overloading.
4.2 LL(1) parsing

We are well prepared by now to tackle the first major challenge: implementing quotations whose syntax is given by an LL(1) grammar. As before, we shall work through a manageable example. This time, we implement arithmetic expressions given by the grammar on the left below (see Aho et al. 2006, p. 193).

\[
\begin{align*}
E & \rightarrow E + T \mid T \\
T & \rightarrow T \ast F \mid F \\
F & \rightarrow ( E ) \mid id \\
\end{align*}
\]

The expression grammar is not LL(1) due to the left recursion; eliminating the left recursion yields the equivalent LL(1) grammar on the right.

The abstract syntax of arithmetic expressions is given by

\[
\text{data} \ Expr = \text{Id String} \mid \text{Add Expr Expr} \mid \text{Mul Expr Expr}
\]

The semantic actions that construct values of type \(Expr\) are straightforward to define for the original expression grammar. They are slightly more involved for the LL(1) grammar: \(E'\) and \(T'\) yield expressions with a hole, where the hole stands for the missing left argument of the operator (see Figures 1 and 2).

The main ingredient of a predictive top-down parser is the parsing table. Here is the table for the LL(1) grammar above (see Aho et al. 2006, p. 225).
The table also includes a column for " », which serves as an end marker.

The state is now a stack of pending symbols, both terminal and nonterminal. The symbol X on top of the stack determines the action of the current active terminal a. If X = a, then the terminal pops X from the stack and passes control to the next active terminal (pop action). If X is a nonterminal, then the terminal looks up the production indexed by X and a in the parsing table, replaces X by the right-hand side of the production, and remains active (expand action). Again, we need not consider error conditions as the type-checker will statically guarantee that parsing does not fail.

```plaintext
instance Id (E a) (T') (E' a') where -- expand: E → T E'
    id s (E ctx) = id s (T (λ t → E' (λ e' → ctx (e' t'))))
instance Id (T a) (T') (a' where) -- expand: T → F T'
    id s (T ctx) = id s (F (λ f → T' (λ t' → ctx (t' f))))
instance Id (F a) a where -- expand: F → id
    id s (F ctx) = id s (I (λ v → ctx v))
instance Id (I a) a where -- pop
    id s (I ctx) = return (ctx (Id s))
instance Add (E a) (T (E' a)) where -- expand: E' → + T E'
    + (E' ctx) = + (A (T (λ t → E' (λ e' → ctx (λ e → e'))))
instance Add (a' a) \Rightarrow Add (T' a) a' where -- expand: T' → e
    + (T' ctx) = + (ctx (λ e → e))
instance Add (A a) a where -- pop
    + (A ctx) = return ctx
instance Mul (T a) (F (T' a)) where -- expand: T' → * F T'
    * (T' ctx) = * (M (F (λ f → T' (λ t' → ctx (λ e → e'))))))
instance Mul (M a) a where -- pop
    * (M ctx) = return ctx
instance Open (E a) (E (C (T' (E' a)))) where -- expand: E → T E'
    (E ctx) = (T (λ t → E' (λ e' → ctx (e' t'))))
instance Open (T a) (E (C (T' a))) where -- expand: T → F T'
    (T ctx) = (F (λ f → T' (λ t' → ctx (t' f))))
instance Open (F a) (E (C a)) where -- expand: F → ( E )
    (F ctx) = (O (E (λ e → C (ctx e))))
instance Open (O a) a where -- pop
    (O ctx) = return ctx
instance Close a a' \Rightarrow Close (E a) a' where -- expand: E' → e
    (E' ctx) = (ctx (λ e → e))
instance Close a a' \Rightarrow Close (T a) a' where -- expand: T' → e
    (T' ctx) = (ctx (λ e → e))
instance Close (C a) a where -- pop
    (C ctx) = return ctx
instance Endquote a \Rightarrow Endquote (E a) where -- expand: E' → e
    » (E' ctx) = » (ctx (λ e → e))
instance Endquote a \Rightarrow Endquote (T a) where -- expand: T' → e
    » (T' ctx) = » (ctx (λ e → e))
instance Endquote Expr where -- pop
    » e = e

Fig. 2. The LL(1) parser for the expression grammar, part 2.

The table also includes a column for “ »,” which serves as an end marker.

The state is now a stack of pending symbols, both terminal and nonterminal. The symbol X on top of the stack determines the action of the current active terminal a.

If X = a, then the terminal pops X from the stack and passes control to the next active terminal (pop action). If X is a nonterminal, then the terminal looks up the production indexed by X and a in the parsing table, replaces X by the right-hand side of the production, and remains active (expand action). Again, we need not consider error conditions as the type-checker will statically guarantee that parsing does not fail.
Since the state is a stack of symbols, we must introduce types both for terminal and nonterminal symbols (we only show some representative examples here, the complete code is listed in Figures 1 and 2):

```haskell
newtype E α = E (Expr → α) -- E
newtype I α = I (Expr → α) -- id
```

Like nonterminals, terminal symbols may carry semantic information: the terminal `id s`, for instance, returns the semantic value `Id s` of type `Expr`. For each terminal symbol including “»,” we introduce a class and an instance that implements the pop action:

```haskell
class Add old new | old → new where
  + :: old → new
instance Add (A α) α where
  + (Act x) = return ctx
```

Finally, each entry of the parsing table gives rise to an instance declaration that implements the expand action. Here are the instances for “+”:

```haskell
instance Add (E' α) (T (E' α)) where
  + (E' ctx) = + (A (T (λt → E' (λe' → ctx (λl → e' (Add l t))))))
instance (Add α α') ⇒ Add (T' α) α' where
  + (T' ctx) = + (ctx (λe → e))
```

Since the look-ahead token is unchanged, the second instance requires an additional context, `Add α α'`, which accounts for the “recursive” call to “+.” The first instance also contains a call to “+” but this occurrence can be statically resolved: it refers to the “pop instance.” In general, an instance for the production `A → ω` requires a context iff `e ∈ L(ω)`, as in this case the stack shrinks. The instance head always reflects the parsing state after the final pop action. Consider the `id` instance

```haskell
instance Id (E α) (T' (E' α)) where
  id s (E ctx) = id s (T (λt → E' (λe' → ctx (e' t))))
```

The expansion phase proceeds `E → T E' → F T' E' → id T' E'`. Consequently, the instance head records that the state changes from `E` to `T' E'`.

It remains to implement `quote` and `antiquote`.

```haskell
quote = return (E (λe → e))
antiquote :: E st → Expr → st
antiquote (E st) e = return (st e)
```

The type of `antiquote` dictates that we can only splice an expression into a position where the nonterminal `E` is expected. This can be achieved by enclosing the antiquotation in “(” and “).”

---

3 The standard construction of parsing tables using `First` and `Follow` sets already provides the necessary information: the parsing table contains the production `A → ω` for look-ahead `a` if either `a ∈ First(ω)`, or `e ∈ L(ω)` and `a ∈ Follow(A)`. Only in the second case is a context required.
Remark 3

We have implemented the parsing actions in a rather ad-hoc way. Alternatively, they can be written using `lift` and monadic composition:

\[
\text{instance } \text{Id} (E \, x) (T' \, (E' \, x)) \text{ where }
\]

\[
\text{id } s = \text{id } s \circ \text{lift } (\lambda(\text{ctx} \rightarrow T \, (\lambda t \rightarrow E' \, (\lambda e' \rightarrow \text{ctx} \, (e' \, t)))))
\]

\[
(\circ) \quad :: (b \rightarrow \text{CPS} \, c) \rightarrow (a \rightarrow \text{CPS} \, b) \rightarrow (a \rightarrow \text{CPS} \, c)
\]

\[
q \circ p = \lambda a \rightarrow p \, a \gg q
\]

The rewrite nicely separates the expansion step from the “recursive call.”

## 5 Bottom-up parsing: LR(0) parsing

Let us move on to our final challenge: quotations whose concrete syntax is based on an LR(0) grammar—LR(1) grammars are also doable but we resist the temptation to spell out the details. Unlike LL parsing, the LR method is generally not suitable for constructing parsers by hand. For that reason, we shall base the treatment on a very simple example, the language of balanced parentheses.

\[
P \rightarrow \epsilon | P \, (P)
\]

The abstract syntax is given by the `Tree` datatype: \( P \rightarrow \epsilon \) constructs a `Leaf`, \( P \rightarrow P \, (P) \) a `Fork`.

An LR parser is similar to a postfix parser in that the state is a stack of symbols the parser has already seen. In contrast, the state of an LL parser is a stack of symbols it expects to see.

For efficiency reasons, an LR parser maintains additional information that summarises the stack configuration in each step. This is accomplished by a finite-state machine, the so-called LR(0) `automaton`. Here is the automaton for the grammar above (\( S \) is a new start symbol, “»” serves as an end marker).

```
\[
S \rightarrow \cdot P
1: P \rightarrow \cdot
P \rightarrow \cdot P(P)
\]

\[
S \rightarrow P \cdot »
2: P \rightarrow P \cdot (P)
\]

\[
P \rightarrow P \cdot (P)
3: S \rightarrow P \cdot »
\]

\[
P \rightarrow P \cdot (P)
5: P \rightarrow \cdot
P \rightarrow \cdot P(P)
\]

\[
P \rightarrow P \cdot (P)
4: P \rightarrow P \cdot (P)
\]

\[
P \rightarrow P \cdot (P)
6: P \rightarrow P(P)
\]
```

The automaton has six states; the production(s) contained in the states illustrate the progress of the parse with the dots marking the borderline between what we have seen and what we expect to see. If the dot appears before a terminal symbol,
we have a shift state (colored in yellow/light gray). If the dot has reached the end in one of the productions, we have a reduce state (colored in blue/dark gray). In a shift state, the parser consumes an input token and pushes it onto the stack. In a reduce state, the right-hand side of a production resides on top of the stack, which is then replaced by the left-hand side.

In our example, the parser first reduces $P \rightarrow \epsilon$ moving from the start state 1 to state 2. Then, it shifts either “)” or “.” Each transition is recorded on the stack. This information is used during a reduction to determine the next state. Consider state 6; two sequences of moves end in this state: $1 \xrightarrow{P} 2 \xrightarrow{5} 6$ and $5 \xrightarrow{4} 6$. Removing $P \langle P \rangle$ from the stack means returning to either state 1 or state 5. Pushing $P$ onto the stack, we move forward to either state 2 or state 4. In short, reducing $P \rightarrow P \langle P \rangle$ is accomplished by replacing the above transitions by either $1 \xrightarrow{P} 2$ or $5 \xrightarrow{P} 4$. The point is that, in general, there are several transitions for a single production.

Turning to the implementation, the parser’s state is a stack of LR(0) states. Each LR(0) state carries the semantic value of the unique symbol that annotates the ingoing edges.

```haskell
data S1 = S1    -- S
data S2 st = S2 Tree st -- P
data S3 st = S3 st -- ”
data S4 st = S4 Tree st -- P
data S5 st = S5 st -- ( data S6 st = S6 st -- )
```

Each state of the automaton is translated into a function that performs the corresponding action. A shift state simply delegates the control to the next active terminal. A reduce state pops the transitions corresponding to the right-hand side from the stack and pushes a transition corresponding to the left-hand side. If there are several possible transitions, then a reduce action is given by a family of functions represented as a type class.

```haskell
quote = state1 S1    -- start
state1 st = state2 (S2 Leaf st)    -- reduce
state2 st = return st    -- shift
state3 (S3 (S2 t S1)) = t    -- accept
state4 st = return st    -- shift
state5 st = state4 (S4 Leaf st)    -- reduce
class State6 old new | old -> new where
state6 :: old -> new
instance State6 (S6 (S4 (S5 (S2 S1))))) (S2 S1) where    -- reduce
    state6 (S6 (S4 u (S5 (S2 t S1)))) = state2 (S2 (Fork t u) S1)
instance State6 (S6 (S4 (S5 (S4 (S5 st)))) (S4 (S5 st))) where    -- reduce
    state6 (S6 (S4 u (S5 (S4 t (S5 st))))) = state4 (S4 (Fork t u) (S5 st))
```

https://doi.org/10.1017/S0956796811000050 Published online by Cambridge University Press
The pattern $S_6 (S_4 u (S_5 (S_4 t (S_5 st))))$ nicely shows the interleaving of states and semantic values. Since the stack is nested to the right, $u$ is the topmost semantic value and consequently becomes the right subtree in $\text{Fork } t u$.

The active terminals implement the shift actions.

```haskell
class Open old new | old -> new where
  ( :: old -> new

instance Open (S_2 st) (S_4 (S_5 (S_2 st))) where
  ( st @(S_2 _) ) = state_5 (S_5 st)

instance Open (S_4 st) (S_4 (S_5 (S_4 st))) where
  ( st @(S_4 _) ) = state_5 (S_5 st)

) st@(S_4 _) = state_6 (S_6 st)

endquote st @(S_2 _) = state_3 (S_3 st)
```

We need a class if a terminal annotates more than one edge. Again, the instance types are not entirely straightforward as they reflect the stack modifications up to the next shift: for instance, “(” moves from $S_2$ to $S_5$ and then to $S_4$, which is again a shift state.

The implementation technique also works for LR(1) grammars. In this case, the active terminals implement both shift and reduce actions.

6 Conclusion

Quotations provide a new, amusing perspective on parsing: terminal symbols turn active and become the driving force of the parsing process. It is quite remarkable that all major syntax analysis techniques can be adapted to this technique.

Typed quotations provide static guarantees: using type-level representations of symbols Haskell’s type-checker is instrumented to scrutinise whether a quotation is syntactically correct. Of course, this means that syntax errors become type errors, which are possibly difficult to decipher. Adding proper error handling is left as the obligatory “instructive exercise to the reader.”

References


https://doi.org/10.1017/S0956796811000050 Published online by Cambridge University Press