[**TODO:** Fill in acknowledgements in cover.tex]

[**TODO:** Move related work to end, so it flows better as explaining various other ways of parsing? Or introduce CFGs earlier.]

[**TODO:** Fill in related work with detailed explanations of various ways of writing parsers]

[**TODO:** Splitter now returns numbers, not strings]

[**TODO:** Explain how soundness can be done without parser extensionality, at the cost of algorithmic complexity elsewhere.]

[**TODO:** Find a citation for Fiat, test with [**fiat**]]

[**TODO:** (Optional) Section on showing that parser has "reasonable" performance on grammars with non-brute-force splitter (by using arrays and native strings)]

[**TODO:** (Really Optional) Section on building parse trees, not just recognizers)]

[**QUESTION FOR ADAM:** Should I include an appendix of all of the code of Fiat and parsers that is used, rendered verbatim?]

An Extensible Framework for Synthesizing Efficient, Verified Parsers

by

Jason S. Gross

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

Author . . . .
Department of Electrical Engineering and Computer Science
July 9, 2015

Certified by . . . .
Adam Chlipala
Associate Professor without Tenure of Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

An Extensible Framework for Synthesizing Efficient, Verified Parsers

by

Jason S. Gross

Submitted to the Department of Electrical Engineering and Computer Science
on July 9, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Parsers have a long history in computer science. This thesis proposes a novel approach to synthesizing efficient, verified parsers by refinement, and presents a demonstration of this approach in the Fiat framework by synthesizing a parser for arithmetic expressions. The benefits of this framework may include more flexibility in the parsers that can be described, more control over the low-level details when necessary for performance, and automatic or mostly automatic correctness proofs.

Thesis Supervisor: Adam Chlipala
Title: Associate Professor without Tenure of Computer Science

Acknowledgments

This is the acknowledgements section. You should replace this with your own acknowledgements. **[TODO:** Fill this in]

Contents

Chapter 1

Parsing Context-Free Grammars

We begin with an overview of the general setting and a description of our approach to parsing.

1.1 Parsing

The job of a parser is to decompose a flat list of characters, called a *string*, into a structured tree, called a *parse tree*, on which further operations can be performed. As a simple example, we can parse `"ab"` as an instance of the regular expression $\texttt{(ab)}^*$, giving this parse tree, where we write $\cdot$ for string concatenation.

$$\dfrac{\dfrac{\dfrac{}{\texttt{"a"} \in \texttt{'a'}} \qquad \dfrac{}{\texttt{"b"} \in \texttt{'b'}} \qquad \dfrac{\dfrac{}{\texttt{""} \in \epsilon}}{\texttt{""} \in \texttt{(ab)}^*}}{\texttt{"a"} \cdot \texttt{"b"} \cdot \texttt{""} \in \texttt{ab(ab)}^*}}{\texttt{"ab"} \in \texttt{(ab)}^*}$$

Our parse tree is implicitly constructed from a set of general inference rules for parsing. There is a naive approach to parsing a string s: run the inference rules as a logic program. Several execution orders work: we may proceed bottom-up, by generating all of the strings that are in the language and not longer than s, checking each one for equality with s; or top-down, by splitting s into smaller parts in a way that mirrors the inference rules. In this paper, we present an implementation based on the second strategy, parameterizing over a "splitting oracle" that provides a list of candidate locations at which to split the string, based on the available inference rules. Soundness of the algorithm is independent of the splitting oracle; each location in the list is attempted. To be complete, if any split of the string yields a valid parse, the oracle must give at least one splitting location that also yields a valid parse. Different splitters yield different simple recursive-descent parsers.

There is a trivial, brute-force splitter that suffices for proving correctness: simply return the list of all locations in the string, the list of all numbers between 0 and the length of the string. Because we construct a parser that terminates no matter what list it is given, and all valid splits are trivially in this list, this splitting "oracle" is enough to fill the oracle-shaped-hole in the correctness proofs. Thus, we can largely separate concerns about correctness and concerns about efficiency. In Chapter 4, we focus only on correctness; we set up the framework we use to achieve efficiency in Chapter 5, and we demonstrate the use of the framework in Chapters 6, 7 and 8.

Although this simple splitter is sufficient for proving the algorithm correct, it is horribly inefficient, running in time $\mathcal{O}(n!)$, where n is the length of the string. We synthesize more efficient splitters in later chapters; we believe that parameterizing the parser over a splitter gives us enough expressiveness to implement essentially all optimizations of interest, while being a sufficiently simple language to make proofs relatively straightforward. For example, to achieve linear parse time on the $\texttt{(ab)}^*$ grammar, we could have a splitter that, when trying to parse $\texttt{'c}_1\texttt{'} \cdot \texttt{'c}_2\texttt{'} \cdot \texttt{s}$ as $\texttt{ab(ab)}^*$, splits the string into ($\texttt{'c}_1\texttt{'}$, $\texttt{'c}_2\texttt{'}$, $\texttt{s}$); and when trying to parse $\texttt{s}$ as ϵ, does not split the string at all.

Parameterizing over a splitting oracle allows us to largely separate correctness concerns from efficiency concerns.

Proving completeness—that our parser succeeds whenever there is a valid parse tree—is conceptually straightforward: trace the algorithm, showing that if the parser returns `false` at a given point, then assuming a corresponding parse tree exists yields a contradiction. The one wrinkle in this approach is that the algorithm, the logic program, is not guaranteed to terminate.

1.1.1 Infinite Regress

Since we have programmed our parser in Coq, our program must be terminating by construction. However, naive recursive-descent parsers do not always terminate!

To see how such parsers can diverge, consider the following example. When defining the grammar $\texttt{(ab)}^*$, perhaps we give the following production rules:

$$\frac{s \in \epsilon}{s \in \texttt{(ab)}^*}\,(\epsilon) \qquad \frac{s_0 \in \texttt{'a'} \qquad s_1 \in \texttt{'b'}}{s_0 s_1 \in \texttt{(ab)}^*}\,(\texttt{"ab"})$$

$$\frac{s_0 \in \texttt{(ab)}^* \qquad s_1 \in \texttt{(ab)}^*}{s_0 s_1 \in \texttt{(ab)}^*}\,(\texttt{(ab)}^*\texttt{(ab)}^*)$$

Now, let us try to parse the string `"ab"` as $\texttt{(ab)}^*$:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\texttt{""} \in \epsilon}}{\texttt{""} \in \texttt{(ab)}^*} \quad \dfrac{\dfrac{\dfrac{\dfrac{}{\texttt{""} \in \epsilon}}{\texttt{""} \in \texttt{(ab)}^*} \quad \dfrac{\iddots}{\texttt{"ab"} \in \texttt{(ab)}^*}}{\texttt{""} \cdot \texttt{"ab"} \in \texttt{(ab)}^*}}{\texttt{"ab"} \in \texttt{(ab)}^*}}{\texttt{""} \cdot \texttt{"ab"} \in \texttt{(ab)}^*}}{\texttt{"ab"} \in \texttt{(ab)}^*}$$

Thus, by making a poor choice in how we split strings and choose productions, we can quickly hit an infinite regress.

Assuming we have a function `split` : `String` → `[String × String]` which is our splitting oracle, we may write out a potentially divergent parser specialized to this grammar.

```
any_parses : [String × String] → Bool
any_parses [] ≔ false
any_parses (("a", "b") :: _) ≔ true
any_parses ((s₁, s₂) :: rest_splits)
    ≔ (parses s₁ && parses s₂) || any_parses rest_splits

parses : String → Bool
parses "" ≔ true
parses str ≔ any_parses (split str)
```

If `split` returns `("", "ab")` as the first item in its list when given `"ab"`, then the code given above will diverge in the way demonstrated above with the infinite derivation tree.

1.1.2 Aborting Early

To work around this wrinkle, we keep track of what nonterminals we have not yet tried to parse the current string as, and we abort early if we see a repeat. Note that this strategy only works for grammars with finite sets of nonterminals, in line with most formalizations of context-free grammars. For our example grammar, since there is only one nonterminal, we only need to keep track of the current string. We refactor the above code to introduce a new parameter `prev_s`, recording the previous string we were parsing. We use `s` < `prev_s` to denote the test that `s` is strictly shorter than prev_s.

```
any_parses : String → [String × String] → Bool
any_parses _ [] := false
any_parses _ (("a","b") :: _) := true
any_parses prev_s ((s₁, s₂) :: rest_splits)
   := (s₁ < prev_s && s₂ < prev_s
       && parses s₁ && parses s₂)
      || any_parses prev_s rest_splits

parses : String → Bool
parses "" := true
parses str := any_parses str (split str)
```

We can convince Coq that this definition is total via well-founded recursion on the length of the string passed to parses. For a more-complicated grammar, we'd need to use a well-founded relation that also included the number of nonterminals not yet tried for this string; we do this in Figure 4-2 in Subsection 4.3.2.

With this refactoring, however, completeness is no longer straightforward. We must show that aborting early does not eliminate good parse trees.

We devote the rest of this paper to describing an elegant approach to proving completeness. Ridge [12] carried out a proof about essentially the same algorithm in HOL4, a proof assistant that does not support dependent types. We instead refine our parser to have a more general polymorphic type signature that takes advantage of dependent types, supporting a proof strategy with a different kind of aesthetic appeal. Relational parametricity frees us from worrying about different control flows with different instantiations of the arguments: when care is taken to ensure that the execution of the algorithm does not depend on the values of the arguments, we are guaranteed that all instantiations succeed or fail together. Freed from this worry, we convince our parser to prove its own soundness and completeness by instantiating its arguments correctly.

1.2 Standard Formal Definitions

Before proceeding, we pause to standardize on terminology and notation for context-free grammars and parsers. In service of clarity for some of our later explanations, we formalize grammars via natural-deduction inference rules, a slightly nonstandard choice.

1.2.1 Context-Free Grammar

A *context-free grammar* consists of *items*, which may be either *terminals* (characters) or *nonterminals*; plus a set of *productions*, each mapping a nonterminal to a sequence of items.

Example: $\texttt{(ab)}^*$

The inference rules of the regular-expression grammar $\texttt{(ab)}^*$ are:

Terminals:

$$\frac{}{\texttt{"a"} \in \texttt{'a'}} \qquad \frac{}{\texttt{"b"} \in \texttt{'b'}}$$

Productions and nonterminals:

$$\frac{s \in \epsilon}{s \in \texttt{(ab)}^*} \qquad \frac{}{\texttt{""} \in \epsilon}$$

$$\frac{s_0 \in \texttt{'a'} \qquad s_1 \in \texttt{'b'} \qquad s_2 \in \texttt{(ab)}^*}{s_0 s_1 s_2 \in \texttt{(ab)}^*}$$

An alternative, more standard, more compact, notation for specifying context free grammars would present this grammar as:

$$\texttt{(ab)}^* \;::=\; \epsilon \mid \texttt{'a'}\ \texttt{'b'}\ \texttt{(ab)}^*$$

This can be read as saying that there is a single nonterminal $\texttt{(ab)}^*$, which parses empty strings, as well as parsing strings which are an `'a'`, followed by a `'b'`, followed by a string which parses as the nonterminal $\texttt{(ab)}^*$.

1.2.2 Parse Trees

A string `s` *parses* as:

- a given terminal `ch` iff `s` = `'ch'`.
- a given sequence of items $\texttt{x}_i$ iff `s` splits into a sequence of strings $\texttt{s}_i$, each of which parses as the corresponding item $\texttt{x}_i$.
- a given nonterminal `nt` iff `s` parses as one of the item sequences that `nt` maps to under the set of productions.

We may define mutually inductive dependent type families of ParseTreeOfs and ParseItemsTreeOfs for a given grammar:

$$\texttt{ParseTreeOf} : \texttt{Item} \to \texttt{String} \to \textbf{Type}$$
$$\texttt{ParseItemsTreeOf} : \texttt{[Item]} \to \texttt{String} \to \textbf{Type}$$

For any terminal character $\texttt{ch}$, we have the constructor

$$(\texttt{'ch'}) : \texttt{ParseTreeOf 'ch' "ch"}$$

For any production $\texttt{rule}$ mapping a nonterminal $\texttt{nt}$ to a sequence of items $\texttt{its}$, and any string $\texttt{s}$, we have this constructor:

$$(\texttt{rule}) : \texttt{ParseItemsTreeOf}\ \ \texttt{its}\ \ \texttt{s} \to \texttt{ParseTreeOf}\ \ \texttt{nt}\ \ \texttt{s}$$

We have the following two constructors of ParseItemsTree. In writing the type of the latter constructor, we adopt a common space-saving convention where we assume that all free variables are quantified implicitly with dependent function (Π) types. We also write constructors in the form of schematic natural-deduction rules, since that notation will be convenient to use later on.

$$\frac{}{\texttt{""} \in \epsilon} : \texttt{ParseItemsTreeOf}\ \ \texttt{[]}\ \ \texttt{""}$$

$$\frac{\texttt{s}_1 \in \texttt{it} \qquad \texttt{s}_2 \in \texttt{its}}{\texttt{s}_1\texttt{s}_2 \in \texttt{it} :: \texttt{its}} : \begin{aligned}&\texttt{ParseTreeOf}\ \ \texttt{it}\ \ \texttt{s}_1\\ &\to \texttt{ParseItemsTreeOf}\ \ \texttt{its}\ \ \texttt{s}_2\\ &\to \texttt{ParseItemsTreeOf}\ \ (\texttt{it} :: \texttt{its})\ \ \texttt{s}_1\texttt{s}_2\end{aligned}$$

For brevity, we will sometimes use the notation $\overline{\texttt{s} \in \texttt{X}}$ to denote both ParseTreeOf $\texttt{X}$ $\texttt{s}$ and ParseItemsTreeOf $\texttt{X}$ $\texttt{s}$, relying on context to disambiguate based on the type of $\texttt{X}$. Additionally, we will sometimes fold the constructors of ParseItemsTreeOf into the ($\texttt{rule}$) constructors of ParseTreeOf, to mimic the natural-deduction trees.

We also define a type of all parse trees, independent of the string and item, as this dependent-pair (Σ) type, using set-builder notation; we use ParseTree to denote the type

$$\{(\texttt{nt}, \texttt{s}) : \texttt{Nonterminal} \times \texttt{String} \mid \texttt{ParseTreeOf}\ \texttt{nt}\ \texttt{s}\}$$

Chapter 2

Completeness and Soundness

Parsers come in a number of flavors. The simplest flavor is the *recognizer*, which simply says whether or not there exists a parse tree of a given string for a given nonterminal; it returns Booleans. There is also a richer flavor of parser that returns inhabitants of `option ParseTree`.

For any recognizer $\texttt{has\_parse} : \texttt{Nonterminal} \to \texttt{String} \to \texttt{Bool}$, we may ask whether it is *sound*, meaning that when it returns `true`, there is always a parse tree; and *complete*, meaning that when there is a parse tree, it always returns `true`. We may express these properties as theorems (alternatively, dependently typed functions) with the following type signatures:

$$\begin{aligned}
\texttt{has_parse_sound} : (\texttt{nt} : \texttt{Nonterminal}) &\to (\texttt{s} : \texttt{String}) \\
&\to \texttt{has_parse nt s} = \texttt{true} \\
&\to \texttt{ParseTreeOf nt s} \\
\texttt{has_parse_complete} : (\texttt{nt} : \texttt{Nonterminal}) &\to (\texttt{s} : \texttt{String}) \\
&\to \texttt{ParseTreeOf nt s} \\
&\to \texttt{has_parse nt s} = \texttt{true}
\end{aligned}$$

For any parser

$$\texttt{parse} : \texttt{Nonterminal} \ \to\ \texttt{String} \ \to\ \texttt{option ParseTree},$$

we may also ask whether it is sound and complete, leading to theorems with the following type signatures, using $\mathtt{p_1}$ to denote the first projection of $\mathtt{p}$:

$$
\begin{aligned}
\mathtt{parse\_sound} : {} & (\mathtt{nt} : \mathtt{Nonterminal}) \\
& \to (\mathtt{s} : \mathtt{String}) \\
& \to (\mathtt{p} : \mathtt{ParseTree}) \\
& \to \mathtt{parse}\ \mathtt{nt}\ \mathtt{s} = \mathtt{Some}\ \mathtt{p} \\
& \to \mathtt{p_1} = (\mathtt{nt}, \mathtt{s}) \\
\mathtt{parse\_complete} : {} & (\mathtt{nt} : \mathtt{Nonterminal}) \\
& \to (\mathtt{s} : \mathtt{String}) \\
& \to \mathtt{ParseTreeOf}\ \mathtt{nt}\ \mathtt{s} \\
& \to \mathtt{parse}\ \mathtt{nt}\ \mathtt{s} \neq \mathtt{None}
\end{aligned}
$$

Since we are programming in Coq, this separation into code and proof actually makes for more awkward type assignments. We also have the option of folding the soundness and completeness conditions into the types of the code. For instance, the following type captures the idea of a sound and complete parser returning parse trees, using the type constructor $+$ for disjoint union (i.e., sum or variant type):

$$
\begin{aligned}
\mathtt{parse} : {} & (\mathtt{nt} : \mathtt{Nonterminal}) \\
& \to (\mathtt{s} : \mathtt{String}) \\
& \to \mathtt{ParseTreeOf}\ \mathtt{nt}\ \mathtt{s} + (\mathtt{ParseTreeOf}\ \mathtt{nt}\ \mathtt{s} \to \bot)
\end{aligned}
$$

That is, given a nonterminal and a string, `parse` either returns a valid parse tree, or returns a *proof* that the existence of any parse tree is *contradictory* (i.e., implies $\bot$, the empty type). Our implementation follows this dependently typed style. Our main initial goal in the project was to arrive at a `parse` function of just this type, generic in an arbitrary choice of context-free grammar, implemented and proven correct in an elegant way.

Chapter 3

Related Work and Other Approaches to Parsing

Stepping back a bit, we describe how our approach to parsing relates to existing work.

[**TODO:** Mention LR (and LL(1)?) parsers: "The field of parsing is one of the most venerable in computer-science. Still with us are a variety of parsing approaches born in times of much more severe constraints on memory and processor speed, including various flavors of LR parsers, which apply only to strict subsets of the context-free grammars, to guarantee ability to predict which production applies based on finite look-ahead into a string."]

3.1 Recursive Descent Parsing

The most conceptually straightforward approaches to parsing fall into the class called recursive descent parsing, where, to parse a string s as a given production p, you attempt to parse various parts of s as each of the items in the list p. The control flow of the code mirrors the structure of the grammar, as well as the structure of the eventual parse tree, descending down the branches of the parse tree, recursively calling itself at each step.

3.1.1 Parser Combinators

A popular approach, called *combinator parsing* [6], to implementing recursive descent parsing involves writing a small set of typed combinators, or higher-order functions, which are then applied to each other in various combinations to write a parser that mimics closely the structure of the grammar.

Essentially, parsers defined via parser combinators answer the question "what prefixes of a given string be parsed as a given item?" Each function returns a list of postfixes of the string they are passed, indicating all of the strings that might remain for the other items in a given rule. [**QUESTION FOR ADAM:** Should I mention semantic actions?]

Basic Combinators

We now define the basic combinators. In the simplest form, each combinator takes in a string, and returns a list of strings (the postfixes); we can define the type `parser` as `String` → `[String]`. We can define the empty-string parser, as well as the parser for a nonterminal with no production rules, which always fails:

$$\begin{aligned}\epsilon &: \texttt{parser}\\ \epsilon\ \texttt{str} &= [\texttt{str}]\\[1em] \texttt{fail} &: \texttt{parser}\\ \texttt{fail}\ \_ &= []\end{aligned}$$

Failure is indicated by returning the empty list; success at parsing the entire string is indicated by returning a list containing the empty string.

The parser for a given terminal fails if the string does not start with that character, and returns all but the first character if it does:

$$\begin{aligned}\texttt{terminal} &: \texttt{Char} \to \texttt{parser}\\ \texttt{terminal}\ \texttt{ch}\ (\texttt{ch} :: \texttt{str}) &= [\texttt{str}]\\ \texttt{terminal}\ \_\ \ \texttt{""} &= []\end{aligned}$$

We now define combinators for sequencing and alternatives:

$$\begin{aligned}(\texttt{>>>}) &: \texttt{parser} \to \texttt{parser} \to \texttt{parser}\\ (p_0 \texttt{ >>> } p_1)\ \texttt{str} &= \texttt{flat\_map}\ p_1\ (p_0\ \texttt{str})\\[1em] (\texttt{|||}) &: \texttt{parser} \to \texttt{parser} \to \texttt{parser}\\ (p_0 \texttt{ ||| } p_1)\ \texttt{str} &= p_0\ \texttt{str} \texttt{ ++ } p_1\ \texttt{str}\end{aligned}$$

where `++` is list concatenation, and `flat_map`, which concatenates the lists returned by mapping its first argument over each of the elements in its second argument, has type `(A → [B]) → [A] → [B]`, where `[A]` is the type denoting a list of elements, all of type `A`.[**TODO:** Check if I've defined this notation already.]

An Example

We can now easily define a parser for the grammar (ab)*:

```
parse-(ab)* : parser
parse-(ab)* = (terminal 'a' >>> terminal 'b' >>> parse-(ab)*) ||| ϵ
```

Note that, by putting ϵ last, we ensure that this parser returns the list in order of longest parse (shortest postfix) to shortest parse (longest postfix).

Proving Correctness and Dealing with Nontermination

Although parser combinators are straightforward, it is easy to make them loop forever. It is well-known that parsers defined naively using parser combinators don't handle grammars with *left recursion*, where the first item in a given production rule is the nonterminal currently being defined. For example, if we have the nonterminal `expr` ::= `number` | `expr '+' expr`, then the parser for `expr '+' expr` will call the parser for `expr`, which will call the parse for `expr '+' expr`, which will quickly loop forever.

The algorithm we presented in Subsection 1.1.2 is essentially the same as the algorithm Ridge presents in [12] to deal with this problem. By wrapping the calls to the parsers, in each combinator, with a function that prunes duplicative calls, Ridge provides a way to ensure that parsers terminate. Also included in [12] are proofs in HOL4 that such wrapped parsers are both sound (and therefore terminating) and complete. Furthermore, Ridge's parser has worst-case $O(n^5)$ running time in the input-string length.

3.1.2 Parsing with Derivatives

Might, Darais, and Spiewak describe an elegant method for recursive descent parsing in [9], based on Brzozowski's derivatives [4], which might be considered a conceptual dual to standard combinator parser. Rather than returning a list of possible string remnants, constructed by recursing down the structure of the grammar, we can iterate down the characters of a string, computing an updated language, or grammar, at each point.

The *language* defined by a grammar is a the set of strings accepted by that grammar. Here we describe the mathematical ideas behind parsing with derivatives. Might et al. take a slightly different approach to ensure termination; where we will describe the mathematical operations on languages, they define these operations on a structural representation of the language, akin to an inductive definition of the grammar.

Much as we defined a parser combinators for the elementary operations of a grammar (ϵ, terminals, sequencing, and alternatives), we can define similar combinators for defining a (lazy, or coinductive) language for a grammar. Defining the type `language` to be a set (or just a coinductive list) of strings, we have:

$$\begin{array}{l}
\epsilon \ : \ \texttt{language} \\
\epsilon \ = \ \{\texttt{""}\} \\
\\
\texttt{terminal} \ : \ \texttt{Char} \ \rightarrow \ \texttt{language} \\
\texttt{terminal ch} \ = \ \{\texttt{ch}\} \\
\\
(\texttt{>>>}) \ : \ \texttt{language} \ \rightarrow \ \texttt{language} \ \rightarrow \ \texttt{language} \\
\mathcal{L}_0 \ \texttt{>>>} \ \mathcal{L}_1 \ = \ \{\ \texttt{s}_0\texttt{s}_1 \mid \texttt{s}_0 \in \mathcal{L}_0 \text{ and } \texttt{s}_1 \in \mathcal{L}_1\} \\
\\
(\texttt{|||}) \ : \ \texttt{language} \ \rightarrow \ \texttt{language} \ \rightarrow \ \texttt{language} \\
\mathcal{L}_0 \ \texttt{|||} \ \mathcal{L}_1 \ = \ \mathcal{L}_0 \cup \mathcal{L}_1
\end{array}$$

The essential operations for computing derivatives are *filtering* and *chopping*. To *filter* a language $\mathcal{L}$ by a character c is to take the subset of strings in $\mathcal{L}$ which start with c. To chop a language $\mathcal{L}$ is to remove the first character from every string. The derivative $D_c(\mathcal{L})$ with respect to c of a language $\mathcal{L}$ is then the language $\mathcal{L}$, filtered by c and chopped:

$$\begin{array}{l}
\texttt{D}_\texttt{c} \ : \ \texttt{language} \ \rightarrow \ \texttt{language} \\
\texttt{D}_\texttt{c} \ \emptyset \ = \ \emptyset \\
\texttt{D}_\texttt{c} \ ((\texttt{ch} :: \texttt{str}) :: \mathcal{L}) \ = \ \texttt{str} :: (\texttt{D}_\texttt{c} \ \mathcal{L}) \\
\texttt{D}_\texttt{c} \ (_ :: \mathcal{L}) \ = \ \texttt{D}_\texttt{c} \ \mathcal{L}
\end{array}$$

We can then define a `has_parse` proposition by taking successive derivatives:

$$\begin{array}{l}
\texttt{has_parse} \ : \ \texttt{language} \ \rightarrow \ \texttt{String} \ \rightarrow \ \textbf{Prop} \\
\texttt{has_parse} \ \mathcal{L} \ \texttt{""} \ = \ \texttt{""} \in \mathcal{L} \\
\texttt{has_parse} \ \mathcal{L} \ (\texttt{ch} :: \texttt{str}) \ = \ \texttt{has_parse} \ (\texttt{D}_\texttt{ch} \ \mathcal{L}) \ \texttt{str}
\end{array}$$

To ensure termination and good performance, Might et al. define the derivative operation on the structure of the grammar, rather than defining combinators that turn a grammar into a language, and furthermore take advantage of laziness and memoization. After adding code to prune the resulting language of useless content, they argue that the cost of parsing with derivatives is $\mathcal{O}(n|G|)$ on average, where n is the length of the input string and $|G|$ is the size of the grammar.

Formal Verification

Almeida et al. formally verify, in Coq, finite automata for parsing the fragment of derivative-based parsing which applies to regular expressions. [1] This fragment dates back to Brzozowski's original presentation of derivatives. [4][**QUESTION FOR ADAM:** Should I define "finite automata?]

[**TODO:** Fill in content] [**QUESTION FOR ADAM:** Am I missing any branches?]

- Say something about recursive descent parsing seeming obvious and trivial only after writing out the types inductively?
- Say something about LR parsers and processor/memory constraints?
- Parser combinators
  - parse by consuming characters from the string.
  - Allow incremental parsing
- Look into what [5]s and [13]s are.
- Approaches to verifying parsers: correct-by-construction, induction. Other way? [**TODO:** Read paper]

[**TODO:** Rewrite the rest of this part so it's not just Adam's words]

However, despite rumors to the contrary, the field of parsing is far from dead. In the twentieth century, the functional-programming world experimented with a variety of approaches to *parser combinators* [6], where parsers are higher-order functions built from a small set of typed combinators. In the twenty-first century alone, a number of new parsing approaches have been proposed or popularized, including parsing expression grammars (PEGs) [5], derivative-based parsing [9], and GLL parsers [13].

However, our approach is essentially the same, algorithmically, as the one that Ridge demonstrated with a verified parser-combinator system [12], taking naive recursive-descent parsing and adding a layer to prune duplicative calls to the parser. His proof was carried out in HOL4, necessarily without using dependent types. Our new work may be interesting for the aesthetic appeal of our unusual application of dependent types to get the parser to generate some of its own soundness proof. Ridge's parser also has worst-case $O(n^5)$ running time in the input-string length. In the context of our verified implementation, we plan to explore a variety of optimizations based on clever, grammar-specific choices of string-splitter functions, which should have a substantial impact on the run-time cost of parsing some relevant grammars, and which we conjecture will not require any changes to the development presented in this paper.

A few other past projects have verified parsers with proof assistants, applying to derivative-based parsing [1] and SLR [2] and LR(1) [7] parsers. Several projects have used proof assistants to apply verified parsers within larger programming-language tools. RockSalt [10] does run-time memory-safety enforcement for x86 binaries, relying on a verified machine-code parser that applies derivative-based parsing for regular expressions. The verified Jitawa [11] and CakeML [8] language implementations include verified parsers, handling Lisp and ML languages, respectively.

Our final parser derivation relies on a relational parametricity property for polymorphic functions in Coq's type theory Gallina. With Coq as it is today, we need to prove this property manually for each eligible function, even though we can prove metatheoretically that it holds for them all. Bernardy and Guilhem [3] have shown how to extend type theories with support for materializing "free theorem" parametricity facts internally, and we might be able to simplify our implementation using such a feature. [**TODO:** Flesh this out, rewrite it so it's not all Adam's words, read the papers]

3.2 What's New and What's Old

The goal of this project is to demonstrate a new approach to generating parsers: incrementally building efficient parsers by refinement.

We begin with naive recursive-descent parsing.[**TODO:** Citation?] [**TODO:** Say something about viewing recursive-descent as an instance of general inhabitation-decision not being anywhere in the literature?] We ensure termination via memoization, a la [12]. We parameterize the parser on a "splitting oracle", which describes how to recurse (Section 1.1). As far as we can tell, the idea of factoring the algorithmic complexity like this is new.

We use Fiat to incrementally build efficient parsers by refinement; we describe Fiat starting in Chapter 5.

Additionally, we take a digression in Chapter 4 to describe how our parser can be used to prove its own completeness; the idea of reusing the parsing algorithm to generate proofs, parsing parse trees rather than strings, is not found in the literature, to the authors' knowledge.

Chapter 4

Completeness, Soundness, and Parsing Parse Trees

4.1 Proving Completeness: Conceptual Approach

Recall from Subsection 1.1.2 that the essential difficulty with proving completeness is dealing with the cases where our parser aborts early; we must show that doing so does not eliminate good parse trees.

The key is to define an intermediate type, that of "minimal parse trees." A "minimal" parse tree is simply a parse tree in which the same (string, nonterminal) pair does not appear more than once in any path of the tree. Defining this type allows us to split the completeness problem in two; we can show separately that every parse tree gives rise to a minimal parse tree, and that having a minimal parse tree in hand implies that our parser succeeds (returns `true` or `Some _`).

Our dependently typed parsing algorithm subsumes the soundness theorem, the minimization of parse trees, and the proof that having a minimal parse tree implies that our parser succeeds. We write one parametrically polymorphic parsing function that supports all three modes, plus the several different sorts of parsers (recognizers, generating parse trees, running semantic actions). That level of genericity requires us to be flexible in which type represents "strings," or inputs to parsers. We introduce a parameter that is often just the normal `String` type, but which needs to be instantiated as the type of *parse trees themselves* to get a proof of parse tree minimizability. That is, we "parse" parse trees to minimize them, reusing the same logic that works for the normal parsing problem.

Before presenting our algorithm's interface, we will formally define and explain minimal parse trees, which will provide motivation for the type signatures of our parser's arguments.

4.2 Minimal Parse Trees: Formal Definition

In order to make tractable the second half of the completeness theorem, that having a minimal parse tree implies that parsing succeeds, it is essential to make the inductive structure of minimal parse trees mimic precisely the structure of the parsing algorithm. A minimal parse tree thus might better be thought of as a parallel trace of parser execution.

As in Subsection 1.2.2, we define mutually inductive type families of `MinParseTreeOf`s and `MinItemsTreeOf`s for a given grammar. Because our parser proceeds by well-founded recursion on the length of the string and the list of nonterminals not yet attempted for that string, we must include both of these in the types. Let us call the initial list of all nonterminals $\texttt{unseen}_0$.

$$\begin{aligned}\texttt{MinParseTreeOf} &: \texttt{String} \to \texttt{[Nonterminal]} \\ &\to \texttt{Item} \to \texttt{String} \to \textbf{Type} \\ \texttt{MinItemsTreeOf} &: \texttt{String} \to \texttt{[Nonterminal]} \\ &\to \texttt{[Item]} \to \texttt{String} \to \textbf{Type}\end{aligned}$$

Much as in the case of parse trees, for any terminal character `ch`, any string $\texttt{s}_0$, and any list of nonterminals `unseen`, we have the constructor

$$\texttt{min\_parse}_{\texttt{'ch'}} : \texttt{MinParseTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ \texttt{'ch'}\ \ \texttt{"ch"}$$

For any production `rule` mapping a nonterminal `nt` to a sequence of items `its`, any string $\texttt{s}_0$, any list of nonterminals `unseen`, and any string `s`, we have two constructors, corresponding to the two ways of progressing with respect to the well-founded relation. Letting $\texttt{unseen}' := \texttt{unseen} - \{\texttt{nt}\}$, we have the following, where we interpret the $<$ relation on strings in terms of lengths.

$$\begin{aligned}(\texttt{rule})_< &: \texttt{s} < \texttt{s}_0 \\ &\to \texttt{MinItemsTreeOf}\ \ \texttt{s}\ \ \texttt{unseen}_0\ \ \texttt{its}\ \ \texttt{s} \\ &\to \texttt{MinParseTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ \texttt{nt}\ \ \texttt{s} \\ (\texttt{rule})_= &: \texttt{s} = \texttt{s}_0 \\ &\to \texttt{nt} \in \texttt{unseen} \\ &\to \texttt{MinItemsTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}'\ \ \texttt{its}\ \ \texttt{s} \\ &\to \texttt{MinParseTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ \texttt{nt}\ \ \texttt{s}\end{aligned}$$

In the first case, the length of the string has decreased, so we may reset the list of not-yet-seen nonterminals, as long as we reset the base of well-founded recursion $\texttt{s}_0$ at the same time. In the second case, the length of the string has not decreased, so we require that we have not yet seen this nonterminal, and we then remove it from the list of not-yet-seen nonterminals.

Finally, for any string $\texttt{s}_0$ and any list of nonterminals `unseen`, we have the following two constructors of `MinItemsTreeOf`.

$$
\begin{aligned}
\texttt{min\_parse}_{[]} &: \texttt{MinItemsTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ \texttt{[]}\ \ \texttt{""} \\
\texttt{min\_parse}_{::} &: \texttt{s}_1\texttt{s}_2 \leq \texttt{s}_0 \\
&\to \texttt{MinParseTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ \texttt{it}\ \ \texttt{s}_1 \\
&\to \texttt{MinItemsTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ \texttt{its}\ \ \texttt{s}_2 \\
&\to \texttt{MinItemsTreeOf}\ \ \texttt{s}_0\ \ \texttt{unseen}\ \ (\texttt{it} :: \texttt{its})\ \ \texttt{s}_1\texttt{s}_2
\end{aligned}
$$

The requirement that $\texttt{s}_1\texttt{s}_2 \leq \texttt{s}_0$ in the second case ensures that we are only making well-founded recursive calls.

Once again, for brevity, we will sometimes use the notation $\overline{\texttt{s} \in \texttt{X}}^{<(\texttt{s}_0,\texttt{v})}$ to denote both `MinParseTreeOf` $\texttt{s}_0$ $\texttt{v}$ $\texttt{X}$ $\texttt{s}$ and `MinItemsTreeOf` $\texttt{s}_0$ $\texttt{v}$ $\texttt{X}$ $\texttt{s}$, relying on context to disambiguate based on the type of $\texttt{X}$. Additionally, we will sometimes fold the constructors of `MinItemsTreeOf` into the two (`rule`) constructors of `MinParseTreeOf`, to mimic the natural-deduction trees.

4.3 Parser Interface

Roughly speaking, we read the interface of our general parser off from the types of the constructors for minimal parse trees. Every constructor leads to one parameter passed to the parser, much as one derives the types of general "fold" functions for arbitrary inductive datatypes. For instance, lists have constructors `nil` and `cons`, so a fold function for lists has arguments corresponding to `nil` (initial accumulator) and `cons` (step function). The situation for the type of our parser is similar, though we need parallel success (managed to parse the string) and failure (could prove that no parse is possible) parameters for each constructor of minimal parse trees.

The type signatures in the interface are presented in Figure 4-1. We explain each type one by one, presenting various instantiations as examples. Note that the interface we actually implemented is also parameterized over a type of `String`s, which we will instantiate with parse trees later in this paper. The interface we present here fixes `String`, for conciseness.

Since we want to be able to specialize our parser to return either `Bool` or `option ParseTree`, we want to be able to reuse our soundness and completeness proofs for both. Our strategy for generalization is to parameterize on dependent type families for "success" and "failure", so we can use relational parametricity to ensure that all instantiations of the parser succeed or fail together. The parser has the rough type signature

$$\texttt{parse} : \texttt{Nonterminal} \ \to\ \texttt{String} \ \to\ \texttt{T}_{\texttt{success}} + \texttt{T}_{\texttt{failure}}.$$

We use ParseQuery to denote the type of all propositions like “"a" ∈ 'a'”; a query consists of a string and a grammar rule the string might be parsed into. We use the same notation for ParseQuery and ParseTree inhabitants. All *_success and *_failure type signatures are implicitly parameterized over a string $\mathtt{s}_0$ and a list of nonterminals unseen. We assume we are given $\mathtt{unseen}_0 : \mathtt{[Nonterminal]}$.

$$\mathtt{T}_{\mathtt{success}},\ \mathtt{T}_{\mathtt{failure}} : \mathtt{String} \to \mathtt{[Nonterminal]} \to \mathtt{ParseQuery} \to \mathbf{Type}$$
$$\mathtt{split} : \mathtt{String} \to \mathtt{[Nonterminal]} \to \mathtt{ParseQuery} \to [\mathbb{N}]$$

$$\begin{aligned}
\mathtt{terminal_success} &: (\mathtt{ch} : \mathtt{Char}) \to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\texttt{"ch"} \in \texttt{'ch'}) \\
\mathtt{terminal_failure} &: (\mathtt{ch} : \mathtt{Char}) \to (\mathtt{s} : \mathtt{String}) \to \mathtt{s} \neq \texttt{"ch"} \to \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \texttt{'ch'}) \\
\mathtt{nil_success} &: \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\texttt{""} \in \epsilon) \\
\mathtt{nil_failure} &: (\mathtt{s} : \mathtt{String}) \to \mathtt{s} \neq \texttt{""} \to \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \epsilon)
\end{aligned}$$

$$\begin{aligned}
\mathtt{cons_success} &: (\mathtt{it} : \mathtt{Item}) \to (\mathtt{its} : \mathtt{[Item]}) \to (\mathtt{s}_1 : \mathtt{String}) \to (\mathtt{s}_2 : \mathtt{String}) \\
&\to \mathtt{s}_1\mathtt{s}_2 \leq \mathtt{s}_0 \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s}_1 \in \mathtt{it}) \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s}_2 \in \mathtt{its}) \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s}_1\mathtt{s}_2 \in \mathtt{it} :: \mathtt{its}) \\
\mathtt{cons_failure} &: (\mathtt{it} : \mathtt{Item}) \to (\mathtt{its} : \mathtt{[Item]}) \to (\mathtt{s} : \mathtt{String}) \\
&\to \mathtt{s} \leq \mathtt{s}_0 \\
&\to \Big(\forall\ (\mathtt{s}_1, \mathtt{s}_2) \in \mathtt{split}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \mathtt{it} :: \mathtt{its}), \\
&\qquad \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s}_1 \in \mathtt{it}) + \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s}_2 \in \mathtt{its})\Big) \\
&\to \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \mathtt{it} :: \mathtt{its})
\end{aligned}$$

$$\begin{aligned}
\mathtt{production_success}_{<} &: (\mathtt{its} : \mathtt{[Item]}) \to (\mathtt{nt} : \mathtt{Nonterminal}) \to (\mathtt{s} : \mathtt{String}) \\
&\to \mathtt{s} < \mathtt{s}_0 \\
&\to (\mathtt{p} : \text{a production mapping } \mathtt{nt} \text{ to } \mathtt{its}) \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}\ \ \mathtt{unseen}_0\ \ (\mathtt{s} \in \mathtt{its}) \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \mathtt{nt}) \\
\mathtt{production_success}_{=} &: (\mathtt{its} : \mathtt{[Item]}) \to (\mathtt{nt} : \mathtt{Nonterminal}) \to (\mathtt{s} : \mathtt{String}) \\
&\to \mathtt{nt} \in \mathtt{unseen} \\
&\to (\mathtt{p} : \text{a production mapping } \mathtt{nt} \text{ to } \mathtt{its}) \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ (\mathtt{unseen} - \{\mathtt{nt}\})\ \ (\mathtt{s} \in \mathtt{its}) \\
&\to \mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \mathtt{nt}) \\
\mathtt{production_failure}_{<} &: (\mathtt{nt} : \mathtt{Nonterminal}) \to (\mathtt{s} : \mathtt{String}) \\
&\to \mathtt{s} < \mathtt{s}_0 \\
&\to \Big(\forall\ (\mathtt{its} : \mathtt{[Item]})\ (\mathtt{p} : \text{a production mapping } \mathtt{nt} \text{ to } \mathtt{its}),\ \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}\ \ \mathtt{unsee} \\
&\to \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{s} \in \mathtt{nt}) \\
\mathtt{production_failure}_{=} &: (\mathtt{nt} : \mathtt{Nonterminal}) \to (\mathtt{s} : \mathtt{String}) \\
&\to \mathtt{s} = \mathtt{s}_0 \\
&\to \Big(\forall\ (\mathtt{its} : \mathtt{[Item]})\ (\mathtt{p} : \text{a production mapping } \mathtt{nt} \text{ to } \mathtt{its}),\ \mathtt{T}_{\mathtt{failure}}\ \ \mathtt{s}_0\ \ (\mathtt{unse}
\end{aligned}$$

To instantiate the parser as a Boolean recognizer, we instantiate everything trivially; we use the fact that $\top + \top \cong$ `Bool`. Just to show how trivial everything is, here is a precise instantiation of the parser, still parameterized over the initial list of nonterminals and the splitter, where $\top$ is the one constructor of the one-element type $\top$:

$$\begin{aligned}
&\mathtt{T}_{\mathtt{success}}\ _\ _\ _ := \top\\
&\mathtt{T}_{\mathtt{failure}}\ _\ _\ _ := \top
\end{aligned}$$

$$\begin{aligned}
&\mathtt{terminal_success}\ _\ _\ _ := ()\\
&\mathtt{terminal_failure}\ _\ _\ _\ _\ _ := ()\\
&\mathtt{nil_success}\ _\ _ := ()\\
&\mathtt{nil_failure}\ _\ _\ _\ _ := ()\\
&\mathtt{cons_success}\ _\ _\ _\ _\ _\ _\ _\ _\ _\ _ := ()\\
&\mathtt{cons_failure}\ _\ _\ _\ _\ _\ _\ _\ _ := ()\\
&\mathtt{production_success}_{<}\ _\ _\ _\ _\ _\ _\ _\ _\ _ := ()\\
&\mathtt{production_success}_{=}\ _\ _\ _\ _\ _\ _\ _\ _\ _\ _ := ()\\
&\mathtt{production_failure}_{<}\ _\ _\ _\ _\ _\ _ := ()\\
&\mathtt{production_failure}_{=}\ _\ _\ _\ _\ _\ _ := ()\\
&\mathtt{production_failure}_{\notin}\ _\ _\ _\ _\ _\ _ := ()
\end{aligned}$$

To instantiate our parser so that it returns `option ParseTree` (rather, the dependently typed flavor, `ParseTreeOf`), we take advantage of the isomorphism $T + \top \cong$ `option` T. We show only the `success` instantiations, as the `failure` ones are identical with the Boolean recognizer. For readability of the code, we write schematic natural-deduction proof trees inline.

$$\mathtt{T}_{\mathtt{success}}\ \_\ \_\ \ (\overline{\mathtt{s} \in \mathtt{X}}) \coloneqq \overline{\mathtt{s} \in \mathtt{X}}$$

$$\begin{aligned}
&\mathtt{terminal_success}\ _\ _\ \mathtt{ch} \coloneqq (\mathtt{'ch'}) \\
&\mathtt{nil_success}\ _\ _ \coloneqq \overline{\mathtt{""} \in \epsilon} \\
&\mathtt{cons_success}\ _\ _\ \mathtt{it}\ \mathtt{its}\ \mathtt{s}_1\ \mathtt{s}_2\ _\ \mathtt{d}_1\ \mathtt{d}_2 \coloneqq \frac{\frac{\mathtt{d}_1}{\mathtt{s}_1 \in \mathtt{it}} \qquad \frac{\mathtt{d}_2}{\mathtt{s}_2 \in \mathtt{its}}}{\mathtt{s}_1\mathtt{s}_2 \in \mathtt{it} :: \mathtt{its}} \\
&\mathtt{production_success}_{<}\ _\ _\ \mathtt{it}\ \mathtt{nt}\ \mathtt{s}\ _\ \mathtt{p}\ \mathtt{d} \coloneqq \frac{\frac{\mathtt{d}}{\mathtt{s} \in \mathtt{its}}}{\mathtt{s} \in \mathtt{nt}}\ {\scriptstyle(\mathtt{p})} \\
&\mathtt{production_success}_{=}\ _\ _\ \mathtt{it}\ \mathtt{nt}\ \mathtt{s}\ _\ \mathtt{p}\ \mathtt{d} \coloneqq \frac{\frac{\mathtt{d}}{\mathtt{s} \in \mathtt{its}}}{\mathtt{s} \in \mathtt{nt}}\ {\scriptstyle(\mathtt{p})}
\end{aligned}$$

What remains is to instantiate the parser in such a way that proving completeness is trivial. The simpler of our two tasks is to show that when the parser fails, no minimal parse tree exists. Hence we instantiate the types as follows, where $\perp$ is the empty type (equivalently, the false proposition).

$$\begin{aligned}
&\mathtt{T}_{\mathtt{success}}\ _\ _\ _ \coloneqq \top \\
&\mathtt{T}_{\mathtt{failure}}\ \mathtt{s}_0\ \mathtt{unseen}\ \ (\overline{\mathtt{s} \in \mathtt{X}}) \coloneqq \left(\overline{\mathtt{s} \in \mathtt{X}}^{\,<(\mathtt{s}_0,\mathtt{unseen})}\right) \to \perp
\end{aligned}$$

Using $\lightning$ to denote deriving a contradiction, we can unenlighteningly instantiate the arguments as

$$\begin{aligned}
&\mathtt{terminal_success}\ _\ _\ _ \coloneqq () \\
&\mathtt{terminal_failure}\ _\ _\ _\ _\ _ \coloneqq \lightning \\
&\mathtt{nil_success}\ _\ _ \coloneqq () \\
&\mathtt{nil_failure}\ _\ _\ _\ _ \coloneqq \lightning \\
&\mathtt{cons_success}\ _\ _\ _\ _\ _\ _\ _\ _\ _\ _ \coloneqq () \\
&\mathtt{cons_failure}\ _\ _\ _\ _\ _\ _\ _ \coloneqq \lightning \\
&\mathtt{production_success}_{<}\ _\ _\ _\ _\ _\ _\ _\ _ \coloneqq () \\
&\mathtt{production_success}_{=}\ _\ _\ _\ _\ _\ _\ _\ _\ _ \coloneqq () \\
&\mathtt{production_failure}_{<}\ _\ _\ _\ _\ _\ _ \coloneqq \lightning \\
&\mathtt{production_failure}_{=}\ _\ _\ _\ _\ _\ _ \coloneqq \lightning \\
&\mathtt{production_failure}_{\notin}\ _\ _\ _\ _\ _\ _ \coloneqq \lightning
\end{aligned}$$

A careful inspection of the proofy arguments to each `failure` case will reveal that there is enough evidence to derive the appropriate contradiction. For example, the $\texttt{s} \neq \texttt{""}$ hypothesis of `nil_failure` contradicts the equalities implied by the type signature of $\texttt{min\_parse}_{\texttt{[]}}$, and the use of `[]` contradicts the equality implied by the use of $\texttt{it}::\texttt{its}$ in the type signature of $\texttt{min\_parse}_{\texttt{[]}}$. Similarly, the $\texttt{s} \neq \texttt{"ch"}$ hypothesis of `terminal_failure` contradicts the equality implied by the usage of the single identifier `ch` in two different places in the type signature of $\texttt{min\_parse}_{\texttt{'ch'}}$.

4.3.1 Parsing Parses

We finally come to the most twisty part of the parser: parsing parse trees. Recall that our parser definition is polymorphic in a choice of `String` type. We proceed with the straw-man solution of literally passing in parse trees as strings to be parsed, such that parsing generates *minimal* parse trees, as introduced in Section 4.1 and defined formally in Section 4.2. Intuitively, we run a top-down traversal of the tree, pausing at each node before descending to its children. During that pause, we *eliminate one level of wastefulness*: if the parse tree is proving $\texttt{s} \in \texttt{X}$, we look for any subtrees also proving $\texttt{s} \in \texttt{X}$. If we find any, we replace the original tree with *the smallest duplicative subtree*. If we do not find any, we leave the tree unchanged. In either case, we then descend into "parsing" each subtree.

We define a function `deloop` to perform the one step of eliminating waste:

$$\texttt{deloop} : \texttt{ParseTreeOf nt s} \to \texttt{ParseTreeOf nt s}$$

This transformation is straightforward to define by structural recursion.

To implement all of the generic parameters of the parser, we must actually augment the result type of `deloop` with stronger types. Define the predicate $\texttt{Unloopy}(t)$ on parse trees t to mean that, where the root node of t proves $\texttt{s} \in \texttt{nt}$, for every subtree proving $\texttt{s} \in \texttt{nt}'$ (same string, possibly different nonterminal), (1) $\texttt{nt}'$ is in the set of allowed nonterminals, `unseen`, associated to the overall tree with dependent types, and (2) if this is not the root node, then $\texttt{nt}' \neq \texttt{nt}$.

We augment the return type of `deloop`, writing:

$$\{t : \texttt{ParseTreeOf nt s} \mid \texttt{Unloopy}(t)\}.$$

We instantiate the generic "string" type parameter of the general parser with this type family, so that, in implementing the different parameters to pass to the parser, we have the property available to us.

Another key ingredient is the "string" splitter, which naturally breaks a parse tree into its child trees. We define it like so:

$$\begin{array}{l}
\mathtt{split}\ _\ _\ (\overline{\mathtt{s} \in \mathtt{it} :: \mathtt{its}}) \coloneqq \\
\quad \mathbf{case}\ \ \mathtt{parse_tree_data}\ \mathtt{s}\ \ \mathbf{of} \\
\quad \left|\ \dfrac{\dfrac{\mathtt{p}_1}{\mathtt{s}_1 \in \mathtt{it}} \qquad \dfrac{\mathtt{p}_2}{\mathtt{s}_2 \in \mathtt{its}}}{\mathtt{s}_1\mathtt{s}_2 \in \mathtt{it} :: \mathtt{its}} \ \rightarrow\ \left[(\mathtt{deloop}\ \mathtt{p}_1, \mathtt{deloop}\ \mathtt{p}_2)\right]\right. \\
\quad \left|\ _ \ \rightarrow\ \lightning\right. \\
\mathtt{split}\ _\ _\ _ \coloneqq []
\end{array}$$

Note that we use `it` and `its` nonlinearly; the pattern only binds if its `it` and `its` match those passed as arguments to `split`. We thus return a nonempty list only if the query is about a nonempty sequence of items. Because we use dependent types to enforce the requirement that the parse tree associated with a string match the query we are considering, we can derive contradictions in the non-matching cases.

This splitter satisfies two important properties. First, it never returns the empty list on a parse tree whose list of productions is nonempty; call this property *nonempty preservation*. Second, it preserves `Unloopy`. We use both facts in the other parameters to the generic parser (and we leave their proofs as exercises for the reader—Coq solutions may be found in our source code).

Now recall that our general parser always returns a type of the form $\mathtt{T}_{\mathtt{success}} + \mathtt{T}_{\mathtt{failure}}$, for some $\mathtt{T}_{\mathtt{success}}$ and $\mathtt{T}_{\mathtt{failure}}$. We want our tree minimizer to return just the type of minimal trees. However, we can take advantage of the type isomorphism $T + \bot \cong T$ and instantiate $\mathtt{T}_{\mathtt{failure}}$ with $\bot$, the uninhabited type; and then apply a simple fix-up wrapper on top. Thus, we instantiate the general parser like so:

$$\begin{array}{l}
\mathtt{T}_{\mathtt{success}}\ \ \mathtt{s}_0\ \ \mathtt{unseen}\ \ (\mathtt{d} : \overline{\mathtt{s} \in \mathtt{X}}) \coloneqq \overline{\mathtt{s} \in \mathtt{X}}^{\ <(\mathtt{s}_0, \mathtt{unseen})} \\
\mathtt{T}_{\mathtt{failure}}\ \ _\ \ _\ \ _ \coloneqq \bot
\end{array}$$

The `success` cases are instantiated in an essentially identical way to the instantiation we used to get `option ParseTree`. The `terminal_failure` and `nil_failure` cases provide enough information ($\mathtt{s} \neq$ "ch" and $\mathtt{s} \neq$ "", respectively) to derive $\bot$ from the existence of the appropriately typed parse tree. In the `cons_failure` case, we make use of the splitter's *nonempty preservation* behavior, after which all that remains is $\bot + \bot \rightarrow \bot$, which is trivial. In the $\mathtt{production\_failure}_<$ and $\mathtt{production\_failure}_=$ cases, it is sufficient to note that every nonterminal is mapped by some production to some sequence of items. Finally, to instantiate the $\mathtt{production\_failure}_{\notin}$ case, we need to appeal to the `Unloopy`-ness of the tree to deduce that $\mathtt{nt} \in \mathtt{unseen}$. Then we can derive $\bot$ from the hypothesis that $\mathtt{nt} \notin \mathtt{unseen}$, and we are done.

We instantiate the general parser with an input type that requires `Unloopy`, so our final tree minimizer is really the composition of the instantiated parser with `deloop`, ensuring that invariant as we kick off the recursion.

4.3.2 Example

In Subsection 1.1.1, we defined an ambiguous grammar for $\texttt{(ab)}^*$ which led our naive parser to diverge. We will walk through the minimization of the following parse tree of `"abab"` into this grammar. For reference, Figure 4-2 contains the fully general implementation of our parser, modulo type signatures.

For reasons of space, define $\overline{T}$ to be the parse tree

$$\dfrac{\dfrac{\dfrac{}{\texttt{""} \in \epsilon}}{\texttt{""} \in \texttt{(ab)}^*} \qquad \dfrac{\dfrac{\dfrac{}{\texttt{"a"} \in \texttt{'a'}} \quad \dfrac{}{\texttt{"b"} \in \texttt{'b'}}}{\texttt{"a"} \cdot \texttt{"b"} \in \texttt{(ab)}^*}}{\texttt{"ab"} \in \texttt{(ab)}^*}}{\texttt{""} \cdot \texttt{"ab"} \in \texttt{(ab)}^*}\,(\texttt{(ab)}^*\texttt{(ab)}^*)$$

Then we consider minimizing the parse tree:

$$\dfrac{\dfrac{\dfrac{\overline{T}}{\texttt{"ab"} \in \texttt{(ab)}^*} \qquad \dfrac{\overline{T}}{\texttt{"ab"} \in \texttt{(ab)}^*}}{\texttt{"ab"} \cdot \texttt{"ab"} \in \texttt{(ab)}^*}\,(\texttt{(ab)}^*\texttt{(ab)}^*)}{\texttt{"abab"} \in \texttt{(ab)}^*}$$

Letting $\overline{T'_m}$ denote the same tree as $\overline{T'}$, but constructed as a `MinParseTree` rather than a `ParseTree`, the tree we will end up with is:

$$\dfrac{\dfrac{\dfrac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]}) \qquad \dfrac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]})}{\texttt{"ab"} \cdot \texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"abab"}, \texttt{[]})}{\texttt{"abab"} \in \texttt{(ab)}^*} < (\texttt{"abab"}, \texttt{[(ab)}^*\texttt{]})$$

To begin, we call `parse`, passing in the entire tree as the string, and $\texttt{(ab)}^*$ as the nonterminal. To transform the tree into one that satisfies `Unloopy`, the first thing `parse` does is call `deloop` on our tree. In this case, `deloop` is a no-op; it promotes the deepest non-root nodes labeled with ($\texttt{"abab"} \in \texttt{(ab)}^*$), of which there are none.

We then take the following execution steps, starting with $\texttt{unseen} := \texttt{unseen}_0 := \texttt{[(ab)}^*\texttt{]}$, the singleton list containing the only nonterminal, and $\texttt{s}_0 := \texttt{"abab"}$.

1. We first ensure that we are not in an infinite loop. We check if $\texttt{s} < \texttt{s}_0$ (it is not, for they are both equal to `"abab"`), and then check if our current nonterminal, $\texttt{(ab)}^*$, is in `unseen`. Since the second check succeeds, we remove $\texttt{(ab)}^*$ from `unseen`; calls made by this stack frame will pass `[]` for `unseen`.

2. We may consider only the productions for which the parse tree associated to

```
parse nt s := parse′ (s_0 := s) (unseen := unseen_0) (s ∈ nt)

parse′ ("ch" ∈ 'ch') := inl terminal_success
parse′ (_ ∈ 'ch') := inr (terminal_failure ↯)
parse′ ("" ∈ ϵ) := inl nil_success
parse′ (_ ∈ ϵ) := inr (nil_failure ↯)
parse′ (s ∈ it :: its) :=
    case any_parse s it its (split (s ∈ it :: its)) of
     | inl ret → inl ret
     | inr ret → inr (cons_failure _ ret)
parse′ (s ∈ nt) :=
    if s < s_0
    then if (parse′ (s_0 := s) (unseen := unseen_0) (s ∈ its)) succeeds returning d
                          for any production p mapping nt to its
         then inl (production_success_< _ p d)
         else inr (production_failure_< _ _)
    else if nt ∈ unseen
         then if (parse′ (unseen := unseen − {nt}) (s ∈ its)) succeeds returning d
                          for any production p mapping nt to its
              then inl (production_success_= _ p d)
              else inr (production_failure_= _ _)
         else inr (production_failure_∉ _ _)

any_parse s it its [] := inr (λ _ : (_ ∈ []). ↯)
any_parse s it its (x :: xs) :=
    case parse′ (take_x s ∈ it), parse′ (drop_x s ∈ its), any_parse s it its xs of
     | inl ret_1, inl ret_2, _       → inl (cons_success _ ret_1 ret_2)
     | _        , _        , inl ret′ → inl ret′
     | ret_1    , ret_2    , inr ret′ → inr _
```

where the hole on the last line constructs a proof of

$$\forall\, \mathtt{x}' \in (\mathtt{x} :: \mathtt{xs}),\ \ \mathtt{T}_{\mathtt{failure}}\ \_\ \_\ (\mathtt{take}_{\mathtt{x}'}\ \mathtt{s} \in \mathtt{it}) + \mathtt{T}_{\mathtt{failure}}\ \_\ \_\ (\mathtt{drop}_{\mathtt{x}'}\ \mathtt{s} \in \mathtt{its})$$

by using $\mathtt{ret}'$ directly when $\mathtt{x}' \in \mathtt{xs}$, and using whichever one of $\mathtt{ret}_1$ and $\mathtt{ret}_2$ is on the right when $\mathtt{x}' = \mathtt{x}$. While straightforward, the use of sum types makes it painfully verbose without actually adding any insight; we prefer to elide the actual term.

Figure 4-2: Pseudo-Implementation of our parser. We take the convention that dependent indices to functions (e.g., `unseen`) are implicit.

the string is well-typed; we will describe the headaches this seemingly innocuous simplification caused us in Subsection 4.4.2. The only such production in this case is the one that lines up with the production used in the parse tree, labeled $\texttt{(ab)}^*\texttt{(ab)}^*$.

3. We invoke `split` on our parse tree.

   (a) The `split` that we defined then invokes `deloop` on the two copies of the parse tree

   $$\frac{\overline{T}}{\texttt{"ab"} \in \texttt{(ab)}^*}$$

   Since there are non-root nodes labeled with ($\texttt{"ab"} \in \texttt{(ab)}^*$), the label of the root node, we promote the deepest one. Letting T' denote the tree

   $$\frac{\overline{\texttt{"a"} \in \texttt{'a'}} \qquad \overline{\texttt{"b"} \in \texttt{'b'}}}{\texttt{"a"} \cdot \texttt{"b"} \in \texttt{(ab)}^*}\ (\texttt{"ab"})$$

   the result of calling `deloop` is the tree

   $$\frac{\overline{T'}}{\texttt{"ab"} \in \texttt{(ab)}^*}$$

   (b) The return of `split` is thus the singleton list containing a single pair of two parse trees; each element of the pair is the parse tree for $\texttt{"ab"} \in \texttt{(ab)}^*$ that was returned by `deloop`.

4. We invoke `parse` on each of the items in the sequence of items associated to $\texttt{(ab)}^*$ via the rule ($\texttt{(ab)}^*\texttt{(ab)}^*$). The two items are identical, and their associated elements of the pair returned by `split` are identical, so we only describe the execution once, on

   $$\frac{\overline{T'}}{\texttt{"ab"} \in \texttt{(ab)}^*}$$

   (a) We first ensure that we are not in an infinite loop. We check if $\texttt{s} < \texttt{s}_0$. This check succeeds, for `"ab"` is shorter than `"abab"`. We thus reset `unseen` and $\texttt{s}_0$; calls made by this stack frame will pass $\texttt{unseen}_0 \equiv \texttt{[(ab)}^*\texttt{]}$ for `unseen`, and $\texttt{s} \equiv \texttt{"ab"}$ for $\texttt{s}_0$.

   (b) We may again consider only the productions for which the parse tree associated to the string is well-typed. The only such production in this case is the one that lines up with the production used in the parse tree T', labeled (`"ab"`).

   (c) We invoke `split` on our parse tree.

      i. The `split` that we defined then invokes `deloop` on the trees $\overline{\texttt{"a"} \in \texttt{'a'}}$ and $\overline{\texttt{"b"} \in \texttt{'b'}}$. Since these trees have no non-root nodes (let alone non-root nodes sharing a label with the root), `deloop` is a no-op.

ii. The return of `split` is thus the singleton list containing a single pair of two parse trees; the first is the parse tree $\overline{\texttt{"a"} \in \texttt{'a'}}$, and the second is the parse tree $\overline{\texttt{"b"} \in \texttt{'b'}}$.

(d) We invoke `parse` on each of the items in the sequence of items associated to $\texttt{(ab)}^*$ via the rule ($\texttt{"ab"}$). Since both of these items are terminals, and the relevant equality check (that "a" is equal to "a", and similarly for "b") succeeds, `parse` returns `terminal_success`. We thus have the two `MinParseTrees`: $\overline{\texttt{"a"} \in \texttt{'a'}}$ and $\overline{\texttt{"b"} \in \texttt{'b'}}$.

(e) We combine these using `cons_success` (and `nil_success`, to tie up the base case of the list). We thus have the tree $\overline{T'_m}$.

(f) We apply $\texttt{production\_success}_<$ to this tree, and return the tree

$$\frac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]})$$

5. We now combine the two identical trees returned by `parse` using `cons_success` (and `nil_success`, to tie up the base case of the list). We thus have the tree

$$\frac{\dfrac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]}) \qquad \dfrac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]})}{\texttt{"ab"} \cdot \texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"abab"}, \texttt{[]})$$

6. We apply $\texttt{production\_success}_=$ to this tree, and return the tree we claimed we would end up with,

$$\frac{\dfrac{\dfrac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]}) \qquad \dfrac{\overline{T'_m}}{\texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"ab"}, \texttt{[(ab)}^*\texttt{]})}{\texttt{"ab"} \cdot \texttt{"ab"} \in \texttt{(ab)}^*} < (\texttt{"abab"}, \texttt{[]})}{\texttt{"abab"} \in \texttt{(ab)}^*} < (\texttt{"abab"}, \texttt{[(ab)}^*\texttt{]})$$

4.3.3 Parametricity

Before we can combine different instantiations of this interface, we need to know that they behave similarly. Inspection of the code, together with relational parametricity, validates assuming the following axiom, which should also be internally provable by straightforward induction (though we have not bothered to prove it).

The *parser extensionality axiom* states that, for any fixed instantiation of `split`, and any arbitrary instantiations of the rest of the interface, giving rise to two different functions $\texttt{parse}_1$ and $\texttt{parse}_2$, we have

$$\begin{aligned}&\forall\ (\texttt{nt} : \texttt{Nonterminal})\ (\texttt{s} : \texttt{String}),\\ &\quad \texttt{bool\_of\_sum}\ (\texttt{parse}_1\ \texttt{nt}\ \texttt{s}) = \texttt{bool\_of\_sum}\ (\texttt{parse}_2\ \texttt{nt}\ \texttt{s})\end{aligned}$$

where `bool_of_sum` is, for any types A and B, the function of type $A + B \to \texttt{Bool}$

obtained by sending everything in the left component to `true`, and everything in the right component to `false`.

4.3.4 Putting It All Together

Now we have parsers returning the following types:

$$
\begin{aligned}
\texttt{has\_parse} &: \texttt{Nonterminal} \to \texttt{String} \to \texttt{Bool} \\
\texttt{parse} &: (\texttt{nt} : \texttt{Nonterminal}) \to (\texttt{s} : \texttt{String}) \\
&\to \texttt{option}\ (\texttt{ParseTreeOf}\ \texttt{nt}\ \texttt{s}) \\
\texttt{has\_no\_parse} &: (\texttt{nt} : \texttt{Nonterminal}) \to (\texttt{s} : \texttt{String}) \\
&\to \top + (\texttt{MinParseTreeOf}\ \texttt{nt}\ \texttt{s} \to \bot) \\
\texttt{min\_parse} &: (\texttt{nt} : \texttt{Nonterminal}) \to (\texttt{s} : \texttt{String}) \\
&\to \texttt{ParseTreeOf}\ \texttt{nt}\ \texttt{s} \\
&\to \texttt{MinParseTreeOf}\ \texttt{nt}\ \texttt{s}
\end{aligned}
$$

Note that we have taken advantage of the isomorphism $\top + \top \cong$ `Bool` for `has_parse`, the isomorphism $A + \top \cong$ `option` A for `parse`, and the isomorphism $A + \bot \cong A$ for `min_parse`.

We can compose these functions to obtain our desired correct-by-construction parser:

$$
\begin{aligned}
&\texttt{parse\_full} : (\texttt{nt} : \texttt{Nonterminal}) \to (\texttt{s} : \texttt{String}) \\
&\qquad\qquad \to \texttt{ParseTreeOf}\ \texttt{nt}\ \texttt{s} + (\texttt{ParseTreeOf}\ \texttt{nt}\ \texttt{s} \to \bot) \\
&\texttt{parse\_full}\ \texttt{nt}\ \texttt{s} \coloneqq \\
&\quad \textbf{case}\ \ \texttt{parse}\ \texttt{nt}\ \texttt{s},\ \texttt{has\_no\_parse}\ \texttt{nt}\ \texttt{s}\ \ \textbf{of} \\
&\quad \mid \texttt{Some}\ \texttt{d},\ \_ \qquad\quad \to\ \texttt{inl}\ \texttt{d} \\
&\quad \mid \_ \qquad\ ,\ \texttt{inr}\ \texttt{nd}\ \to\ \texttt{inr}\ (\texttt{nd} \circ \texttt{min\_parse}) \\
&\quad \mid \_ \qquad\ ,\ \_ \qquad\quad \to\ \lightning
\end{aligned}
$$

In the final case, we derive a contradiction by applying the parser extensionality axiom, which says that `parse` and `has_no_parse` must agree on whether or not `s` parses as `nt`.

4.4 Missteps, Insights, and Dependently Typed Lessons

We will now take a step back from the parser itself, and briefly talk about the process of coding it. We encountered a few pitfalls that we think highlight some key aspects of dependently typed programming, and our successes suggest benefits to be reaped from using dependent types.

4.4.1 The Trouble of Choosing the Right Types

Although we began by attempting to write the type-signature of our parser, we found that trying to write down the correct interface, without any code to implement it, was essentially intractable. Giving your functions dependent types requires performing a nimble balancing act between being uselessly general on the one hand, and too overly specific on the other, all without falling from the highropes of well-typedness onto the unforgiving floor of type errors.

We have found what we believe to be the worst sin the typechecker will let you get away with: having different levels of generality in different parts of your code base, which are supposed to interface with each other without a thoroughly vetted abstraction barrier between them. Like setting your highropes at different tensions, every trip across the interface will be costly, and if the abstraction levels get too far away, recovering your balance will require Herculean effort.

We eventually gave up on writing a dependently typed interface from the start, and decided instead to implement a simply typed Boolean recognizer, together with proofs of soundness and completeness. Once we had in hand these proofs, and the data types required to carry them out, we found that it was mostly straightforward to write down the interface and refine our parser to inhabit its newly generalized type.

4.4.2 Misordered Splitters

One of our goals in this presentation was to hide most of the abstraction-level mismatch that ended up in our actual implementation, often through clever use of notation overloading. One of the most significant mismatches we managed to overcome was the way to represent the set of productions. In this paper, we left the type as an abstract mathematical set, allowing us to forgo concerns about ordering, quantification, and occasionally well-typedness.

In our Coq implementation, we fixed the type of productions to be a list very early on, and paid the price when we implemented our parse-tree parser. As mentioned in the execution of the example in Subsection 4.3.2, we wanted to restrict our attention to certain productions, and rule out the other ones using dependent types. This should be possible if we parameterize over not just a splitter, but a production-selector, and only require that our string type be well-typed for productions given by the production-selector. However, the implementation that we currently have requires a well-typed string type for all productions; furthermore, it does not allow the order in which productions are considered to depend on the augmented string data. We paid for this with the extra 300 lines of code we had to write to interleave two different splitters, so that we could handle the cases that we dismissed above as being ill-typed and therefore not necessary to consider. That is, because our types were not formulated in a way that actually made these cases ill-typed, we had to deal with them, much to our displeasure.

4.4.3 Minimal Parse Trees vs. Parallel Traces

Taking another step back, our biggest misstep actually came before we finished the completeness proof for our simply typed Boolean recognizer.

When first constructing the type `MinParseTree`, we thought of them genuinely as minimal parse trees (ones without a duplicate label in any single path). After much head-banging, of knowledge that a theorem was obviously true, against proof goals that were obviously impossible, we discovered the single biggest insight—albeit a technical one—of the project. The type of "minimal parse trees" we had originally formulated did not match the parse trees produced by our algorithm. A careful examination of the algorithm execution in Subsection 4.3.2 should reveal the difference.[1] Our insight, thus, was to conceptualize the data type as the type of traces of parallel executions of our particular parser, rather than as truly minimal parse trees.

This may be an instance of a more general phenomenon present when programming with dependent types: subtle friction between what you think you are doing and what you are actually doing often manifests as impossible proof goals.

[1] For readers wanting to skip that examination: the algorithm we described allows a label ($\mathtt{s} \in \mathtt{nt}$) to appear one extra time along a path if, the first time it appears, its parent node's label, ($\mathtt{s}' \in \mathtt{nt}'$), satisfies $\mathtt{s} < \mathtt{s}'$. That is, whenever the string being parsed shrinks, the first nonterminal the shrunken string is parsed as may be duplicated once before shrinking the string again.

Chapter 5

Refining Splitters by Fiat

5.1 Splitters at a Glance

We have now finished describing the general parsing algorithm, as well as its correctness proofs; we have an algorithm, that decides whether or not a given structure can be imposed on any block of unstructured text. The algorithm is parametrized on an "oracle" that describes how to split the string for each rule; essentially all of the algorithmically interesting content is in the splitters. For the remainder of this paper, we will focus on how to implement the splitting oracle. Correctness is not enough, in general; algorithms also need to be fast to use. We thus focus primarily on efficiency when designing splitting algorithms, and work in a framework that guarantees correctness.

The goals of this work, as mentioned in Section 3.2, are to present a framework for constructing proven-correct parsers incrementally, and argue for its eventual feasibility. To this end, we build on the previous work of Fiat [**fiat**], to allow us to build programs incrementally while maintaining correctness guarantees. This section will describe Fiat, and how it is used in this project. The following sections will focus more on the details of the splitting algorithms, and less on Fiat itself.

5.2 What counts as efficient?

To guide our implementations, we characterize efficient splitters informally, as follows. Although our eventual concrete efficiency target is to be competitive with extant open source JavaScript parsers, when designing algorithms, we aim at the asymptotic efficiency target of linearity in the length of the string. In practice, the dominating concern is that doubling the length of the string should only double the duration of the parse, and not quadruple it (or more!). [**TODO:** CITATION NEEDED] To be efficient, it suffices to have the splitter return at most one index. In this case, the parsing time is $\mathcal{O}$(length of string $\times$ (product over all nonterminals of the number of possible rules for that nonterminal)).

Here is an example of hitting the worst-case scenario. **[TODO:** Is this actually possible?]

To avoid hitting this worst-case scenario, we can use a nonterminal-picker, which returns the list of possible production rules for a given string and nonterminal. As long as it returns at most one possible rule in most cases, in constant time, the parsing time will be $\mathcal{O}$(length of string); backtracking will never happen. This is future work.

5.3 Introducing Fiat

5.3.1 Incremental Construction by Refinement

Efficiency targets in hand, we move on to incremental construction. The key idea is that parsing rules tend to fall into clumps that are similar between grammars. For example, many grammars use delimiters (such as whitespace, commas, or binary operation symbols) as splitting points, but only between well-balanced brackets (such as double quotes, parentheses, or comment markers). We can take advantage of these similarities by baking the relevant algorithms into basic building blocks, which can then be reused across different grammars. To allow this reuse, we construct the splitters incrementally, allowing us to deal with different rules in different ways.

The Fiat framework [**fiat**] is the scaffolding of our splitter implementations. As a framework, the goal of Fiat is to enable library-writers to construct algorithmic building blocks packaged with correctness guarantees, in such a way that users can easily and mostly-automatically make use of these building blocks when they apply.

5.3.2 The Fiat Mindset

The correctness guarantees of Fiat are based on specifications in the form of propositions in Gallina, the mathematical language used by Coq. For example, the specification of a valid `has_parse` method is that `has_parse nt str = true` $\longleftrightarrow$ `inhabited (ParseTreeOf nt s)`. Fiat allows incremental construction of algorithms by providing a language for seamlessly mixing specifications and code. The language is a light-weight monadic syntax with one extra operator: a non-deterministic choice operator; we define the following combinators:

$\mathtt{x} \leftarrow \mathtt{c};\ \ \mathtt{c}'$	Run c and store the result in x; continue with c′, which may mention x
$\mathtt{c};;\ \ \mathtt{c}'$	Run c. If it terminates, throw away the result, and run c′
`ret x`	A computation that immediately returns the value x
$\{\mathtt{x} \mid \mathtt{P}(\mathtt{x})\}$	Nondeterministically choose a value of x satisfying P. If none exists, the program is considered to not terminate.

An algorithm starts out as a nondeterministic choice of a value satisfying the specification. Coding then proceeds by refinement. Formally, we say that a computation c' *refines* a computation c, written $c' \subseteq c$, if every value that c' can compute to, c can also compute to. We freely generate the relation "the computation c can compute to the value v", written $c \rightsquigarrow v$, by the rules:

$$
\begin{array}{rcll}
\texttt{ret}\ v & \rightsquigarrow & v & \\
\{x \mid P(x)\} & \rightsquigarrow & v & \text{iff } x \text{ satisfies } P \\
(c;;\ \ c') & \rightsquigarrow & v & \text{iff there is a } v' \text{ such that } c \rightsquigarrow v' \text{ and } c' \rightsquigarrow v \\
(x \leftarrow c;\ \ c'(x)) & \rightsquigarrow & v & \text{iff there is a } v' \text{ such that } c \rightsquigarrow v' \text{ and } c'(v') \rightsquigarrow v
\end{array}
$$

In our use case, we express the specification of the splitter as a nondeterministic choice of a list of split locations, such that any splitting location that results in a valid parse tree is contained in the list. More formally, we can define the proposition

$$
\begin{array}{l}
\texttt{split\_list\_is\_complete} : \texttt{Grammar} \to \texttt{String} \to \texttt{[Item]} \to [\mathbb{N}] \to \textbf{Prop} \\
\texttt{split\_list\_is\_complete}\ G\ str\ \texttt{[]}\ splits\ = \bot \\
\texttt{split\_list\_is\_complete}\ G\ str\ (it :: its)\ splits \\
\quad = \forall\, n,\ n < \texttt{length}\ str \\
\qquad\quad \to\ (\texttt{has\_parse}\ it\ (\texttt{take}\ n\ str)\ \wedge\ \texttt{has\_parse}\ its\ (\texttt{drop}\ n\ str)) \\
\qquad\quad \to\ n \in splits
\end{array}
$$

where we overload `has_parse` to apply to items and productions alike. In practice, we pass the first item, and the rest of the items, as separate arguments, so that we don't have to deal with the empty list case.

Let `production_is_reachable` $G\ p$ to be the proposition that p could show up during parsing, i.e., that p is a tail of a rule associated to some valid nonterminal in the grammar; we define this by folding over the list of valid nonterminals. The specification of the splitter, as a nondeterministic computation, for a given grammar G, a given string str, and a given rule $it{::}its$, is then:

$$
\begin{array}{l}
\{\ splits\ :\ \texttt{list}\ \mathbb{N} \\
\mid \texttt{production\_is\_reachable}\ G\ (it :: its) \\
\quad \to\ \texttt{split\_list\_is\_complete}\ G\ str\ it\ its\ splits\ \}
\end{array}
$$

We then refine this into a choice of a splitting location for each rule actually in the grammar (checking for equality with the given rule), and then can refine (implement) the splitter for each rule separately. For example, for the grammar `(ab)`*, defined to have a single nonterminal `(ab)`* which can either be empty, or be mapped to 'a' 'b' (ab)*, we would refine this to the computation:

```
If [(ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str (ab)* [] splits }
Else If ['b', (ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str 'b' [(ab)*] splits }
Else If ['a', 'b', (ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str 'a' ['b', (ab)*] splits }
Else
    { dummy_splits : list ℕ | ⊤ }
```

where $=_p$ refers to a boolean equality test for productions. Note that in the final case, we permit any list to be picked, because whenever the production we are handling is reachable, it will never be used.

We can now refine each of these cases separately, using `setoid_rewrite`. The key to doing this, to making Fiat work, is that the refinement rules package their correctness properties, so users don't have to worry about correctness when programming by refinement. We use Coq's setoid rewriting machinery to automatically glue together the various correctness proofs when refining only a part of a program.

For example, we might have a lemma `singleton` which says that returning the length of the string is a valid refinement for any rule that has only one nonterminal; it's type, for a particular grammar `G`, a particular string `str`, and a particular nonterminal `nt`, would be

```
singleton G str nt
    : (ret [length str])
      ⊆
      { splits : list ℕ
      | split_list_is_complete G str nt [] splits }
```

Then `setoid_rewrite (singleton _ _ (ab)*)` would refine

```
If [(ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str (ab)* [] splits }
Else If ['b', (ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str 'b' [(ab)*] splits }
Else If ['a', 'b', (ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str 'a' ['b', (ab)*] splits }
Else
    { dummy_splits : list ℕ | ⊤ }
```

into

```
If [(ab)*] =p it :: its Then
    ret [length str]
Else If ['b', (ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str 'b' [(ab)*] splits }
Else If ['a', 'b', (ab)*] =p it :: its Then
    { splits : list ℕ
    | split_list_is_complete G str 'a' ['b', (ab)*] splits }
Else
    { dummy_splits : list ℕ | ⊤ }
```

We now describe the refinements that we do within this framework, to implement efficient splitters.

5.4 Optimizations

5.4.1 An Easy First Optimization: Indexed Representation of Strings

One optimization that is always possible is to represent the current string being parsed in this recursive call as a pair of indices into the original string. This allows us to optimize the code doing string manipulation, as it will no longer need to copy strings around, only do index arithmetic.

5.4.2 Upcoming Optimizations

In the next few sections, we build up various strategies for splitters. Although our eventual target is JavaScript, we cover only a more modest target of very simple arithmetical expressions in this paper. We begin by tying up the `(ab)*` grammar, and then moving on to parse numbers, parenthesized numbers, expressions with only numbers and `'+'`, and then expressions with numbers, `'+'` and parentheses.

[**TODO:** The following is mostly redundant with what was said above; remove most of it.] [**QUESTION FOR ADAM:** Is there any part here worth keeping, that isn't taken care of by the above.] For each grammar, the Fiat framework presents us with goals describing the unimplemented portion of the splitter for this particular grammar. For example, the goal for the grammar that parses arithmetic expressions involving plusses and parentheses, after taking care of the trivial obligations that we describe in the next chapter, looks like this:

```
1 focused subgoals (unfocused: 3)
, subgoal 1 (ID 3491)

r_n : string * (nat * nat)
n : item ascii * production ascii
H := ?88 : hiddenT
============================
 refine
   (ls <- If ([NonTerminal "pexpr"] =p fst n :: snd n)
             || (([NonTerminal "expr"] =p fst n :: snd n)
             || ([NonTerminal "number"] =p fst n :: snd n)
             || (([Terminal ")"] =p fst n :: snd n)
             || ([Terminal "0"] =p fst n :: snd n)
             || ([Terminal "1"] =p fst n :: snd n)
             || ([Terminal "2"] =p fst n :: snd n)
             || ([Terminal "3"] =p fst n :: snd n)
             || ([Terminal "4"] =p fst n :: snd n)
             || ([Terminal "5"] =p fst n :: snd n)
             || ([Terminal "6"] =p fst n :: snd n)
             || ([Terminal "7"] =p fst n :: snd n)
             || ([Terminal "8"] =p fst n :: snd n)
             || ([Terminal "9"] =p fst n :: snd n)
             || ([NonTerminal "digit"] =p fst n :: snd n)))
          Then ret [ilength r_n]
          Else (If [NonTerminal "pexpr"; Terminal "+";
                   NonTerminal "expr"] =p fst n :: snd n
          Then {splits : list nat
               | ParserInterface.split_list_is_complete
                 plus_expr_grammar
```

```
              (string_of_indexed r_n)
              (NonTerminal "pexpr")
              [Terminal "+"; NonTerminal "expr"]
              splits}
       Else
         ret [If ([Terminal "+"; NonTerminal "expr"]
                     =p fst n :: snd n)
                  || ([Terminal "("; NonTerminal "expr"; Terminal ")"]
                     =p fst n :: snd n)
                  || ([NonTerminal "digit"; NonTerminal "number"]
                     =p fst n :: snd n)
              Then 1
              Else (If [NonTerminal "expr"; Terminal ")"]
                          =p fst n :: snd n
              Then  pred (ilength r_n)
              Else (If [NonTerminal "number"]
                          =p fst n :: snd n
              Then ilength r_n
              Else 0))]);
  ret (r_n, ls)) (H r_n n)
```

The important part of this goal is:

```
{splits : list nat
| ParserInterface.split_list_is_complete
    plus_expr_grammar
    (string_of_indexed r_n)
    (NonTerminal "pexpr") [Terminal "+"; NonTerminal "expr"]
    splits}
```

It says that we have to find a refinement of non-deterministically picking a complete list of splits for the rule “a `pexpr`, followed by a `'+'`, followed by a `expr`.”

To get to this goal, we write this code:

```
Lemma ComputationalSplitter'
: FullySharpened (string_spec plus_paren_expr_grammar).
Proof.
  start honing parser using indexed representation.

  hone method "splits".
  {
    simplify parser splitter.
```

The first line of the proof takes care of the trivial rules, and `hone method "splits"` says that we want to refine the splitter. The tactic `simplify parser splitter` performs a number of straightforward and simple optimizations, such as combining nested `if` statements which return the same value.

We begin the next section by describing the strategies for the "trivial" rules.

Chapter 6

Fixed Length Items, Parsing (ab)*; Parsing #s; Parsing #, ()

In this chapter, we explore the Fiat framework with a few example grammars, which we describe how to parse. Because these rules are so straightforward, they can be handled automatically, in the very first step of the derivation; we will explain how this works, too.

Recall the grammar for the regular expression $\texttt{(ab)}^*$:

$$\texttt{(ab)}^* ::= \epsilon \mid \texttt{'a'}\ \texttt{'b'}\ \texttt{(ab)}^*$$

In addition to parsing this grammar, we will also be able to parse the grammar for non-negative parenthesized integers:

$$\begin{aligned}
\texttt{pexpr} &::= \texttt{'('}\ \texttt{pexpr}\ \texttt{')'} \mid \texttt{number} \\
\texttt{number} &::= \texttt{digit}\ \texttt{number?} \\
\texttt{number?} &::= \epsilon \mid \texttt{number} \\
\texttt{digit} &::= \texttt{'0'} \mid \texttt{'1'} \mid \texttt{'2'} \mid \texttt{'3'} \mid \texttt{'4'} \mid \texttt{'5'} \mid \texttt{'6'} \mid \texttt{'7'} \mid \texttt{'8'} \mid \texttt{'9'}
\end{aligned}$$

6.1 Parsing (ab)*: At Most One Nonterminal

The simpler of these grammars is the one for $\texttt{(ab)}^*$. The idea is that if any rule has at most one nonterminal, then there is only one possible split: we assign one character to each terminal, and the remaining characters to the single nonterminal.

For any given rule, we can compute straightforwardly whether or not this is the case; Haskell-like pseudocode for doing so is:

```
has-at-most-one-nt : [Item] -> Bool
has-at-most-one-nt [] = true
has-at-most-one-nt (Terminal _)::xs = has-at-most-one-nt xs
has-at-most-one-nt (NonTerminal _)::xs = has-only-terminals xs

has-only-terminals : [Item] -> Bool
has-only-terminals [] = true
has-only-terminals (Terminal _)::xs = has-only-terminals xs
has-only-terminals (NonTerminal _)::xs = false
```

The code for determining the split location is even easier: if the first item of the rule is a terminal, then split at character 1; if the first item of the rule is a nonterminal, and there are n remaining items in the rule, then split n characters before the end of the string.

6.2 Parsing Parenthesized Numbers: Fixed Lengths

The grammar for parenthsized numbers has only one rule with multiple nonterminals: the rule for `number` ::= `digit number?`. The strategy here is also simple: because `digit` only accepts strings of length exactly 1, we always want to split after the first character.

The following pseudocode computes the determines whether or not all strings parsed by a given item are a fixed length, and, if so, what that length is: [**TODO:** Colorize code; use macros to ensure it matches the conventions of the rest of the document]

```
fixed-length-of : [String] -> [Item] -> Maybe Int
fixed-length-of valid_nonterminals [] = just 0
fixed-length-of valid_nonterminals (Terminal _::xs)
  = case fixed-length-of valid_nonterminals xs of
      | just k -> just (1 + k)
      | nothing -> nothing
fixed-length-of valid_nonterminals (NonTerminal nt::xs)
  if nt ‘in‘ valid_nonterminals
  then let lengths = map (fixed-length-of (remove nt valid_nonterminals)) (Lookup nt)
       in if all of the elements of lengths are equal to L for some L
          then case L, fixed-length-of valid_nonterminals xs of
                | just L’, just k -> just (L’ + k)
                | _, _ -> nothing
          else nothing
  else nothing
```

We have proven that for any nonterminal for which this method returns `just` k, the only valid split of the any string for this rule is at location k. This is the correctness obligation that Fiat demands of us to be able to use this rule.

6.3 Putting It Together

Both of these rules are simple and complete for the rules they handle; if a rule has at most one nonterminal, or if the first element of a rule has a fixed length, then we can't do any better than these rules. Therefore, we incorperate them into the initial invocation of `start honing parser using indexed representation`. To do this, we express the splitter by folding `if` statements over all of the rules of the grammar that are reachable from valid nonterminals. The `if` statements check equality of the rule against the one we were given, and, if they match, looks to see if either of these strategies applies. If either does, than we return the appropriate singleton value. If neither applies, then we leave over a non-determinisitic pick of a list containing all possible valid splits. The results of applying this procedure without treating any rules specially was shown in Subsection 5.3.2. The results of applying this procedure, including the rules of this chapter, was shown in Subsection 5.4.2.

[**QUESTION FOR ADAM:** Is it worth including more code? I don't actually write out the refinement rules in the code I use, since I never need to setoid_rewrite with them; it's all baked into the initial first-step refinement rule. The general code for defining the computation, in case it's worth including is:]

```
Definition expanded_fallback_list'
           (P : String -> item Ascii.ascii -> production Ascii.ascii -> item Ascii.as
           (s : T)
           (it : item Ascii.ascii) (its : production Ascii.ascii)
           (dummy : list nat)
: Comp (T * list nat)
  := (ls <- (forall_reachable_productions
               G
               (fun p else_case
                => If production_beq ascii_beq p (it::its)
                   Then (match p return Comp (list nat) with
                           | nil => ret dummy
                           | _::nil => ret [ilength s]
                           | (Terminal _):: _ :: _ => ret [1]
                           | (NonTerminal nt):: p'
                             => If has_only_terminals p'
                                Then
                                  ret [(ilength s - Datatypes.length p')%natr]
                                Else
                                (option_rect
                                   (fun _ => Comp (list nat))
                                   (fun (n : nat) => ret [n])
```

```
                                  { splits : list nat
                                  | P
                                      (string_of_indexed s)
                                      (NonTerminal nt)
                                      p'
                                      it
                                      its
                                      splits }%comp
                                  (length_of_any G nt))
                     end)
                 Else else_case)
             (ret dummy));
ret (s, ls))%comp.
```

Chapter 7

Disjoint Items, Parsing #, +

Consider now the following grammar for arithmetic expressions involving '+' and numbers:

$$
\begin{aligned}
\texttt{expr} &::= \texttt{number +expr?} \\
\texttt{+expr?} &::= \epsilon \mid \texttt{'+' expr} \\
\texttt{number} &::= \texttt{digit number?} \\
\texttt{number?} &::= \epsilon \mid \texttt{number} \\
\texttt{digit} &::= \texttt{'0'} \mid \texttt{'1'} \mid \texttt{'2'} \mid \texttt{'3'} \mid \texttt{'4'} \mid \texttt{'5'} \mid \texttt{'6'} \mid \texttt{'7'} \mid \texttt{'8'} \mid \texttt{'9'}
\end{aligned}
$$

The only rule not handled by the strategies of the previous chapter is the rule `expr` ::= `number +expr?`. We can handle this rule by noticing that the set of characters in the strings accepted by `number` is disjoint from the set of possible first characters of the strings accepted by `+expr?`. Namely, all characters in strings accepted by `number` are digits, while the first character of a string accepted by `+expr?` can only be '+'.

The following code computes the set of possible characters of a rule:

```
possible-terminals' : [String] -> [Item] -> [Char]
possible-terminals' _ [] = []
possible-terminals' valid_nonterminals (Terminal ch :: xs)
  = [ch] `union` possible-terminals' valid_nonterminals xs
possible-terminals' valid_nonterminals (NonTerminal nt :: xs)
  = if nt `in` valid_nonterminals
    then fold
           union
           (possible-terminals' valid_nonterminals xs)
           (map (possible-terminals' (remove nt valid_nonterminals)) (Lookup nt))
```

```
    else possible-terminals' valid_nonterminals xs

possible-terminals : Grammar -> [Item] -> [Char]
possible-terminals G its = possible-terminals' (valid_nonterminals_of G) its
```

In the case where the nonterminal is not in the list of valid nonterminals, we assume that we have already seen that nonterminal higher up the chain of recursion (which we will have, if it is valid according to the initial list), and thus don't have to recompute its possible terminals.

The following code computes the set of possible first characters of a rule:

```
possible-first-terminals' : [String] -> [Item] -> [Char]
possible-first-terminals' _ [] = []
possible-first-terminals' valid_nonterminals (Terminal ch :: xs)
  = [ch]
possible-first-terminals' valid_nonterminals (NonTerminal nt :: xs)
  = (if nt `in` valid_nonterminals
     then fold
            union
            []
            (map (possible-first-terminals' (remove nt valid_nonterminals))
                 (Lookup nt))
     else [])
    `union`
    (if has_parse nt ""
     then possible-first-terminals' valid_nonterminals xs
     else [])


possible-first-terminals : Grammar -> [Item] -> [Char]
possible-first-terminals G its = possible-first-terminals' (valid_nonterminals_of G)
```

We can decide `has_parse` at refinement-time with the brute-force parser, which tries every split; when the string we're parsing is empty, $\mathcal{O}(\text{length!})$ is not that long. The idea is that we recursively look at the first element of each production; if it is a terminal, then that is the only possible first terminal of that production. If it's a nonterminal, then we have to fold the recursive call over the alternatives. Additionally, if the nonterminal might end up parsing the empty string, then we have to also move on to the next item in the production, and see what its first characters might be.

By computing whether or not these two lists are disjoint, we can decide whether or not this rule applies. When it applies, we can either look for the first character not in the first list (in this example, the list of digits), or we can look for the first character which is in the second list (in this case, the `'+'`). Since there are two alternatives, we leave it up to the user to decide which one to use.

For this grammar, we choose the shorter list. We define a function:

$$\texttt{index\_of\_first\_character\_in} : \texttt{String} \to \texttt{[Char]} \to \mathbb{N}$$

by folding over the string. We can then phrase the refinement rule as having type

$$\begin{array}{l}\texttt{is\_disjoint}\ (\texttt{possible-terminals}\ \texttt{G}\ [\texttt{it}])\ (\texttt{possible-first-terminals}\ \texttt{G}\ \texttt{its}) = \texttt{true} \\ \to\ \texttt{ret}\ [\texttt{index\_of\_first\_character\_in}\ \texttt{str}\ (\texttt{possible-first-terminals}\ \texttt{G}\ \texttt{its})] \\ \quad \subseteq \\ \quad \{\ \texttt{splits} : \texttt{list}\ \mathbb{N} \\ \quad \mid \texttt{split\_list\_is\_complete}\ \texttt{G}\ \texttt{str}\ \texttt{it}\ \texttt{its}\ \texttt{splits}\ \}\end{array}$$

Chapter 8

Parsing well-parenthesized expressions

8.1 At a Glance

We finally get to a grammar that requires a non-trivial splitting strategy. In this section, we describe how to parse strings for a grammar that accepts arithmetical expressions involving numbers, pluses, and well-balanced parentheses. More generally, this strategy handles any binary operation with guarded brackets.

8.2 Grammars we can parse

Consider the following two grammars, with `digit` denoting the nonterminal that accepts any single decimal digit.

Parenthesized addition:

$$
\begin{aligned}
\texttt{expr} &::= \texttt{pexpr +expr} \\
\texttt{+expr} &::= \epsilon \mid \texttt{'+'}\ \texttt{expr} \\
\texttt{pexpr} &::= \texttt{number} \mid \texttt{'('}\ \texttt{expr}\ \texttt{')'} \\
\texttt{number} &::= \texttt{digit number?} \\
\texttt{number?} &::= \epsilon \mid \texttt{number} \\
\texttt{digit} &::= \texttt{'0'} \mid \texttt{'1'} \mid \texttt{'2'} \mid \texttt{'3'} \mid \texttt{'4'} \mid \texttt{'5'} \mid \texttt{'6'} \mid \texttt{'7'} \mid \texttt{'8'} \mid \texttt{'9'}
\end{aligned}
$$

We have carefully constructed this grammar so that the first character of the string suffices to uniquely determine which rule of any given nonterminal to apply.

S-expressions are a notation for nested space-separated lists. By replacing `digit` with a nonterminal that accepts any symbol in a given set, which must not contain either of the brackets, nor whitespace, and replacing `'+'` with a space character `' '`, we get a grammar for S-expressions:

$$
\begin{aligned}
\texttt{expr} &::= \texttt{pexpr}\ \texttt{sexpr} \\
\texttt{sexpr} &::= \epsilon \mid \texttt{whitespace}\ \texttt{expr} \\
\texttt{pexpr} &::= \texttt{atom} \mid \texttt{'('}\ \texttt{expr}\ \texttt{')'} \\
\texttt{atom} &::= \texttt{symbol}\ \texttt{atom?} \\
\texttt{atom?} &::= \epsilon \mid \texttt{atom} \\
\texttt{whitespace} &::= \texttt{whitespace-char}\ \texttt{whitespace?} \\
\texttt{whitespace?} &::= \epsilon \mid \texttt{whitespace} \\
\texttt{whitespace-char} &::= \texttt{' '} \mid \texttt{'\textbackslash n'} \mid \texttt{'\textbackslash t'} \mid \texttt{'\textbackslash r'}
\end{aligned}
$$

8.3 The Splitting Strategy

8.3.1 The Main Idea

The only rule not already handled is the rule that says that a `pexpr +expr` is an `expr`. The key insight here is that, to know where to split, we need to know where the next `'+'` at the current level of parenthesization is. If we can compute an appropriate lookup table in time linear in the length of the string, then our splitter overall with be linear.

8.3.2 Building the Lookup Table

We build the table by reading the string from right to left, storing for each character the location of the next `'+'` at the current level of parenthesization. To compute this location we keep a list of the location of next `'+'` at every level of parenthesization.

[**QUESTION FOR ADAM:** Should this example be set off with a `\subsubsection` or a bold **Example.** or something?] Let's start with a very simple example, before moving to a more interesting one. To parse `"4+5"`, we are primarily interested in the case where we are parsing something that is a number, or parenthesized on the left, followed by a `'+'`, followed by any expression. For this item, we want to split the string right before the `'+'`, and say that the `"4"` can be parsed as a number (or parenthesized expression), and that the `"+5"` can be parsed as a `'+'` followed by an expression.

To do this, we build a table that keeps track of the location of the next `'+'`, starting from the right of the string. We will end up with the table:

Table Entries:		1	0	∅	
String:	"	4	+	5	"

At the '5', there is no next '+', and we are not parenthesized at all, so we record this as $\emptyset$. At the '+', we record that there is a '+' at the current level of parenthesization, with 0. Then, since the '4' is not a '+', we increment the previous number, and store 1. This is the correct location to split the string: we parse 1 character as a `number`, and the rest as `+expr`.

Now let's try a more complicated example: "(1+2)+3". We want to split this string into "(1+2)" and "+3". The above strategy is insufficient to do this; we need to keep track of the next '+' at all levels of parenthesization at each point. We will end up with the following table, where the bottom-most element is the current level, and the ones above that are higher levels. We use lines to indicate chains of numbers at the same level of parenthesization.

```
                            4 – 3 – 2 – 1
Table Entries:            /               \
                        5   1 – 0 – ∅ – ∅   0 – ∅
String:            "    (   1   +   2   )   +   3   "
```

We again start from the right. Since there are no '+'s that we have seen, we store the singleton list [$\emptyset$], indicating that we know about only the current level of parenthesization, and that there is no '+' to the right. At the '+' before the "3", we store the singleton list [0], indicating that the current character is a '+', and we only know about one level of parenthesization. At the ')', we increment the counter for '+', but we also now know about a new level of parenthesization. So we store the two element list [$\emptyset$, 1]. At the '3', we increment all numbers, storing [$\emptyset$, 2]. At the '+' before the "2", we store 0 at the current level, and increment higher levels, storing [0, 3]. At the '1', we simply increment all numbers, storing [1, 4]. Finally, at the '(', we pop a level of parenthesization, and increment the remaining numbers, storing [5]. This is correct; we have 5 characters in the first string, and when we go to split "1+2" into "1" and "+2", we have the list [1, 4], and the first string does indeed contain 1 character.

As an optimization, we can drop all but the first element of each list once we're done computing, and, in fact, can do this as we construct the table. However, for correctness, it is easier to reason about a list located at each location.

8.3.3 Table Correctness

What is the correctness condition on this table? The correctness condition Fiat gives us is that the splits we compute must be the only ones that give rise to valid parses. This is hard to reason about directly, so we use an intermediate correctness condition: for any cell of the table [**QUESTION FOR ADAM:** IS MEANING OF CELL OBVIOUS?], if it is empty (is $\emptyset$, or does not exist at all), then there must not be a well-parenthesized fragment of the string starting at that point and ending with a '+', which closes the appropriate number of parentheses for this level of parenthesization. [**TODO:** FIXME, THIS PREVIOUS SENTENCE SEEMS LIKE A CONFUSING EXPLANATION] [**QUESTION FOR ADAM:** Any suggestions?] If the cell points to a given location, then that location must contain a `'+'`, and the fragment of the string starting at the current location and going up to but not including the `'+'`, must not contain any `'+'`s which are "exposed". [**QUESTION FOR ADAM:** THIS ALSO SEEMS CONFUSING. Ideas? Should I put code/math first?]

More formally, we can define a notation of paren-balanced and paren-balanced-hiding-`'+'`. Say that a string is paren-balanced at level n if it closes exactly n more parentheses than it opens, and there is no point at which it has closed more than n more parentheses than it has opened. So the string `"1+2)"` is paren-balanced at level 1 (because it closes 1 more parenthesis than it opens), and the string `"1+2)+(3"` is not paren balanced at any level (though the string `"1+2)+(3)"` is paren-balanced at level 1). A string is paren-balanced-hiding-`'+'` at level n if it is paren-balanced at level n, and, at any point at which there is a `'+'`, at most $n - 1$ more parentheses have been closed than opened. So `"(1+2)"` is paren-balanced-hiding-`'+'` at level 0, and `"(1+2))"` is paren-balanced-hiding-`'+'` at level 1, and `"(1+2)+3"` is not paren-balanced-hiding-`'+'` at any level, though it is paren-balanced at level 0. [**QUESTION FOR ADAM:** THIS IS VERBOSE. MAYBE I SHOULD START WITH CODE?]

Then, the formal correctness condition is that if a cell at parenthesis level n points to a location ℓ, then the string from the cell up to but not including ℓ must be paren-balanced-hiding-`'+'` at level n, and the character at location ℓ must be a `'+'`. If the cell is empty, then the string up to but not including any subsequent `'+'` must not be paren-balanced at level n.

The table computed by the algorithm given above satisfies this correctness condition, and this correctness condition implies that the splitting locations given by the table are the only ones that produce valid parse trees; there is a unique table satisfying this correctness condition (because it picks out the *first* `'+'` at the relevant level), and any split location which is not appropriately paren-balanced/paren-balanced-hiding results in no parse tree. [**QUESTION FOR ADAM:** THIS SEEMS REPETITIVE?]

8.3.4 Diving into Refinement Code

[**TODO:** Check the following assumptions: that the reader has seen:

- what Fiat goals look like
- what the initial representation change to the indexed representation looks like
- the general structure of remaining splitter goals

]

[**TODO:** Remove me. GOALS: rule is non-trivial, but use should be trivial. pieces to making it trivial: no representation change, automatic discovery of binary operation and brackets, reflective side-conditions]

Although the rule itself is non-trivial, our goal is to make using this rule as trivial as possible; we now explain how this refinement rule requires only one line of code to use (modulo long tactic names):

```
  setoid_rewrite refine_binop_table;
  [ presimpl_after_refine_binop_table | reflexivity.. ].
```

[**QUESTION FOR ADAM:** How much explanation at this point? Mention that `presimpl...` only unfolds things?]

There are three components to making a rule that can be used with a one-liner: not requiring a change of representation; reflective computation of the side-conditions, allowing them all to be discharged with `reflexivity`; and automatic discovery of any arguments to the rule. We cover each of these separately.

Doing Without a New Representation

Recall from [**TODO:** secref] that the first step in any parser/splitter [**TODO:** decide on which one to use] is to use an indexed representation of strings, where splitting a string only involves updating the indices into the original string, as well as taking care of all of the straightforward grammar rules. Naively, we'd need to store a fresh table every time we split the string.

How do we get around computing a new table on splits? We pull the same trick here that we pulled on strings; we refer only to the table that is built from the initial string, and describe the correctness condition on the table in terms of arbitrary substrings of that string.

Fiat presents us with a goal involving a statement of the form "non-deterministically pick a list of splitting locations that is complete for the substring of `str` starting at `n` and ending at `m`, for the rule `pexpr+expr`." In code form, this is:

```
{splits : list nat
| split_list_is_complete
    G
    (substring n m str)
    (NonTerminal nt)
    (Terminal ch :: its)
    splits}
```

[**QUESTION FOR ADAM:** How much should I say about this code? For example, the current Coq code doesn't line up perfectly with the explanation; the code only handles the case where the rule is "a nonterminal, followed by a terminal, followed by other things", not the form presented above, where the nonterminal is hidden by one layer of indirection (to prevent backtracking).]

The final refinement rule, which we use with `setoid_rewrite`, says that this is refined by a lookup into a suitably defined `table`: [**TODO:** Standardize notation for goals, or drop code]

```
refine {splits : list nat
       | split_list_is_complete
           G
           (substring n m str)
           (NonTerminal nt)
           (Terminal ch :: its)
           splits}
       (ret [match List.nth n table None with
               | Some idx => idx
               | None => dummy_value
             end])
```

By phrasing the rule in terms of `substring n m str`, rather than in terms of an arbitrary string, the computation of the table is the same in every call to the splitter. All that remains is to lift the computation outside of the recursive call to the parsing function during the extraction phase, which we plan to do soon. [**TODO:** Section reference future work. Add bit to future work.]

Before moving on to the other components of making usage of this rule require only one line of code, we note that we make use of the essential property that removing characters from the end of the string doesn't add new locations where splitting could yield a valid parse; if a given location is the first `'+'` at the current level of parenthesization, this does not change when we remove characters from the end of the string. [**QUESTION FOR ADAM:** Should there be more content here?]

Discharging the Side-Conditions Trivially

To prove the correctness condition on the table, we need to know some things about the grammar that we are handling. In particular, we need to know that if we are trying to parse a string as a rule analogous to `pexpr`, then there can be no exposed `'+'` characters, and, furthermore, that every such parseable string has well-balanced parentheses. To allow discharging the side conditions trivially, we define a function that computes whether or not this is the case for a given nonterminal in a given grammar. We then prove that, whenever this function returns `true`, valid tables yield complete lists of splitting locations.

To make things simple, we approximate which grammars are valid; we require that every open parenthesis be closed in the same rule (rather than deeply nested in further nonterminals). In Haskell-like pseudocode, the function we use to check validity of a grammar can be written as:

```
grammar-and-nonterminal-is-valid : Grammar -> NonTerminal -> Bool
grammar-and-nonterminal-is-valid G nt
  = fold (&&) paren-balanced-hiding G (G.(Lookup) nt)

pb' : Grammar -> Char -> Nat -> [Item] -> Bool
pb' G n []          = (n == 0)
pb' G n (NonTerminal nt :: s)
  = fold (&&) (pb' G 0) (G.(Lookup) nt) && pb' G n s
pb' G n ('(' :: s) = pb' G (n + 1) s
pb' G n (')' :: s) = n > 0 && pb' G (n - 1) s
pb' G n (_ :: s)   = pb' G n s

paren-balanced G = pb' G 0

pbh' : Grammar -> Char -> Nat -> [Item] -> Bool
pbh' G n []          = (n == 0)
pbh' G n (NonTerminal nt :: s)
  = fold (pbh' G 0) (G.(Lookup) nt) && pb' G n s
pbh' G n ('+' :: s) = n > 0 && pbh' G n s
pbh' G n ('(' :: s) = pbh' G (n + 1) s
pbh' G n (')' :: s) = n > 0 && pbh' G (n - 1) s
pbh' G n (_ :: s)   = pbh' G n s

paren-balanced-hiding G = pbh' G 0
```

[**QUESTION FOR ADAM:** Does this code need more explanation?]

There is one subtlety here, that was swept under the rug in the above code: this computation might not terminate! We could deal with this by memoizing this computation in much the same way that we memoized the parser to deal with potentially infinite parse trees. Rather than dealing with the subtleties of figuring out what to do when we hit repeated nonterminals, we perform the computation in two steps. First, we trace the algorithm above, building up a list of which nonterminals need to be paren-balanced, and which ones need to be paren-balanced-hiding. Second, we fold the computation over these lists, replacing the recursive calls for nonterminals with list membership tests. [**QUESTION FOR ADAM:** Does this need code?]

Automatic Discovery of Arguments

Throughout this chapter, we've been focusing on the arithmetic expression example. However, the exact same rule can handle S-expressions, with just a bit of generalization. There are two things to be computed: the binary operation, and the parenthesis characters.[1]

We require that the first character of any string parsed by the nonterminal analogous to `+expr` be uniquely determined; that character will be the binary operator; we can reuse the code from Chapter 7 to compute this character and ensure that it is unique.

To find the parenthesis characters, we first compute a list of candidate character pairs: for each rule associated to the nonterminal analogous to `pexpr`, we consider the pair of the first character and the last character (assuming both are terminals) to be a candidate pair.[2] We then filter the list for characters which yield the conclusion that this rule is applicable to the grammar, according to the algorithm of the previous subsubsection. We then require, as a side-condition, that the length of this list be positive.

[1] Currently, our code requires the binary operation to be exposed as a terminal in the rule we are handling. We plan on generalizing this to the grammars described in this chapter shortly.

[2] Again, generalizing this to characters hidden by nested nonterminals should be straightforward.

Chapter 9

Future work

- Grammars and efficiency: The eventual target for this demonstration of the framework is the JavaScript grammar, and we aim to be competitive, performance-wise, with popular open-source JavaScript implementations. <lookup list of JS implementations> We plan to profile our parser against these on <lookup test suite>
  - Description of anticipated challenges, based on the JavaScript grammar <lookup JS grammar>
- Generating Parse Trees
  - We plan to eventually generate parse trees, and error messages, rather than just booleans, in the complete pipeline. We have already demonstrated that this requires only small adjustments to the algorithm in the section on the dependently typed parser.
- Validating extraction
  - By adapting <Clement's work>, our parsers will be able to be compiled to verified bedrock/assembly, within Coq

9.1 Future work with dependent types

Recall from Chapter 4 that dependent types have allowed us to refine our parsing algorithm to prove its own soundness and completeness.

However, we still have some work left to do to clean up the implementation of the dependently typed version of the parser.

Formal extensionality/parametricity proof To completely finish the formal proof of completeness, as described in this paper, we need to prove the parser extensionality axiom from Subsection 4.3.3. We need to prove that the parser does not make any decisions based on any arguments to its interface other than `split`, internalizing the obvious parametricity proof. (Alternatively, as mentioned above, we could hope to use an extension of Coq with internalized parametricity [3].)

Even more self-reference We might also consider reusing the same generic parser to generate the extensionality proofs, by instantiating the type families for success and failure with families of propositions saying that all instantiations of the parser, when called with the same parsing problem, always return values that are equivalent when converted to Booleans. A more specialized approach could show just that `has_parse` agrees with `parse` on successes and with `has_no_parse` on failures:

$$
\begin{aligned}
&\mathtt{T}_{\mathtt{success}}\ \_\ \_\ (\overline{\mathtt{s} \in \mathtt{nt}}) \\
&\quad := \mathtt{has\_parse}\ \mathtt{nt}\ \mathtt{s} = \mathtt{true} \wedge \mathtt{parse}\ \mathtt{nt}\ \mathtt{s} \neq \mathtt{None} \\
&\mathtt{T}_{\mathtt{failure}}\ \_\ \_\ (\overline{\mathtt{s} \in \mathtt{nt}}) \\
&\quad := \mathtt{has\_parse}\ \mathtt{nt}\ \mathtt{s} = \mathtt{false} \wedge \mathtt{has\_no\_parse} \neq \mathtt{inl}\ ()
\end{aligned}
$$

Synthesizing dependent types automatically? Although finding sufficiently general (dependent) type signatures was a Herculean task before we finished the completeness proof and discovered the idea of using parallel parse traces, it was mostly straightforward once we had proofs of soundness and completeness of the simply typed parser in hand; most of the issues we faced involving having to figure out how to thread additional hypotheses, which showed up primarily at the very end of the proof, through the entire parsing process. Subsequently instantiating the types was also mostly straightforward, with most of our time and effort being spent writing transformations between nearly identical types that had slightly different hypotheses, e.g., converting a `Foo` involving strings shorter than s_1 into another analogous `Foo`, but allowing strings shorter than s_2, where s_1 is not longer than s_2. Our experience raises the question of whether it might be possible to automatically infer dependently typed generalizations of an algorithm, which subsume already-completed proofs about it, and perhaps allow additional proofs to be written more easily.

Further generalization Finally, we believe our parser could be generalized even further; the algorithm we have implemented is essentially an algorithm for inhabiting arbitrary inductive type families, subject to some well-foundedness, enumerability, and finiteness restrictions on the arguments to the type family. The interface we described is, conceptually, a composition of this inhabitation algorithm with recursion and inversion principles for the type family we are inhabiting (`ParseTreeOf` in this paper). Our techniques for refining this algorithm so that it could prove itself sound and complete should therefore generalize to this viewpoint.

Bibliography

[1] José Bacelar Almeida et al. "Partial Derivative Automata Formalized in Coq". In: *Proceedings of the 15th International Conference on Implementation and Application of Automata*. CIAA'10. Winnipeg, MB, Canada: Springer-Verlag, 2011, pp. 59–68. ISBN: 978-3-642-18097-2. URL: `http://dl.acm.org/citation.cfm?id=1964285.1964292`.

[2] Aditi Barthwal and Michael Norrish. "Verified, Executable Parsing". In: *Proceedings of the 18th European Symposium on Programming Languages and Systems: Held As Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2009*. ESOP '09. York, UK: Springer-Verlag, 2009, pp. 160–174. ISBN: 978-3-642-00589-3. DOI: `10.1007/978-3-642-00590-9_12`. URL: `http://dx.doi.org/10.1007/978-3-642-00590-9_12`.

[3] Jean-Philippe Bernardy and Moulin Guilhem. "Type-theory in Color". In: *Proceedings of the 18th ACM SIGPLAN International Conference on Functional Programming*. ICFP '13. Boston, Massachusetts, USA: ACM, 2013, pp. 61–72. ISBN: 978-1-4503-2326-0. DOI: `10.1145/2500365.2500577`. URL: `http://doi.acm.org/10.1145/2500365.2500577`.

[4] Janusz A. Brzozowski. "Derivatives of Regular Expressions". In: *Journal of the ACM (JACM)* 11.4 (Oct. 1964), pp. 481–494. ISSN: 0004-5411. DOI: `10.1145/321239.321249`. URL: `http://doi.acm.org/10.1145/321239.321249`.

[5] Bryan Ford. "Parsing Expression Grammars: A Recognition-based Syntactic Foundation". In: *Proceedings of the 31st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '04. Venice, Italy: ACM, 2004, pp. 111–122. ISBN: 1-58113-729-X. DOI: `10.1145/964001.964011`. URL: `http://doi.acm.org/10.1145/964001.964011`.

[6] Graham Hutton. "Higher-order Functions for Parsing". In: *Journal of Functional Programming* 2.3 (July 1992), pp. 323–343.

[7] Jacques-Henri Jourdan, François Pottier, and Xavier Leroy. "Validating LR(1) Parsers". In: *Proceedings of the 21st European Conference on Programming Languages and Systems*. ESOP'12. Tallinn, Estonia: Springer-Verlag, 2012, pp. 397–416. ISBN: 978-3-642-28868-5. DOI: `10.1007/978-3-642-28869-2_20`. URL: `http://dx.doi.org/10.1007/978-3-642-28869-2_20`.

[8] Ramana Kumar et al. "CakeML: A Verified Implementation of ML". In: *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '14. San Diego, California, USA: ACM, 2014, pp. 179–191. ISBN: 978-1-4503-2544-8. DOI: `10.1145/2535838.2535841`. URL: `http://doi.acm.org/10.1145/2535838.2535841`.

[9] Matthew Might, David Darais, and Daniel Spiewak. "Parsing with Derivatives: A Functional Pearl". In: *Proceedings of the 16th ACM SIGPLAN International Conference on Functional Programming*. ICFP '11. Tokyo, Japan: ACM, 2011, pp. 189–195. ISBN: 978-1-4503-0865-6. DOI: `10.1145/2034773.2034801`. URL: `http://doi.acm.org/10.1145/2034773.2034801`.

[10] Greg Morrisett et al. "RockSalt: better, faster, stronger SFI for the x86". In: *Proceedings of the 33rd ACM SIGPLAN conference on Programming Language Design and Implementation*. PLDI '12. Beijing, China: ACM, 2012, pp. 395–404. ISBN: 978-1-4503-1205-9. DOI: `10.1145/2254064.2254111`. URL: `http://doi.acm.org/10.1145/2254064.2254111`.

[11] Magnus O. Myreen and Jared Davis. "A Verified Runtime for a Verified Theorem Prover". In: *Proceedings of the Second International Conference on Interactive Theorem Proving*. ITP'11. Berg en Dal, The Netherlands: Springer-Verlag, 2011, pp. 265–280. ISBN: 978-3-642-22862-9. URL: `http://dl.acm.org/citation.cfm?id=2033939.2033961`.

[12] Tom Ridge. "Simple, Functional, Sound and Complete Parsing for All Context-free Grammars". In: *Proceedings of the First International Conference on Certified Programs and Proofs*. CPP'11. Kenting, Taiwan: Springer-Verlag, 2011, pp. 103–118. ISBN: 978-3-642-25378-2. DOI: `10.1007/978-3-642-25379-9_10`. URL: `http://dx.doi.org/10.1007/978-3-642-25379-9_10`.

[13] Elizabeth Scott and Adrian Johnstone. "GLL Parsing". In: *Proceedings of the Ninth Workshop on Language Descriptions Tools and Applications*. LDTA '09. 2009, pp. 177–189.