

Chapter 1

PetitParser

with the participation of:

Lukas Renggli (renggli@gmail.com)

Building parsers to analyze data is a common task in software development. In this chapter we present a powerful parser framework named PetitParser. And contrary to its name, PetitParser is a really powerful parsing framework combining several technologies (scannerless parsers, parser combinators, packrat..). PetitParser is written by Lukas Renggli as part of his work on the Helvetia system but it can be used as a standalone tool.

1.1 Writing Parsers with PetitParser

PetitParser is a parsing framework different to many other popular parser generators. For example, it is not table based such as SmaCC or ANTLR. Instead it uses a unique combination of four alternative parser methodologies: scannerless parsers, parser combinators, parsing expression grammars and packrat parsers. As such PetitParser is more powerful in what it can parse and it arguably fits better the dynamic nature of Smalltalk. Let's have a quick look at these four parser methodologies:

Scannerless Parsers combine what is usually done by two independent tools (scanner and parser) into one. This makes writing a grammar much simpler and avoids common problems when grammars are composed.

Parser Combinators are building blocks for parsers modeled as a graph of composable objects; they are modular and maintainable, and can be changed, recomposed, transformed and reflected upon.

Parsing Expression Grammars (PEGs) provide the notion of ordered choices. Unlike in parser combinators, the ordered choice of PEGs always follows the first matching alternative and ignores other alternatives. Valid input always results in exactly one parse-tree, the result of a parse is never ambiguous.

Packrat Parsers give linear parse time guarantees and avoid common problems with left-recursion in PEGs.

Loading PetitParser

Enough talking, let's get started. PetitParser is developed in Pharo. A ready made image can be downloaded.¹ To load PetitParser into an existing image evaluate the following Gofer expression:

Script 1.1: *Installing PetitParser*

```
Gofer new
  renggli: 'petit';
  package: 'PetitParser';
  package: 'PetitTests';
  load.
```

There are other packages in the same repository that provide additional features, for example PetitSmalltalk is a Smalltalk grammar, PetitXml is an XML grammar, PetitJson is a JSON grammar, PetitAnalyzer provides functionality to analyze and transform grammars, and PetitGui is a Glamour IDE (see Chapter 1.5) for writing complex grammars. We are not going to use any of these packages for now.

More information on how to get PetitParser can be found on the website of the project.²

Writing a Simple Grammar

Writing grammars with PetitParser is as simple as writing Smalltalk code. For example a grammar parsing identifiers that start with a letter followed by zero or more letters or digits is defined as follows:

Script 1.2: *Creating our first parser to parse identifiers*

```
identifier := #letter asParser , #word asParser star.
```

To invoke such parser we use the message `parse:` (we will explain into details just after) as shown in the following snippet

¹<http://hudson.lukas-renggli.ch/job/PetitParser/lastSuccessfulBuild/artifact/petitparser>

²<http://scg.unibe.ch/research/helvetia/petitparser>



Figure 1.1: Syntax diagram representation for the identifier parser defined in script 1.2

```

|identifier|
identifier := #letter asParser , #word asParser star.
identifier parse: 'a987jlkjl'
  → #($a #($9 $8 $7 $j $l $k $j $l))
  
```

A Graphical Notation

Figure 1.1 presents a graphical representation, called a syntax diagram, of the identifier parser. Such a syntax diagram contains squared boxes (not shown on the figure) that represent non terminals, round boxes that represent terminals, arrows between these boxes, some entry points and some exit points. In the context of PetitParser, the squared boxes (non terminals) are in fact parsers composed of other parsers, and round boxes (terminals) are elementary parsers.

The idea behind this representation is that if an input can go from the entry point to the exit point by following the arrows and being consumed by the boxes, this input is matched by the represented parser.

If you inspect the object `identifier` of the previous script, you'll notice that it is an instance of a `PPSequenceParser`. If you dive further into the object you notice the following simple tree of different parser objects:

Script 1.3: Composition of parsers used for the identifier parser

```

PPSequenceParser (accepts a sequence of parsers)
  PPPredicateObjectParser (accepts a single letter)
  PPPossessiveRepeatingParser (accepts zero or more instances of another parser)
    PPPredicateObjectParser (accepts a single word character)
  
```

The root parser is a sequence parser because the `,` (comma) operator creates a sequence from (1) a letter parser and (2) zero or more word character parser. The root parser first child is a predicate object parser created by the `#letter asParser` expression. This parser is capable of parsing a single letter as defined by the `Character.isLetter` method. The root parser second child is a repeating parser created by the star call. This parser uses its child parser (another predicate object parser) as much as possible on the input (*i.e.*, it is a *greedy* parser). Its child parser is a predicate object parser created by the

`#word` asParser expression. This parser is capable of parsing a single digit or letter as defined by the `Character»isDigit` and `Character»isLetter` methods.

Parsing Some Input

To actually parse a string (or stream) we use the method `PPParser»parse:` as follows:

Script 1.4: *Parsing some input strings with the identifier parser*

```
identifier parse: 'yeah'.      —>  #($y #($e $a $h))
identifier parse: 'f123'.     —>  #($f #($1 $2 $3))
```

While it seems odd to get these nested arrays with characters as a return value, this is the default decomposition of the input into a parse tree. We'll see in a while how that can be customized.

If we try to parse something invalid we get an instance of `PPFailure` as an answer:

Script 1.5: *Parsing invalid input results in a failure*

```
identifier parse: '123'.      —>  letter expected at 0
```

This parsing results in a failure because the first character (1) is not a letter. Instances of `PPFailure` are the only objects in the system that answer with `true` when you send the message `#isPetitFailure`. Alternatively you can also use `PPParser»parse: onError:` to throw an exception in case of an error:

```
identifier
  parse: '123'
  onError: [ :msg :pos | self error: msg ].
```

If you are only interested if a given string (or stream) matches or not you can use the following constructs:

Script 1.6: *Checking that some inputs are identifiers*

```
identifier matches: 'foo'.      —>  true
identifier matches: '123'.     —>  false
identifier matches: 'foo()'.   —>  true
```

The last result can be quite surprising: indeed, a parenthesis is neither a digit nor a letter as was specified by the `#word` asParser expression. In fact, the identifier parser matches “foo” and this is enough for the `matches:` call to return `true`. The result would be similar with the use of `parse:` which would return `#($f #($o $o))`.

If you want to be sure that the complete input is matched, use the message `PPParser»end` as follows:

Script 1.7: Ensuring that the whole input is matched using end

```
identifier end matches: 'foo()'.    →    false
```

The end message creates a new parser that matches the end of input. To be able to compose parsers easily, it is important that parsers do not match the end of input by default. Because of this, you might be interested to find all the places that a parser can match using the message matchesSkipIn: and matchesIn:.

Script 1.8: Finding all matches in an input

```
identifier matchesSkipIn: 'foo 123 bar12'.
```

```
→    an OrderedCollection(#($f #($o $o)) #($b #($a $r $1 $2)))
```

```
identifier matchesIn: 'foo 123 bar12'.
```

```
→    an OrderedCollection(#($f #($o $o)) #($o #($o)) #($o #()) #($b #($a $r $1 $2))
    #($a #($r $1 $2)) #($r #($1 $2)))
```

The `PPParser»matchesSkipIn:` method returns a collection of arrays containing what has been matched (e.g., `#($f #($o $o))` and `#($b #($a $r $1 $2))` in this case. This function avoids parsing the same character twice. The method `PPParser»matchesIn:` does a similar job but returns a collection with all possible sub-parsed elements: e.g., evaluating `identifier matchesIn: 'foo 123 bar12'` returns a collection of 6 elements: `#($f #($o $o))`, `#($o #($o))`, `#($o #())`, `#($b #($a $r $1 $2))`, `#($a #($r $1 $2))`, and `#($r #($1 $2))`.

Script 1.9: Finding all matche ranges in an input

```
identifier matchingSkipRangesIn: 'foo 123 bar12'.
```

```
→    an OrderedCollection((1 to: 3) (9 to: 13))
```

```
identifier matchingRangesIn: 'foo 123 bar12'.
```

```
→    an OrderedCollection((1 to: 3) (2 to: 3) (3 to: 3) (9 to: 13) (10 to: 13) (11 to: 13))
```

Similarly, to find all the matching ranges (index of first character and index of last character) in the given input one can use either `PPParser»matchingSkipRangesIn:` or `PPParser»matchingRangesIn:` as shown by the script above.

Different Kinds of Parsers

PetitParser provide a large set of ready-made parser that you can compose to consume and transform arbitrarily complex languages. The terminal parsers are the most simple ones. We've already seen a few of those, some more are defined in Table 1.1.

| Terminal Parsers | Description |
|---------------------|---|
| \$a asParser | Parses the character \$a. |
| 'abc' asParser | Parses the string 'abc'. |
| #any asParser | Parses any character. |
| #digit asParser | Parses one digit (0..9). |
| #letter asParser | Parses one letter (a..z and A..Z). |
| #word asParser | Parses a digit or letter. |
| #blank asParser | Parses a space or a tabulation. |
| #newline asParser | Parses the carriage return or line feed characters. |
| #space asParser | Parses a space character. |
| #lowercase asParser | Parses a lowercase character. |
| #uppercase asParser | Parses an uppercase character. |

Table 1.1: PetitParser pre-defines a multitude of terminal parsers

| Parser Combinators | Description |
|--------------------|---|
| p1 , p2 | Parses p1 followed by p2 (sequence). |
| p1 / p2 | Parses p1, if that doesn't work parses p2. |
| p star | Parses zero or more p. |
| p plus | Parses one or more p. |
| p optional | Parses p if possible. |
| p and | Parses p but does not consume its input. |
| p negate | Parses p and succeeds when p fails. |
| p not | Parses p and succeeds when p fails, but does not consume its input. |
| p end | Parses p and succeeds only at the end of the input. |

Table 1.2: PetitParser pre-defines a multitude of parser combinators

The class side of `PPPredicateObjectParser` provides a lot of other factory methods that can be used to build more complex terminal parsers. To use them, send the message `asParser` to a symbol containing the name of the factory method (such as `#punctuation asParser`).

The next set of parsers are used to combine other parsers together and is defined in Table 1.2.

As a simple example of parser combination the expression is equivalent to the `#word asParser`.

Script 1.10: *Creating our first parser to parse identifiers*

```
identifier2 := #letter asParser , (#letter asParser / #digit asParser) star.
```

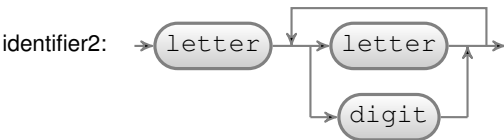


Figure 1.2: Syntax diagram representation for the identifier2 parser defined in script 1.10

Parser Action

In addition to the messages `parse:`, `matches:`, and `matchesSkipIn:` whose output is a collection of matched elements, we can also performs more operation on the parser output. To define an action or transformation to a parser we can use one of the messages `==>`, `flatten`, `token` or `trim` defined in Table 1.3.

| Action Parsers | Description |
|------------------------------|---|
| <code>p ==> aBlock</code> | Performs the transformation given in <code>aBlock</code> . |
| <code>p flatten</code> | Creates a string from the result of <code>p</code> . |
| <code>p token</code> | Similar to <code>flatten</code> but returns a <code>PPToken</code> with more information. |
| <code>p trim</code> | Trims whitespaces before and after <code>p</code> . |

Table 1.3: PetitParser pre-defines a multitude of action parsers

To return a string of the parsed identifier instead of getting a collection of matched element, configure the parser by sending it the message `flatten`.

Script 1.11: Using `flatten` so that the parsing result is a string

```
|identifier|
identifier := (#letter asParser , (#letter asParser / #digit asParser) star) flatten.
identifier parse: 'ajka0'
    -> 'ajka0'
identifier := (#letter asParser , (#letter asParser / #digit asParser) star).
identifier parse: 'ajka0'
    -> #($a #($j $k $a $0))
```

Sending the message `trim` configure the parser to remove whitespaces before and after. Study the following example: The first example leads to an error because the parser does not accept starting spaces. The second example handles perfectly the fact that the string starts with a couple of spaces.

Script 1.12: Using `trim` to remove spaces

```
|identifier|
identifier := (#letter asParser , (#letter asParser / (#space asParser) / #digit asParser)
    star) flatten.
```



Figure 1.3: Syntax diagram representation for the number parser defined in script 1.14

```

identifier parse: ' aj ka0 '
  → letter expected at 0

|identifier|
identifier := (#letter asParser , (#letter asParser / (#space asParser) / #digit asParser)
  star) trim flatten.
identifier parse: ' aj ka0 '
  → ' aj ka0 '

```

The message `==>` is really important since it lets you specify a block to be executed on the result of a parser. The next section presents several examples. Here is a simple way to get a number from its string representation.

Script 1.13: *Parsing integers*

```

number := #digit asParser plus trim flatten
  ==> [ :str | str asNumber ].
number parse: ' 122'
  → 122

```

The table 1.3 shows the basic elements to build parsers. There are a few more well documented and tested factory methods in the operators protocols of `PPParser`. If you want to know more about these factory methods, browse these protocols. Handy ones are `delimitedBy: aParser` which answers a new parser that parses the receiver one or more times, separated and possibly ended by `aParser`, `separatedBy: aParser` which answers a new parser that parses the receiver one or more times, separated by `aParser`.

Writing a More Complicated Grammar

Now let us start to write a more complicated grammar for evaluating simple arithmetic expressions. Within a workspace we start with the grammar for a number (actually an integer):

Script 1.14: *Parsing integers*

```

number := #digit asParser plus token trim ==> [ :token | token value asNumber ].
number parse: '123'
  → 123

```

The `token` call allows one to get a token in the action block instead of an array. The `trim` call asks the parser to handle spaces before and after the

number such as in '123'. The action block asks the parser to convert any parsed string (fetched from the token using `PPToken»value`) to a number using the `String»asNumber` method.

The next step is to define the productions for addition and multiplication in order of precedence. Note that we instantiate the productions as `PPDelegateParser` upfront, because they recursively refer to each other. The method `#setParser:` then resolves this recursion. The following script defines three parsers for the addition, multiplication and parenthesis (see Figure 1.4 for the related syntax diagram):

Script 1.15: *Parsing arithmetic expressions*

```
term := PPDelegateParser new.
prod := PPDelegateParser new.
prim := PPDelegateParser new.

term setParser: (prod , $+ asParser trim , term ==> [ :nodes | nodes first + nodes last ])
/ prod.
prod setParser: (prim , $* asParser trim , prod ==> [ :nodes | nodes first * nodes last ])
/ prim.
prim setParser: ($ ( asParser trim , term , $ ) asParser trim ==> [ :nodes | nodes second ])
/ number.
```

The term parser is defined as being either (1) a prod followed by '+', followed by another term or (2) a prod. In case (1), an action block asks the parser to compute the arithmetic addition of the value of the first node (a prod) and the last node (a term). The prod parser is similar to the term parser. The prim parser is interesting in that it accepts left and right parenthesis before and after a term and has an action block that simply ignores them.

To understand the precedence of productions, see Figure 1.5. The root of the tree in this figure (term), is the production that is tried first. A term is either a + or a prod. The term production comes first because + as the lowest priority in mathematics.

To make sure that our parser consumes all input we wrap it with the end parser into the start production:

```
start := term end.
```

That's it, we can now test our parser:

Script 1.16: *Trying our arithmetic expressions evaluator*

```
start parse: '1 + 2 * 3'.    —→  7
start parse: '(1 + 2) * 3'. —→  9
```

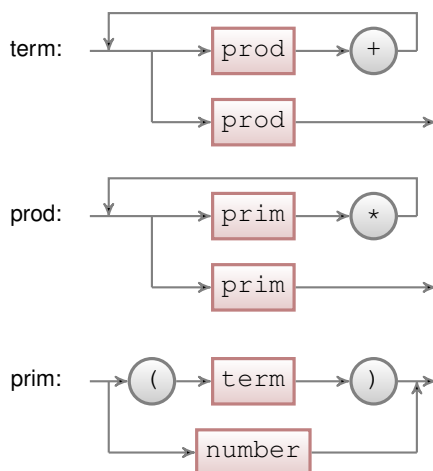


Figure 1.4: Syntax diagram representation for the term, prod, and prim parsers defined in script 1.15

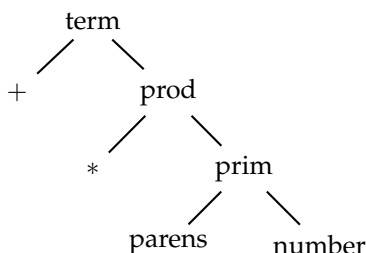


Figure 1.5: Explains how to understand the precedence of productions. An expression is a term which is either a sum or a production. It is necessary to recognize sums first as they have the lowest priority. A production is either a multiplication or a primitive. A primitive is either a parenthesised expression or a number. **Stéf** ► I do not get this picture ◀ **DamienC** ► is it better now? ◀

1.2 Composite Grammars with PetitParser

In the previous section we saw the basic principles of PetitParser and gave some introductory examples. In this section we are going to present a way to define more complicated grammars. We continue where we left off with the arithmetic expression grammar.

Writing parsers as a script as we did previously can be cumbersome, especially when grammar productions are mutually recursive and refer to each other in complicated ways. Furthermore a grammar specified in a sin-

gle script makes it unnecessary hard to reuse specific parts of that grammar. Luckily there is PPCompositeParser to the rescue.

Defining the Grammar

As an example let's create a composite parser using the same expression grammar we built in the last section but this time we define it inside a class subclass of PPCompositeParser.

Script 1.17: *Creating a class to hold our arithmetic expression grammar*

```
PPCompositeParser subclass: #ExpressionGrammar
  instanceVariableNames: "
  classVariableNames: "
  poolDictionaries: "
  category: 'PetitTutorial'
```

Again we start with the grammar for an integer number. Define the method ExpressionGrammar>>number as follows:

Script 1.18: *Implementing our first parser as a method*

```
ExpressionGrammar>>number
  ↑ #digit asParser plus token trim ==> [ :token | token value asNumber ]
```

Every production in ExpressionGrammar is specified as a method that returns its parser. Therefore we define the productions term, prod, mul, and prim.

Now there is a trick to be able to create recursive grammars. Productions refer to each other by reading the respective instance variable of the same name. The instance variables themselves are typically not written to and initialized as PetitParser takes care to initialize them for you automatically. We let Pharo automatically add the necessary instance variables as we refer to them for the first time. We obtain the following class definition:

Script 1.19: *Creating a class to hold our arithmetic expression grammar*

```
PPCompositeParser subclass: #ExpressionGrammar
  instanceVariableNames: 'add prod tern mul parens number'
  classVariableNames: "
  poolDictionaries: "
  category: 'PetitTutorial'
```

Script 1.20: *Defining more expression grammar parsers, this time with no associated action*

```
ExpressionGrammar>>term
  ↑ add / prod
```

```

ExpressionGrammar>>add
  ↑ prod , $+ asParser trim , term

ExpressionGrammar>>prod
  ↑ mul / prim

ExpressionGrammar>>mul
  ↑ prim , $* asParser trim , prod

ExpressionGrammar>>prim
  ↑ parens / number

ExpressionGrammar>>parens
  ↑ $( asParser trim , term , $) asParser trim

```

Contrary to our previous implementation we do not define the production actions yet (what we previously did by using `PPParser»==>`); and we factor out the parts for addition (`add`), multiplication (`mul`), and parenthesis (`parens`) into separate productions. This will give us better reusability later on. For example, a subclass may override such methods to produce slightly different production output. Usually, production methods are categorized in a protocol named `grammar` (which can be refined into more specific protocol names when necessary such as `grammar-literals`).

Last but not least we define the starting point of the expression grammar. This is done by overriding `PPCompositeParser»start` in the `ExpressionGrammar` class:

Script 1.21: *Defining the starting point of our expression grammar parser*

```

ExpressionGrammar>>start
  ↑ term end

```

Instantiating the `ExpressionGrammar` gives us an expression parser that returns a default abstract-syntax tree:

Script 1.22: *Testing our parser on simple arithmetic expressions*

```

parser := ExpressionGrammar new.
parser parse: '1 + 2 * 3'.    →  #(1 $+ #(2 $* 3))
parser parse: '(1 + 2) * 3'. →  #(#($ ( #(1 $+ 2) $)) $* 3)

```

Defining an Evaluator

Now that we have defined a grammar we can reuse this definition to implement an evaluator. To do this we create a *subclass* of `ExpressionGrammar` called `ExpressionEvaluator`.

Script 1.23: *Separating the grammar from the evaluator by creating a subclass*

```
ExpressionGrammar subclass: #ExpressionEvaluator
  instanceVariableNames: "
  classVariableNames: "
  poolDictionaries: "
  category: 'PetitTutorial'
```

We then redefine the implementation of add, mul and parens with our evaluation semantics. This is accomplished by calling the super implementation and adapting the returned parser as shown in the following methods.

Script 1.24: *Refining the definition of some parsers to evaluate arithmetic expressions*

```
ExpressionEvaluator>>add
  ↑ super add ==> [ :nodes | nodes first + nodes last ]

ExpressionEvaluator>>mul
  ↑ super mul ==> [ :nodes | nodes first * nodes last ]

ExpressionEvaluator>>parens
  ↑ super parens ==> [ :nodes | nodes second ]
```

The evaluator is now ready to be tested:

Script 1.25: *Testing our evaluator on simple arithmetic expressions*

```
parser := ExpressionEvaluator new.
parser parse: '1 + 2 * 3'.      → 7
parser parse: '(1 + 2) * 3'.    → 9
```

Defining a Pretty-Printer

We can reuse the grammar for example to define a simple pretty printer. This is as easy as subclassing ExpressionGrammar again!

Script 1.26: *Separating the grammar from the pretty printer by creating a subclass*

```
ExpressionGrammar subclass: #ExpressionPrinter
  instanceVariableNames: "
  classVariableNames: "
  poolDictionaries: "
  category: 'PetitTutorial'

ExpressionPrinter>>add
  ↑ super add ==> [:nodes | nodes first, ' + ', nodes third]

ExpressionPrinter>>mul
```

```
↑ super mul ==> [:nodes | nodes first, ' * ', nodes third]
```

```
ExpressionPrinter>>number
```

```
↑ super number ==> [:num | num printString]
```

```
ExpressionPrinter>>parens
```

```
↑ super parens ==> [:node | '(', node second, ')']
```

This pretty printer can be tried out as shown by the following expressions.

Script 1.27: *Testing our pretty printer on simple arithmetic expressions*

```
parser := ExpressionPrinter new.
```

```
parser parse: '1+2 *3'.      → '1 + 2 * 3'
```

```
parser parse: '(1+ 2) * 3'.  → '(1 + 2) * 3'
```

1.3 Testing a Grammar

The PetitParser contains a framework dedicated to testing your grammars. Testing a grammar is done by subclassing `PPCompositeParserTest` as follows:

Script 1.28: *Creating a class to hold the tests for our arithmetic expression grammar*

```
PPCompositeParserTest subclass: #ExpressionGrammarTest
```

```
instanceVariableNames: "
```

```
classVariableNames: "
```

```
poolDictionaries: "
```

```
category: 'PetitTutorial'
```

It is then important that the test case class references the parser class: this is done by overriding the `PPCompositeParserTest>>parserClass` method in `ExpressionGrammarTest`:

Script 1.29: *Linking our test case class to our parser*

```
ExpressionGrammarTest>>parserClass
```

```
↑ ExpressionGrammar
```

Writing a test scenario is done by implementing new methods in `ExpressionGrammarTest`:

Script 1.30: *Implementing tests for our arithmetic expression grammar*

```
ExpressionGrammarTest>>testNumber
```

```
self parse: '123' rule: #number.
```

```
ExpressionGrammarTest>>testAdd
```

```
self parse: '123+77' rule: #add.
```

These tests ensure that the ExpressionGrammar parser can parse some expressions using a specified production rule. Testing the evaluator and pretty printer is similarly easy:

Script 1.31: *Testing the evaluator and pretty printer*

```
ExpressionGrammarTest subclass: #ExpressionEvaluatorTest
```

```
instanceVariableNames: "
```

```
classVariableNames: "
```

```
poolDictionaries: "
```

```
category: 'PetitTutorial'
```

```
ExpressionEvaluatorTest>>parserClass
```

```
↑ ExpressionEvaluator
```

```
ExpressionEvaluatorTest>>testAdd
```

```
super testAdd.
```

```
self assert: result equals: 200
```

```
ExpressionEvaluatorTest>>testNumber
```

```
super testNumber.
```

```
self assert: result equals: 123
```

```
ExpressionGrammarTest subclass: #ExpressionPrinterTest
```

```
instanceVariableNames: "
```

```
classVariableNames: "
```

```
poolDictionaries: "
```

```
category: 'PetitTutorial'
```

```
ExpressionPrinterTest>>parserClass
```

```
↑ ExpressionPrinter
```

```
ExpressionPrinterTest>>testAdd
```

```
super testAdd.
```

```
self assert: result equals: '123 + 77'
```

```
ExpressionPrinterTest>>testNumber
```

```
super testNumber.
```

```
self assert: result equals: '123'
```

1.4 Case Study: A JSON Parser

In this section we illustrate PetitParser through the development of a JSON parser.³ JSON is a lightweight data-interchange format defined in <http://www.>

³Written by Lukas Renggli and available for download at <http://source.lukas-renggli.ch/petit/>

json.org. We are going to use the specification on this website to define our own JSON parser.

JSON is a simple format based on nested pairs and arrays. The following script gives an example taken from Wikipedia <http://en.wikipedia.org/wiki/JSON>

Script 1.32: *An example of JSON*

```
{  "firstName": "John",
   "lastName": "Smith",
   "age": 25,
   "address":
     { "streetAddress": "21 2nd Street",
       "city": "New York",
       "state": "NY",
       "postalCode": "10021" },
   "phoneNumber":
     [
       { "type": "home",
         "number": "212 555-1234" },
       { "type": "fax",
         "number": "646 555-4567" } ] }
```

JSON consists of object definitions (between curly braces “{ }”) and arrays (between square brackets “[]”). An object definition is a set of key/value pairs whereas an array is a list of values. The previous JSON example then represents an object (a person) with several key/value pairs (e.g., for the person firstname, name, age...). The address of the person is represented by another object while the phone number is represented by an array of objects.

First we define a grammar as subclass of PPCompositeParser. Let us call it PPJsonGrammar

Script 1.33: *Defining the JSON grammar class*

```
PPCompositeParser subclass: #PPJsonGrammar
  instanceVariableNames: ""
  classVariableNames: 'CharacterTable'
  poolDictionaries: ""
  category: 'PetitJson-Core'
```

We define the CharacterTable class variable since we will later use it to parse strings.

Parsing Objects and Arrays

Stéf ► what is an element? I rename members into elements to be in sync with the code◀ Stéf
 ► I'm quite sure that this code was not run.◀

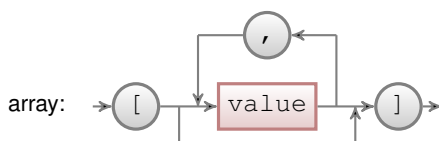


Figure 1.7: Syntax diagram representation for the JSON array parser defined in script 1.35

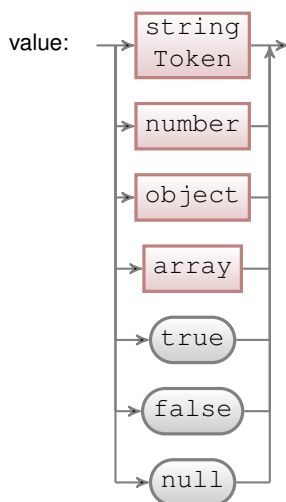


Figure 1.8: Syntax diagram representation for the JSON value parser defined in script 1.36

Script 1.36: *Defining the JSON parser for value as represented in Figure 1.8*

```

PPJsonGrammar>>value
↑ stringToken / numberToken / object / array /
  trueToken / falseToken / nullToken
  
```

A string requires quite some work to parse. A string starts and end with double-quotes. What is inside these double-quotes is a sequence of characters. Any character can either be an escape character, an octal character, or a normal character. An escape character is composed of a backslash immediately followed by a special character (e.g., '\n' to get a new line in the string). An octal character is composed of a backslash, immediately followed by the letter 'u', immediately followed by 4 hexadecimal digits. Finally, a normal character is any character except a double quote (used to end the string) and a backslash (used to introduce an escape character).

Script 1.37: *Defining the JSON parser for string as represented in Figure 1.9*

```

PPJsonGrammar>>stringToken
  ↑ string token trim
PPJsonGrammar>>string
  ↑ "$" asParser , char star , "$" asParser
PPJsonGrammar>>char
  ↑ charEscape / charOctal / charNormal
PPJsonGrammar>>charEscape
  ↑ "$\" asParser , (PPPredicateObjectParser anyOf: (String withAll: CharacterTable keys))
PPJsonGrammar>>charOctal
  ↑ '\u' asParser , (#hex asParser min: 4 max: 4)
PPJsonGrammar>>Normal
  ↑ PPPredicateObjectParser anyExceptAnyOf: ""

```

Special characters allowed after a slash and their meanings are defined in the CharacterTable dictionary that we initialize in the initialize class method.

Script 1.38: *Defining the JSON special characters and their meaning*

```

PPJsonGrammar class>>initialize
  CharacterTable := Dictionary new.
  CharacterTable
    at: $\" put: $\";
    at: $/ put: $/;
    at: $" put: $" ;
    at: $b put: Character backspace;
    at: $f put: Character newPage;
    at: $n put: Character lf;
    at: $r put: Character cr;
    at: $t put: Character tab

```

Parsing numbers is only slightly simpler as a number can be positive or negative and integral or decimal. Additionally, a decimal number can be expressed with a floating number syntax.

Script 1.39: *Defining the JSON parser for number as represented in Figure 1.10*

```

PPJsonGrammar>>numberToken
  ↑ number token trim
PPJsonGrammar>>number
  ↑ "$-" asParser optional ,
    ($0 asParser / #digit asParser plus) ,
    ($. asParser , #digit asParser plus) optional ,
    (($e asParser / $E asParser) , ($- asParser / $+ asParser) optional , #digit asParser
    plus) optional

```

The attentive reader will have noticed a small difference between the syntax diagram in Figure 1.10 and the code in script 1.39. Numbers in JSON can

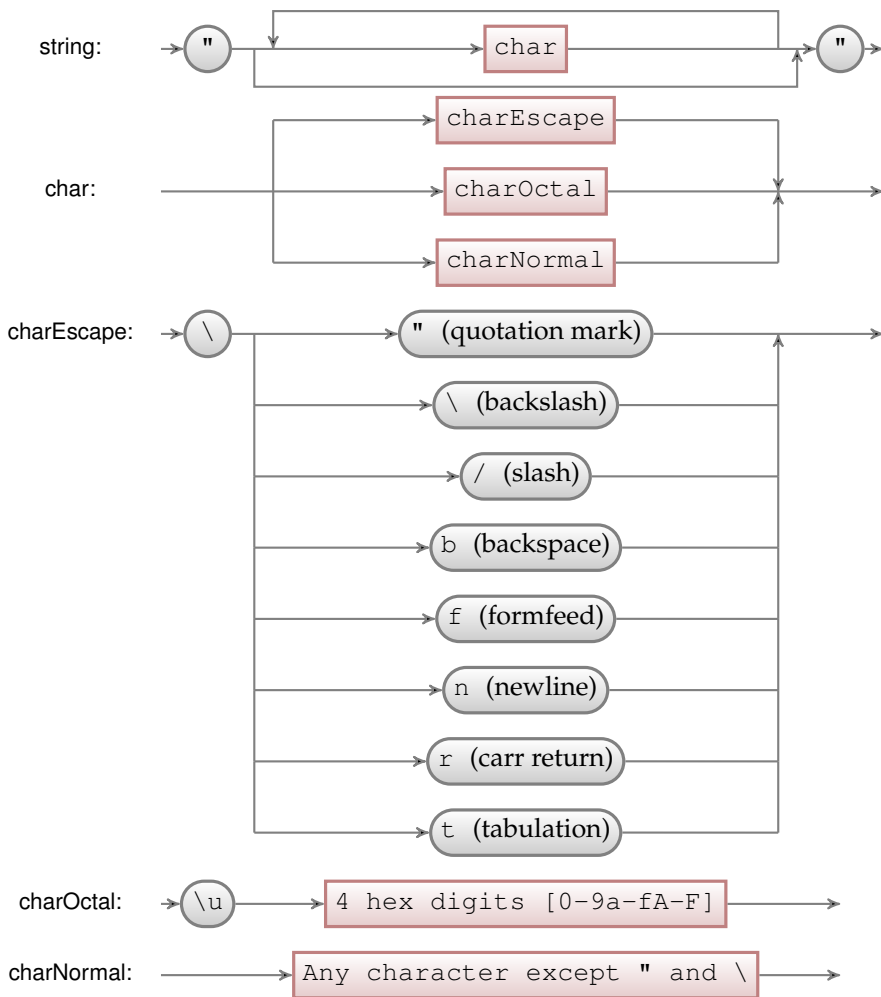


Figure 1.9: Syntax diagram representation for the JSON string parser defined in script 1.37

not contain leading zeros: *i.e.*, strings such as `"01"` do not represent valid numbers. The syntax diagram makes that particularly explicit by allowing either a 0 or a digit between 1 and 9. In the above code, the rule is made implicit by relying on the fact that the parser combinator `$/` is ordered: the parser on the right of `$/` is only tried if the parser on the left fails: thus, `($0 asParser / #digit asParser plus)` defines numbers as being just a 0 or a sequence of digits not starting with 0.

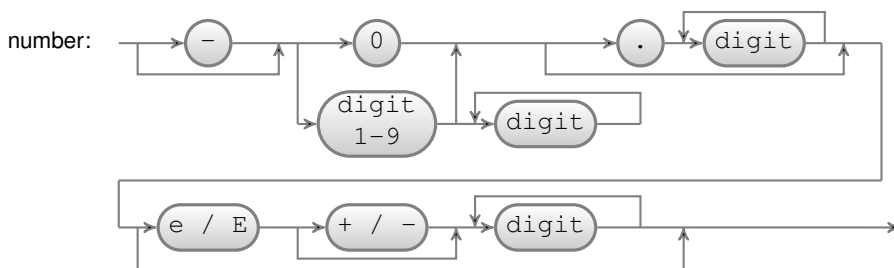


Figure 1.10: Syntax diagram representation for the JSON number parser defined in script 1.39

Stéf ► why (\$0 asParser / #digit asParser plus) why #digit does not handle 0? ◀ **DamienC** ► is it clearer now? ◀

The other parsers are fairly trivial:

Script 1.40: Defining missing JSON parsers

```

PPJsonGrammar>>>falseToken
↑ 'false' asParser token trim
PPJsonGrammar>>>nullToken
↑ 'null' asParser token trim
PPJsonGrammar>>>trueToken
↑ 'true' asParser token trim
  
```

The only piece missing is the start parser.

Script 1.41: Defining the JSON start parser as being a value (Figure 1.8) with nothing following

```

PPJsonGrammar>>start
↑ value end
  
```

Stéf ► what value? where does it come from? ◀

1.5 Chapter Conclusion

This concludes our tutorial of PetitParser. Please note that this tutorial is not meant to give an exhaustive overview of PetitParser, but is merely intended to introduce the reader to its usage. For a more extensive view of PetitParser, its concepts and implementation, the Moose book⁴ and Lukas Renggli's PhD have a dedicated chapter dedicated.

⁴<http://www.themoosebook.org/book>