
MOUSE: FROM PARSING EXPRESSIONS TO A PRACTICAL PARSER

Version 2.3

Roman R. Redziejowski

Parsing Expression Grammar (PEG) is a way to specify recursive-descent parsers with limited backtracking. The use of backtracking lifts the $LL(1)$ restriction usually imposed by top-down parsers. In addition, PEG can define parsers with integrated lexing.

Mouse is a tool to transcribe PEG into an executable parser written in Java. Unlike some existing PEG generators (e.g., *Rats!*), *Mouse* does not produce a storage-hungry "packrat parser", but a collection of transparent recursive procedures.

An integral feature of *Mouse* is the mechanism for specifying semantics (also in Java). This makes *Mouse* a convenient tool if one needs an ad-hoc language processor. Being written in Java, the processor is operating-system independent.

Starting with Version 2.0, *Mouse* offers some support for left-recursive grammars.

This is a user's manual in the form of a tutorial that introduces the reader to *Mouse* through a hands-on experience.

April 30, 2021

Copyright © 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2019, 2020, 2021
by Roman R. Redziejowski (www.romanredz.se).

The author gives unlimited permission to copy, translate and/or distribute this document, with or without modifications, as long as this notice is preserved, and information is provided about any changes.

This document is available at mousepeg.sourceforge.net/Documents/Manual.pdf.

Contents

1	Introduction	1
2	Parsing Expression Grammar	2
3	Getting started	4
4	The first steps	4
5	Adding semantics	6
6	Understanding the "right-hand side"	10
7	Getting more realistic	10
8	Let's go floating	12
9	What about backtracking?	13
10	A mouse, not a pack rat	14
11	Full arithmetic	15
12	Want a tree?	17
13	Calculator with memory	18
14	Get error handling right	20
15	Input from file	22
16	Error recovery	23
17	Left Recursion	24
18	Extensions to PEG	26
19	Checking the grammar	27
20	Miscellaneous features	28
21	Deploying the parser	31
A	Appendix: The grammar of <i>Mouse</i> PEG	32
B	Appendix: Helper methods	33
C	Appendix: Your parser class	34
D	Appendix: <i>Mouse</i> tools	35
E	References	39

Changes from version of July 1, 2020:

- Added sections [18](#) and [19](#).
- Removed second section on backtracking.
- Introduced new operators ":" and "!", described in Section [18](#).
- Description of operators "*" and "+" moved to Section [18](#).

1 Introduction

Parsing Expression Grammar (PEG), introduced by Ford in [7], is a way to define the syntax of a programming language. It encodes a recursive-descent parser for that language. *Mouse* is a tool to transcribe PEG into an executable parser. Both *Mouse* and the resulting parser are written in Java, which makes them operating-system independent. An integral feature of *Mouse* is the mechanism for specifying semantics (also in Java).

Recursive-descent parsing with backtracking

Recursive-descent parsers consist of procedures that correspond to syntax rules. The procedures call each other recursively, each being responsible for recognizing input strings defined by its rule. The syntax definition and its parser can be then viewed as the same thing, presented in two different ways. This design is very transparent and easy to modify when the language evolves. The idea can be traced back to Lucas [11] who suggested it for grammars that later became known as the Extended Backus-Naur Form (EBNF).

The problem is that procedures corresponding to certain types of syntax rules must decide which procedure to call next. It is easily solved for a class of languages that have the so-called LL(1) property: the decision can be made by looking at the next input symbol. Some syntax definitions can be transformed to satisfy this property. However, forcing the language into the LL(1) mold can make the grammar – and the parser – unreadable.

One can always use brute force: trial and error. It means trying the alternatives one after another and backtracking after a failure. But, an exhaustive search may require exponential time, so a possible option is *limited* backtracking: never return after a partial success. This approach has been used in actual parsers [4, 13] and is described in [1–3, 10]. It has been eventually formalized by Ford [7] under the name of Parsing Expression Grammar (PEG).

Wikipedia has an article on PEG [21], and a dedicated Web site [5] contains list of publications about PEGs and a link to discussion forum.

PEG programming

Parsers defined by PEG do not require a separate "lexer" or "scanner". Together with lifting of the LL(1) restriction, this gives a very convenient tool when you need an ad-hoc parser for some application. However, the limitation of backtracking may have unexpected effects that give an impression of PEG being unpredictable.

In its external appearance, PEG is very like an EBNF grammar. For some grammars, the limited backtracking is "efficient", in the sense that it finds everything that would be found by full backtracking. Parsing Expression Grammar with efficient backtracking defines exactly the same language as its EBNF look-alike. It agrees well with intuition and is therefore easier to understand. Some conditions for efficient backtracking have been identified in [17]. The *Mouse* package includes a tool, the *PEG Explorer*, that assists you in checking whether your PEG has efficient backtracking. You find its documentation at mousepeg.sourceforge.net/explorer.htm.

PEG has a feature that does not have an EBNF counterpart: the syntactic predicates `!e` and `&e`. Not being part of EBNF, they have to be understood on their own.

Mouse - not a pack rat

Even the limited backtracking may require a lot of time. In [6], PEG was introduced together with a technique called *packrat parsing*. Packrat parsing handles backtracking by extensive *memoization*: storing all results of parsing procedures. It guarantees linear parsing time at a large memory cost. (The name "packrat" comes from *pack rat* – a small rodent *Neotoma cinerea* known for hoarding unnecessary items.)

Wikipedia [20] lists a number of generators producing packrat parsers from PEG.

The amount of backtracking does not matter in small interactive applications where the input is short and performance not critical. Moreover, the usual programming languages do not require much backtracking. Experiments reported in [15, 16] demonstrated a moderate backtracking activity in PEG parsers for programming languages Java and C.

In view of these facts, it is useful to construct PEG parsers where the complexity of packrat technology is abandoned in favor of simple and transparent design. This is the idea of *Mouse*: a parser generator that transcribes PEG into a set of recursive procedures that closely follow the grammar. The name *Mouse* was chosen in contrast to *Rats!*, one of the first generators producing packrat parsers [8]. Optionally, *Mouse* can offer a small amount of memoization using the technique described in [15].

Left recursion

The recursive-descent process used by PEG parsers can not be applied to a grammar that contains left recursion, because rule such as $A = A"a"/"a"$ would result in an infinite descent. But, for many reasons, this pattern is often used for defining the syntax of programming languages.

Converting left recursion to iteration is possible, but is tedious, error-prone, and obscures the spirit of the grammar. Several ways of extending PEG to handle left recursion have been suggested [9, 14, 18, 19].

Version 2.0, *Mouse* supports left recursion using an experimental method of *recursive ascent*, which continues the idea from [9]. It boils down to constructing behind the scenes an alternative PEG that takes care of left-recursive portions. It is explained in mousepeg.sourceforge.net/Documents/RecursiveAscent.pdf. The support is limited to the Choice and Sequence expressions, and imposes some restrictions on expressions and semantic actions.

After a short presentation of PEG in the following section, sections 3 through 17 have the form of a tutorial, introducing the reader to *Mouse* by hands-on experience. They are followed by descriptions of some extensions and details not covered in the tutorial.

2 Parsing Expression Grammar

2.1 Parsing expressions

Parsing expressions are instructions for parsing strings. You can think of a parsing expression as shorthand for a procedure that carries out such instruction.

Parsing expression is applied to input text – a character string – and tries to recognize the initial portion of that text. If it succeeds, it "consumes" the recognized portion and indicates "success"; otherwise, it indicates "failure" and does not consume anything.

The following five kinds of parsing expressions work directly on the input text:

"s"	where s is a nonempty character string. If the text starts with the string s , consume that string and indicate success. Otherwise indicate failure.
[s]	where s is a nonempty character string. If the text starts with a character appearing in s , consume that character and indicate success. Otherwise indicate failure.
^[s]	where s is a nonempty character string. If the text starts with a character <i>not</i> appearing in s , consume that character and indicate success. Otherwise indicate failure.
[c ₁ -c ₂]	where c_1, c_2 are two characters. If the text starts with a character from the range c_1 through c_2 , consume that character and indicate success. Otherwise indicate failure.
_	where $_$ is the underscore character. If there is a character ahead, consume it and indicate success. Otherwise (that is, at the end of input) indicate failure.

These expressions are analogous to terminals of a classical context-free grammar, and are referred to as "terminals". The remaining kinds of parsing expressions invoke other expressions to do their job:

$e?$	Invoke the expression e and indicate success whether it succeeded or not.
e^*	Invoke the expression e repeatedly as long as it succeeds. Indicate success even if it did not succeed a single time.
e^+	Invoke the expression e repeatedly as long as it succeeds. Indicate success if e succeeded at least once. Otherwise indicate failure.
$\&e$	Invoke the expression e and then reset the input as it was before the invocation of e (do not consume anything). Indicate success if e succeeded or failure if e failed.
$!e$	Invoke the expression e and then reset the input as it was before the invocation of e (do not consume anything). Indicate success if e failed or failure if e succeeded.
$e_1 \dots e_n$	Invoke expressions e_1, \dots, e_n , in this order, as long as they succeed. Indicate success if all succeeded. Otherwise reset input as it was before the invocation of e_1 (do not consume anything) and indicate failure.
$e_1 / \dots / e_n$	Invoke expressions e_1, \dots, e_n , in this order, until one of them succeeds. Indicate success if one of expressions succeeded. Otherwise indicate failure.

The expressions e, e_1, \dots, e_n above can be specified either explicitly or by name (the way of naming expressions will be explained in a short while). An expression specified explicitly within another expression with the same or higher precedence (see below) must be enclosed in parentheses. The following table summarizes all forms of parsing expressions.

expression	name	precedence
" s "	String Literal	5
$[s]$	Character Class	5
$\sim [s]$	Not Character Class	5
$[c_1 - c_2]$	Character Range	5
$-$	Any Character	5
$e?$	Optional	4
e^*	Iterate	4
e^+	One or More	4
$\&e$	Ahead	3
$!e$	Not Ahead	3
$e_1 \dots e_n$	Sequence	2
$e_1 / \dots / e_n$	Choice	1

Backtracking takes place in the Sequence expression. If e_1, \dots, e_i in $e_1 \dots e_i \dots e_n$ succeed and consumed some input, but then e_{i+1} fails, the input is reset as it was before trying e_1 ; the whole expression fails. If this expression was invoked from a Choice expression (and was not the last there), Choice has an opportunity to try another alternative on the same input.

However, the opportunities to try another alternative are limited: once e_i in the Choice $e_1 / \dots / e_i / \dots / e_n$ succeeded, none of the alternatives e_{i+1}, \dots, e_n will ever be tried on the same input, even if the parse fails later on. This is the limited backtracking.

2.2 The grammar

Parsing Expression Grammar is a list of one or more "rules" of the form:

$$name = expr ;$$

where $expr$ is a parsing expression, and $name$ is a name given to it. The $name$ is a string of one or more letters (a-z, A-Z) and/or digits, starting with a letter. White space is allowed everywhere except inside names. Comments starting with a double slash and extending to the end of a line are also allowed.

The order of the rules does not matter, except that the expression specified first is the "start expression", invoked at the start of the parser.

A specific grammar may look like this:

```
Sum    = Number ("+" Number)* !_ ;
Number = [0-9]+ ;
```

It consists of two named expressions: `Sum` and `Number`. They define a parser consisting of two procedures named `Sum` and `Number`. The parser starts by invoking `Sum`. The `Sum` invokes `Number`, and if this succeeds, repeatedly invokes `("+" Number)` as long as it succeeds. Finally, `Sum` invokes a sub-procedure for the predicate `"!_"`, which succeeds only if it does not see any character ahead – that is, only at the end of input. The `Number` reads digits in the range from 0 through 9 as long as it succeeds, and is expected to find at least one such digit. One can easily see that the parser accepts strings like `"2+2"`, `"17+4711"`, or `"2"`; in general, one or more integers separated by `"+"`.

It is quite obvious that the expression e in e^* and e^+ must never succeed in consuming an empty string; otherwise, the parser could run into an infinite loop. *Mouse* checks this and gives you a warning if needed.

3 Getting started

For hands-on experience with *Mouse* you will need the executables and examples. All materials are packaged in the TAR file `Mouse-2.3.tar.gz`, available from sourceforge.net/projects/mousepeg. Download and unpack this file to obtain directory `Mouse-2.3`. The directory contains, among other items, JAR file `Mouse-2.3.jar` and sub-directory `examples`. Add a path to that JAR in the `CLASSPATH` system variable.

The `examples` directory contains ten directories named `example1` through `example10` that you will use in your exercise.

All the examples use unnamed packages and assume that you do all the compilation and execution in a current work directory. You may create such directory specifically for this purpose; it is in the following called `work`. You need, of course, to have "current work directory" in the `CLASSPATH`.

4 The first steps

We start with the sample grammar from Section 2.2. You find it as the file `myGrammar.txt` in `example1`. Copy it to the `work` directory, and enter at the command prompt:

```
java mouse.Generate -G myGrammar.txt -P myParser
```

This should produce the answer:

```
2 rules
1 inner
3 terminals
```

and generate in the `work` directory a file `myParser.java` containing class `myParser`. This is your parser.

If you open the file in a text editor, you will see that the parsing expressions `Sum` and `Number` have been transcribed into parsing procedures in the form of methods in the class `myParser`. The unnamed inner expression `("+" Number)` has been transcribed into the method `Sum_0`:


```

//=====
// Sum = Number ("+" Number)* !_ ;
//=====
private boolean Sum()
{
    begin("Sum");
    if (!Number()) return reject();
    while (Sum_0());
    if (!aheadNot()) return reject();
    return accept();
}

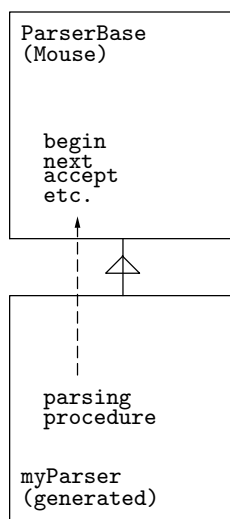
//-----
// Sum_0 = "+" Number
//-----
private boolean Sum_0()
{
    begin("");
    if (!next('+')) return rejectInner();
    if (!Number()) return rejectInner();
    return acceptInner();
}

//=====
// Number = [0-9]+ ;
//=====
private boolean Number()
{
    begin("Number");
    if (!nextIn('0','9')) return reject();
    while (nextIn('0','9'));
    return accept();
}
}

```

The methods `begin()`, `accept()`, `reject()`, `next()`, etc. are standard services provided by *Mouse*. The first three maintain, behind the scenes, the information needed to define semantics; it will be discussed in the next section. As you may guess, `accept()` returns *true* and `reject()` returns *false*. The methods `next()` and `nextIn()` implement terminals. For example, `next('+')` implements the parsing expression "+": if the next input character is +, consumes it and returns *true*, otherwise returns *false*.

As you can see by inspecting the file, class `myParser` is defined as a subclass of `mouse.runtime.ParserBase` provided by *Mouse*. The service methods are inherited from that superclass. The structure of your parser is thus as follows:



Compile your parser by typing:

```
javac myParser.java
```

To try the parser, type:

```
java mouse.TryParser -P myParser
```

This invokes a *Mouse* tool for running the parser. It responds with "> " and waits for input. If you enter a correct line, it prints another "> " to give you another chance. If your entry is incorrect, you get an error message before another "> ". You exit by pressing Enter just after "> ".

The session may look like this:

```
java mouse.TryParser -P myParser
> 1+2+3
> 17+4711
> 2#2
After '2': expected [0-9] or '+' or end of text
>
```

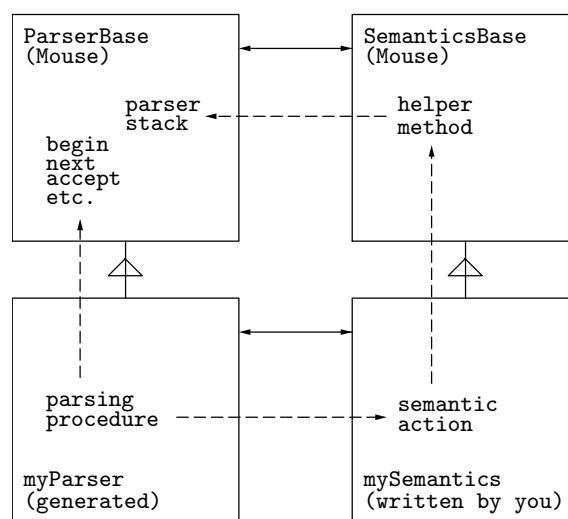
5 Adding semantics

This hasn't been very exciting. You certainly want the parser to do something more than checking the syntax. For example, to actually compute the sum. For this purpose, you attach "semantic actions" to rules of the grammar.

As you have seen, *Mouse* transcribes each rule of the grammar into a parsing procedure in the form of a Java method. A semantic action is a Java method, written by you, that is called by a successful parsing procedure as the last thing before it exits. You may use this method to process the text consumed by the procedure.

In order to compute the sums, you may define a semantic action for *Number* that will convert the digits consumed by *Number* into an integer. You may then define a semantic action for *Sum* that will receive these integers, add them, and print the result.

In most parser generators, the code for semantic actions is inserted directly into the grammar. *Mouse* is a departure from that tradition. You provide semantic actions as methods in a separate class with any name of your choice, for example "mySemantics". It must be a subclass of `mouse.runtime.SemanticsBase` provided by *Mouse*. The superclass `SemanticsBase` contains "helper methods" that give you access to results of the parsing. The structure of your parser together with semantic actions is as follows:



All four classes are instantiated by a constructor that is included in the generated parser. The constructor establishes references between the instantiated objects.

In the grammar, you indicate the presence of a semantic action by inserting its name in curly brackets at the end of a rule:

```
Sum    = Number ("+" Number)* !_ {sum} ;
Number = [0-9]+ {number} ;
```

This tells that Sum will have semantic action sum(), and Number will have semantic action number(). You find this grammar as myGrammar.txt in example2. Copy it into the work directory (replacing the old one), and type at the command prompt:

```
java mouse.Generate -G myGrammar.txt -P myParser -S mySemantics
```

This generates a new file myParser.java. By inspecting this file, you can see that parsing procedures Sum() and Number() contain calls to their semantic actions, as shown below. The qualifier sem is a reference to your semantics class mySemantics; you had to supply its name in order to generate this reference.

```
//=====
// Sum = Number ("+" Number)* !_ {sum} ;
//=====
private boolean Sum()
{
    begin("Sum");
    if (!Number()) return reject();
    while (Sum_0());
    if (!aheadNot()) return reject();
    sem.sum(); // semantic action
    return accept();
}
```

```
//=====
// Number = [0-9]+ {number} ;
//=====
private boolean Number()
{
    begin("Number");
    if (!nextIn('0','9')) return reject();
    while (nextIn('0','9'));
    sem.number(); // semantic action
    return accept();
}
```

Your semantics class is going to appear like this:

```
class mySemantics extends mouse.runtime.SemanticsBase
{
    //-----
    // Sum = Number ("+" Number)* !_
    //-----
    void sum()
    {...}

    //-----
    // Number = [0-9]+
    //-----
    void number()
    {...}
}
```

If you wish, you may generate a file containing the above skeleton at the same time as you generated your parser, by specifying the option `-s` on your command:

```
java mouse.Generate -G myGrammar.txt -P myParser -S mySemantics -s
```

It replies:

```
2 rules
1 inner
3 terminals
2 semantic procedures
```

indicating that it found two semantic actions, and produces the skeleton file `mySemantics.java` in the work directory, in addition to the parser file `myParser.java`.

It remains now to write the two methods. In order to do this, you need to know how to access the result of parsing.

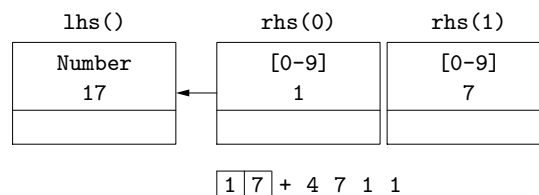
Behind the scenes, hidden in the service methods, each call to a parsing procedure, or to a service that implements a terminal, returns a `Phrase` object. The object is an instance of class `Phrase`. It describes the portion of input consumed by the call – the consumed "phrase".

Still behind the scenes, each parsing procedure collects the `Phrase` objects returned by the service methods and procedures that it calls. At the end, it uses information from them to construct its own resulting `Phrase`. The semantic action is called just after this latter has been constructed, and can access it, together with the `Phrases` representing the partial results.

As an illustration, we are going to describe what happens when your parser processes the input "17+4711".

The process starts by calling `Sum()`, which immediately calls `Number()`. `Number()`, in turn, makes two calls to `nextIn()` that consume, respectively, the digits "1" and "7", and return one `Phrase` object each. `Number()` (or, more precisely, the service methods called in the process) uses information from these objects to construct a new `Phrase` object that represents the text "17" thus consumed. This is going to be the result of `Number()`.

All three objects are accessible to the semantic action `number()`, and can be visualized as follows:



Each box represents one `Phrase` object. The text highest in the box identifies the parsing procedure or the service that constructed it. The text below is the consumed portion of input.

In the semantic action, you obtain references to the three objects by calls to helper methods shown above the boxes. They should be read as "left-hand side", "right-hand side element 0", and "right-hand side element 1": they correspond to the left- and right-hand sides of the rule "Number = [0-9]+".

According to the initial plan, `number()` is expected to convert the text "17" consumed by `Number()` into the integer 17. This could be done by obtaining the digits represented by `rhs(0)` and `rhs(1)` and processing them to obtain the integer. However, the integer can be obtained directly from the text represented by `lhs()`, by using a method from the Java class `Integer`. To obtain the text represented by a `Phrase` object, you apply the helper method `text()` to the object. The integer 17 is thus obtained as `Integer.valueOf(lhs().text())`.

Each `Phrase` object has a field of type `Object` where you can insert the "semantic value" of the represented phrase. It is shown here as an empty slot at the bottom of each box. You can insert there the integer 17 as semantic value of the result. You insert semantic value into a `Phrase` object by applying the helper method `put(...)` to that object, so your `number()` can be written as follows:

```

//-----
// Number = [0-9]+
//-----
void number()
{ lhs().put(Integer.valueOf(lhs().text())); }

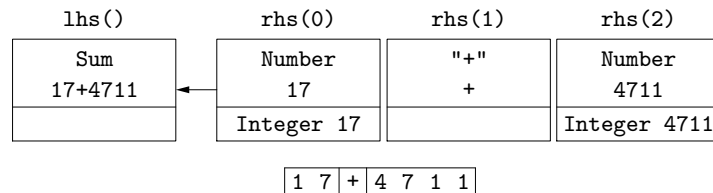
```

After completed semantic action, `Number()` returns to `Sum()`, delivering this object as its result:

Number
17
Integer 17

After receiving control back, `Sum()` proceeds, via `Sum_0()`, to call `next()` and `Number()` that consume, respectively, "+" and "4711", and return their resulting `Phrase` objects. `Sum()` uses information from them to construct a new `Phrase` object that represents the text "17+4711" thus consumed.

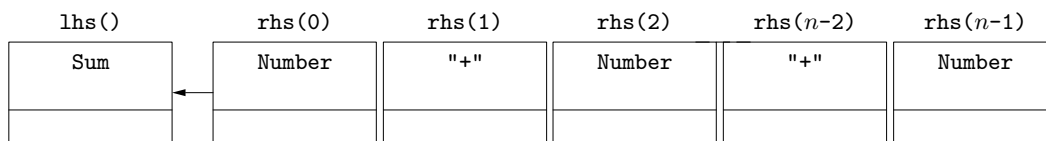
All four objects are accessible to the semantic action `sum()`, and can be visualized as follows:



Both `Number()` results carry as semantic values the integers computed by `number()`. To obtain the semantic value of a `Phrase` you apply the helper method `get()` to it. You can now easily compute the sum as follows (you need the casts because `get()` returns an `Object`):

```
(Integer)rhs(0).get() + (Integer)rhs(2).get();
```

However, in the general case you may have any number $p \geq 0$ of ("+" `Number`) pairs resulting from p successful calls to `Sum_0()`. Your `sum()` must thus handle the following general situation with $n = 2p + 1$ `Phrases` on the right-hand side:



To obtain the number n you use the helper method `rhsSize()`. Your method `sum()` may thus look like this:

```

//-----
// Sum = Number "+" Number ... "+" Number !_
//         0   1  2     n-2 n-1
//-----
void sum()
{
    int s = 0;
    for (int i=0;i<rhsSize();i+=2)
        s += (Integer)rhs(i).get();
    System.out.println(s);
}

```

You find the complete semantics class, with the methods `number()` and `sum()` just described, as file `mySemantics.java` in `example2`; copy it to the work directory. To obtain the new parser, compile the semantics class and the parser class you have generated. You can try it in the same way as before. The session may look like this:

```

javac myParser.java
javac mySemantics.java
java mouse.TryParser -P myParser
> 17+4711
4728
> 12+2+3
17
>

```

6 Understanding the "right-hand side"

As can be seen from the above example, you need to know the exact appearance of the "right-hand side" in order to correctly write your semantic action. Your method must be prepared to handle a varying configuration that depends on the parsed text. To have this configuration in front of you while writing the code, it is convenient to describe it by a comment as shown above. (The author has learned this style from his colleague Bertil Steinholtz, who used it to write very clear yacc++ code.)

The objects appearing on the "right-hand side" correspond to the names and terminals appearing to the right of "=" in the rule. You can find the possible configurations of the "right-hand side" like this:

Take the expression to the right of "=" and remove from it all sub-expressions of the form $&e$ and $!e$. Replace also each "/" by "|". The result is a regular expression on an alphabet consisting of names and terminals treated as single letters. The possible configurations of the "right-hand side" are exactly the strings defined by this regular expression.

As an example, the regular expression obtained for Sum is:

```
Number ("+" Number)*
```

which indeed defines these strings of symbols Number and "+":

```
Number "+" Number ... "+" Number
 0      1      2      n-2 n-1
```

where $n = 2p + 1$ for $p = 0, 1, 2, \dots$

7 Getting more realistic

Our example so far is not very realistic. Any decent parser will allow white space in the input. This is simply achieved in PEG by defining Space as a sequence of zero or more blanks, and inserting Space in strategic places. We are going to do this, and on this occasion add two features: an optional minus sign in front of the first Number, and an alternative minus between the Numbers:

```
Sum      = Space Sign Number (AddOp Number)* !_ {sum} ;
Number   = Digits Space {number} ;
Sign     = ("-" Space)? ;
AddOp    = ["+"] Space ;
Digits   = [0-9]+ ;
Space    = " "* ;
```

The semantic action number() will now see this configuration of the "right-hand side":

```
Digits Space
 0      1
```

The text to be converted is now represented by rhs(0) rather than lhs(), so number() appears as follows:

```
//-----
// Number = Digits Space
//           0      1
//-----
void number()
{ lhs().put(Integer.valueOf(rhs(0).text())); }
```

Note that we could have defined Number as $[0-9]^+ \text{Space}$, but this would present number() with a "right-hand side" of variable length, and give it a job of collecting all the digits. It is much easier to let the parser do the job.

The semantic action `sum()` will see this configuration of the "right-hand side":

```
Space Sign Number AddOp Number ... AddOp Number
  0     1     2     3     4         n-2  n-1
```

where $n = 2p + 3$ for $p = 0, 1, 2, \dots$

The text for `Sign` is either an empty string or a minus followed by zero or more blanks. You can check which is the case with the helper method `isEmpty()`. The text for `AddOp` is either a plus or a minus followed by zero or more blanks. You can obtain the first character of this text with the helper method `charAt()`. This gives the following new version of `sum()`:

```
//-----
// Sum = Space Sign Number AddOp Number ... AddOp Number
//           0     1     2     3     4         n-2  n-1
//-----
void sum()
{
    int n = rhsSize();
    int s = (Integer)rhs(2).get();
    if (!rhs(1).isEmpty()) s = -s;
    for (int i=4; i<n; i+=2)
    {
        if (rhs(i-1).charAt(0)=='+')
            s += (Integer)rhs(i).get();
        else
            s -= (Integer)rhs(i).get();
    }
    System.out.println(s);
}
```

Note again that we could have `("-" Space)?` instead of `Sign` in the definition of `Sum`. But, according to the rule given in Section 6, `sum()` would then have to handle two different configuration of the "right-hand side": one starting with `(Space "-" Space Number...)` and one starting with `(Space Number...)`. In our definition of `Sum`, `Sign` is always there, even if the underlying `("-" Space)` failed. Note that `Space` is also present even if it did not consume any blanks.

The above illustrates the fact that you have to exercise a lot of care when defining rules expected to have semantic actions. Introducing `Digits` and `Sign` as separate named expressions made the "right-hand side" in each case easier to handle. For reasons that should now be obvious, you are not recommended to have repetitions or optional expressions inside repetitions. A clever programming of semantic action can always handle them, but the job is much better done by the parser.

You find the grammar and the semantics class for this example as `myGrammar.txt` and `mySemantics.java` in `example3`. Copy them to the work directory (replacing the old ones), generate the parser, and compile it as before. A session with the new parser may look like this:

```
java mouse.TryParser -P myParser
> 17 + 4711
4728
> -1 + 2 - 3
-2
>
```

To close this section, note that you implemented here some functions traditionally delegated to a separate "lexer": conversion of integers and ignoring white space. If you wanted, you could easily include newlines and comments in `Space`.

8 Let's go floating

To illustrate some more features of *Mouse*, we shall add floating-point numbers to your primitive calculator. The calculations will now be done on numbers in `double` form. The only thing needed in the grammar is the possibility to specify such numbers. We decide for a format with one or more digits after the decimal point, and optional digits before: `Digits? "." Digits`. We are going to keep the integer format as an alternative, which means defining `Number` as

```
Number = Digits? "." Digits Space / Digits Space ;
```

The two alternatives produce two quite different configurations of the "right-hand side". The semantic action would have to begin by identifying them. But, this has already been done by the parser. To exploit that, you specify a separate semantic action for each alternative in a rule that defines choice expression. It means modifying your grammar as follows:

```
Sum      = Space Sign Number (AddOp Number)* !_ {sum} ;
Number   = Digits? "." Digits Space {fraction}
          / Digits Space {integer} ;
Sign     = ("-" Space)? ;
AddOp    = ["+"] Space ;
Digits   = [0-9]+ ;
Space    = " "* ;
```

You specify here two semantic actions, `fraction()` and `integer()`, for the two alternatives of `Number`.

The first of them is invoked to handle input of the form `Digits? "." Digits Space`. According to the rule given in Section 6, it can result in two possible configurations of the "right-hand side":

```
    Digits "." Digits Space      or      "." Digits Space
      0     1     2     3          0     1     2
```

You need to convert the text that represents the floating-point number into a `Double` object. You can obtain this text by using the helper method `rhsText(i,j)`. The method returns the text identified by the right-hand side elements i through $j - 1$. As you can easily see, `rhsText(0,rhsSize()-1)` will do the job for each of the two configurations, so `fraction()` can be coded as follows:

```
//-----
// Number = Digits? "." Digits Space
//           0     1(0) 2(1) 3(2)
//-----
void fraction()
{ lhs().put(Double.valueOf(rhsText(0,rhsSize()-1))); }
```

As computation will now be done in floating point, `integer()` should also return a `Double`:

```
//-----
// Number = Digits Space
//           0     1
//-----
void integer()
{ lhs().put(Double.valueOf(rhs(0).text())); }
```

The semantic action for `Sum` remains unchanged, except that `s` becomes `double` and the casts are to `Double` instead of `Integer`.

You may copy the new grammar and semantics from `example4`, generate new parser (replacing the old), and compile both. A session with the new parser may look like this:

```
java mouse.TryParser -P myParser
> -123 + 456
333.0
> 1.23 - 4.56
-3.3299999999999996
>
```

(Who said floating-point arithmetic is exact?)

9 What about backtracking?

As you can easily see, `Number` in the last grammar does not satisfy the classical $LL(1)$ condition: seeing a digit as the first character, you do not know which alternative it belongs to. Presented with `"123"`, the parser will start with the first alternative of `Number` (`"fraction"`) and process `"123"` as its `Digits`. Not finding the decimal point, the parser will backtrack and try the other alternative (`"integer"`), again processing the same input as `Digits`. This re-processing of input is the price for not bothering to make the grammar $LL(1)$.

The loss of performance caused by reprocessing is uninteresting in this particular application. However, circumventing $LL(1)$ in this way may cause a more serious problem. The reason is the limited backtracking of PEG. As indicated in Section 2.1, if the parser successfully accepts the first alternative of e_1/e_2 and fails later on, it will never return to try e_2 . As a result, some part of the language of e_2 may never be recognized.

Just imagine for a while that the two alternatives in `Number` are reversed. Presented with `"123.45"`, `Number` consumes `"123"` as an `"integer"` and returns to `Sum`, reporting success. The `Sum` finds a point instead of `AddOp`, and terminates prematurely. The other alternative of `Number` is never tried. All fractional numbers starting with `Digits` are hidden by greedy `"integer"`.

Your rule for `Number` is almost identical to the following production in Extended Backus-Naur Form (EBNF):

```
Number ::= Digits? "." Digits Space | Digits Space
```

and you certainly expect that the language accepted by your PEG rule is exactly one defined by this production. This is indeed true in this case. A sufficient condition for PEG rule $A = e_1/e_2$ defining the same language as EBNF production $A ::= e_1|e_2$ is:

$$\mathcal{L}(e_1)\Sigma^* \cap \mathcal{L}(e_2)\text{Tail}(A) = \emptyset, \quad (*)$$

where $\mathcal{L}(e)$ is the language defined by e (in EBNF), Σ^* is the set of all strings, and $\text{Tail}(A)$ is the set of all strings that can follow A in correct input¹.

You can see that in this case $\text{Tail}(A)$ is $\mathcal{L}((\text{AddOp } \text{Number})^*)$, so the condition becomes:

$$\mathcal{L}(\text{Digits? "." Digits Space})\Sigma^* \cap \mathcal{L}(\text{Digits Space } (\text{AddOp } \text{Number})^*) = \emptyset.$$

It is obviously satisfied as every string to the left of the intersection operator \cap has a decimal point in front of `Digits`, and none of them to the right does. If you reverse the alternatives in `Number`, the condition becomes:

$$\mathcal{L}(\text{Digits Space})\Sigma^* \cap \mathcal{L}(\text{Digits? "." Digits Space } (\text{AddOp } \text{Number})^*) = \emptyset,$$

which obviously does not hold.

There is no general method to check (*). However, you can often verify it by inspection as this was done above. The *Mouse* package includes an interactive tool, the *PEG Explorer*, to facilitate such inspection. It is described in <http://mousepeg.sourceforge.net/explorer.htm>.

You should also note that (*) always holds if A satisfies $LL(1)$.

To watch the backtracking activity of your parser, you may generate an instrumented version of it. To do this, type these commands:

```
java mouse.Generate -G myGrammar.txt -P myParser -S mySemantics -T
javac myParser.java
```

The option `-T` instructs the Generator to construct a `"test version"`. (You may choose another name for this version if you wish, the semantics class remains the same.)

¹This condition was formally proved in [17] using an approach invented by Medeiros [12].

Invoking the test version with `mouse.TryParser` produces the same results as before. In order to exploit the instrumentation, you have to use `mouse.TestParser`; the session may look like this:

```
java mouse.TestParser -P myParser
> 123 + 4567
4690.0

50 calls: 33 ok, 15 failed, 2 backtracked.
11 rescanned.
backtrack length: max 4, average 3.5.

>
```

This output tells you that to process your input "123 + 4567", the parser executed 50 calls to parsing procedures, of which 33 succeeded, 15 failed, and two backtracked. (We treat here the services that implement terminals as parsing procedures.) As expected, the parser backtracked 3 characters on the first `Number` and 4 on the second, so the maximum backtrack length was 4 and the average backtrack length was 3.5. You can also see that 11 of the procedure calls were "re-scans": the same procedure called again at the same input position. You can get more detail by specifying the option `-d`:

```
java mouse.TestParser -P myParser -d
> 123 + 4567
4690.0

50 calls: 33 ok, 15 failed, 2 backtracked.
11 rescanned.
backtrack length: max 4, average 3.5.

Backtracking, rescan, reuse:

procedure      ok  fail  back  resc  reuse  totbk  maxbk  at
-----
Number_0        0    0    2     0    0     7     4  After '123 + '
Digits          4    0    0     2    0     0     0    0
[0-9]          14    4    0     9    0     0     0    0

>
```

You see here statistics for individual procedures that were involved in backtracking and rescanning. `Number_0` is the internal procedure for the first alternative of `Number`. As you can guess, "totbk" stands for total backtrack length and "maxbk" for length of the longest backtrack; "at" tells where this longest backtrack occurred. The meaning of "reuse" will be clear in a short while.

10 A mouse, not a pack rat

As mentioned in the introduction, *Mouse* can provide a small degree of memoization. You can exercise this possibility by specifying the option `-m` to `TestParser`. You follow it by a digit from 1 to 9, indicating how many most recent results you want to keep. For example, your session from the previous section, with procedures caching one result, will appear like this:

```
java mouse.TestParser -PmyParser -d -m1
> 123 + 4567
4690.0

41 calls: 24 ok, 13 failed, 2 backtracked.
0 rescanned, 2 reused.
backtrack length: max 4, average 3.5.

Backtracking, rescan, reuse:

procedure      ok  fail  back  resc  reuse  totbk  maxbk  at
-----
Number_0        0    0    2     0    0     7     4  After '123 + '
Digits          2    0    0     0    2     0     0    0

>
```

You can see here that the parser reused the cached result of `Digits` on two occasions, thus eliminating the unnecessary rescanning by `[0-9]`.

If you decide you want memoization, you can generate a version of the parser that allows it, without all the instrumentation overhead. You do it by specifying the option `-M` on `mouse.Generate`. You specify then the amount of memoization by means of `-m` option in the same way as above.

When deciding whether you want memoization or not, you should consider the fact that it introduces some overhead. It may cost more in performance than some moderate rescanning. In the current version of *Mouse*, memoization is not applied to the services for terminals because the overhead of memoization was felt to be larger than the job of rescanning a terminal.

11 Full arithmetic

As the next development, we suggest adding multiplication, division and parentheses:

```

Input  = Space Sum !_ {result} ;
Sum    = Sign Product (AddOp Product)* {sum} ;
Product = Factor (MultOp Factor)* {product} ;
Factor = Digits? "." Digits Space {fraction}
        / Digits Space {integer}
        / Lparen Sum Rparen {unwrap}
        / Name Space {retrieve} ;
Sign   = ("-" Space)? ;
AddOp  = ["+"] Space ;
MultOp = (["*/"] Space)? ;
Lparen = "(" Space ;
Rparen = ")" Space ;
Digits = [0-9]+ ;
Space  = " " * ;

```

We have replaced here `Number` by `Factor` that can be a fraction, an integer, or a sum in parentheses.

We have added `Product` with multiplication and division to ensure, in a syntactical way, the priority of these operations over addition and subtraction. The omitted `MultOp` is intended to mean multiplication.

As `Sum` can now appear in `Factor`, we had to add `Input` as new top level, moving there the initial `Space` and the check for end of input.

The semantic action `sum()` cannot now print its result; instead, the result will be inserted as semantic value into the `Sum`'s result, to be retrieved by the semantic action of `Input` or `Factor`:

```

//-----
// Sum = Sign Product AddOp Product ... AddOp Product
//           0       1       2       3       n-2   n-1
//-----
void sum()
{
    int n = rhsSize();
    double s = (Double)rhs(1).get();
    if (!rhs(0).isEmpty()) s = -s;
    for (int i=3; i<n; i+=2)
    {
        if (rhs(i-1).charAt(0)=='+')
            s += (Double)rhs(i).get();
        else
            s -= (Double)rhs(i).get();
    }
    lhs().put(Double.valueOf(s));
}

```

This assumes, of course, that `Product` delivers its result as semantic value of its result. The semantic action `product()` follows the same pattern as `sum()`:

```
//-----
// Product = Factor MultOp Factor ... MultOp Factor
//           0      1      2      n-2    n-1
//-----
void product()
{
    int n = rhsSize();
    double s = (Double)rhs(0).get();
    for (int i=2;i<n;i+=2)
    {
        if (!rhs(i-1).isEmpty() && rhs(i-1).charAt(0)=='/')
            s /= (Double)rhs(i).get();
        else
            s *= (Double)rhs(i).get();
    }
    lhs().put(Double.valueOf(s));
}
```

(`MultOp` means division if it is a nonempty string starting with with `/`; otherwise it is multiplication.)

Note that the precedence of multiplication and division over addition and subtraction is here defined by the syntax. However, the order of operations within each class (left to right) is *not* defined by the syntax, but by semantic actions.

The actions `fraction()` and `integer()` remain the same as in the preceding example. The function of `unwrap()` is to take the semantic value of `Sum` and deliver it as the semantic value of `Factor`:

```
//-----
// Factor = Lparen Sum Rparen
//           0      1      2
//-----
void unwrap()
{ lhs().put(rhs(1).get()); }
```

No casts are needed. It remains to print the result:

```
//-----
// Input = Space Sum !_
//           0      1
//-----
void result()
{ System.out.println((Double)rhs(1).get()); }
```

The modified grammar and semantics are available in `example5`; you may copy them into the work directory, then generate and compile the parser as before. The session may now look like this:

```
java mouse.TryParser -P myParser
> 2(1/10)
0.2
> (1+2)*(3-4)
-3.0
> 1.23(-4+.56)
-4.2312
>
```

Notice that the multiplication operator can, in particular, be the empty string. This would cause an ambiguity in the classical context-free grammar, where `"234"` could be parsed as meaning `2*3*4`, `23*4`, or `2*34`. However, there is no place for ambiguity in the parser specified by PEG: it will always read the string as integer `234`. Which is, by the way, exactly how a human reader sees it.

12 Want a tree?

Many existing parsers deliver their result in the form of a syntax tree, showing how the input was parsed according to the grammar rules. The preceding examples show that much of this information may be uninteresting if the grammar is specified at the level of input characters rather than "tokens" produced by a "lexer". For example, the structure of our fractional number in terms of individual characters is not interesting as long as we only want the complete string. For this reason, *Mouse* does not provide automatic construction of syntax trees.

If you wish, you may still construct syntax tree at the desired level of detail with the help of semantic actions. In the following, we show how the preceding example can be modified to produce an operand-operator tree instead of calculating the result. Such tree can be then further processed (for example, to generate executable code). All you need to construct the tree is write another semantics class, without changing the parser.

Of course, you also need to define elements of the tree to be constructed. You find their definitions in the file `Node.java` in `example6`. It defines an abstract class `Node` with two subclasses: `Node.Num` and `Node.Op`, to represent, respectively, a number and a binary operator. In order to simplify the example, we do not provide separate subclasses for different kinds of numbers and operators. A number is stored in `Node.Num` as the text string retrieved from the input; an operator is stored in `Node.Op` as one character.

The semantic values will now be partial trees, represented by their top nodes. The semantic actions will appear like this:

```
//-----  
// Factor = Digits? "." Digits Space  
//           0   1(0) 2(1) 3(2)  
//-----  
void fraction()  
    { lhs().put(new Node.Num(rhsText(0,rhsSize()-1))); }
```

```
//-----  
// Sum = Sign Product AddOp Product ... AddOp Product  
//       0     1     2     3     n-2     n-1  
//-----  
void sum()  
    {  
        int n = rhsSize();  
        Node N = (Node)rhs(1).get();  
        if (!rhs(0).isEmpty())  
            N = new Node.Op(new Node.Num("0"), '-', N);  
        for (int i=3; i<n; i+=2)  
            N = new Node.Op(N, rhs(i-1).charAt(0), (Node)rhs(i).get());  
        lhs().put(N);  
    }
```

(Minus as `Sign` is represented by subtraction from 0.)

You find complete semantics in `example6`. As you can see there, the final action, `result()`, saves the top node of the resulting tree in the variable `tree`, local to `mySemantics`. You cannot use the tool `mouse.TryParser` to obtain the tree thus stored; you have to code invocation of the parser on your own. It will be an occasion to learn how this is done.

To use the parser you need first to instantiate it, for example like this:

```
myParser parser = new myParser();
```

This instantiates both `myParser` and `mySemantics`. The string to be parsed has to be wrapped in an object of class `SourceString` supplied by *Mouse*:

```
SourceString src = new SourceString(line);
```

which wraps the string `line`. Finally, you invoke the parser like this:

```
boolean ok = parser.parse(src);
```

which sets `ok` to `true` or `false` depending on the result. To access `mySemantics` object you invoke the method `semantics()` on the parser.

A complete code to parse the lines that you type is shown below.

```
import mouse.runtime.SourceString;
import java.io.BufferedReader;
import java.io.InputStreamReader;

class Try
{
    public static void main(String argv[])
        throws Exception
    {
        BufferedReader in = new BufferedReader(new InputStreamReader(System.in));
        myParser parser = new myParser();           // Instantiate Parser+Semantics
        while (true)
        {
            System.out.print("> ");                // Print prompt
            String line = in.readLine();            // Read line
            if (line.length()==0) return;          // Quit on empty line
            SourceString src = new SourceString(line); // Wrap up the line
            boolean ok = parser.parse(src);         // Apply Parser to it
            if (ok)                                 // If succeeded:
            {
                mySemantics sem = parser.semantics(); // Access Semantics
                System.out.println(sem.tree.show()); // Get the tree and show it
            }
        }
    }
}
```

The above code is found in `example6` under the name `Try.java`. Copy the files `Node.java`, `mySemantics.java`, and `Try.java` into the work directory and compile them. You may now type `java Try` at the command prompt, and enter your input after `>`. You will get the resulting tree as a nested expression where each operator node with its sub-nodes is enclosed by a pair of brackets `[]`. Your session may now look like this:

```
java Try
> 1+2+3
[[[1+2]+3]
> 2(1/10)
[2*[1/10]]
> (1+2)*(3-4)
[[[1+2]*[3-4]]
> 1.23(-4+.56)
[1.23*[[0-4]+.56]]
>
```

This is not the most elegant presentation of a tree, but the purpose is just to check what was constructed.

13 Calculator with memory

Back to the calculator from Section 11. We suggest one more improvement: the ability to store computed values for subsequent use. The idea is that a session may look like this:

```
java Calc
> n = 3/2
> n = (3/n + n)/2
> (3/n + n)/2
1.7321428571428572
```

In the first line we store the number $3/2$ under name `n`. In the second, we replace it by $(3/n + n)/2$, computed using the stored value. Finally, we use the newest `n` to compute one more value and print it. (Is it just an accident that the result resembles $\sqrt{3}$?)

To achieve this kind of functionality, you can modify the grammar as follows:

```

Input  = Space (Store / Print) !_ ;
Store  = Name Space Equ Sum {store} ;
Print  = Sum {print} ;
Sum    = Sign Product (AddOp Product)* {sum} ;
Product = Factor (MultOp Factor)* {product} ;
Factor = Digits? "." Digits Space {fraction}
        / Digits Space {integer}
        / Lparen Sum Rparen {unwrap}
        / Name Space {retrieve} ;
Sign   = ("-" Space)? ;
AddOp  = ["+"] Space ;
MultOp = (["*/"] Space)? ;
Lparen = "(" Space ;
Rparen = ")" Space ;
Equ    = "=" Space ;
Digits = [0-9]+ ;
Name   = [a-z]+ ;
Space  = " " * ;

```

We have here two alternative forms of Input to perform the two tasks, and one more way to obtain a Factor: by retrieving a stored result. The results will be stored in a hash table dictionary that resides in the semantics object.

The semantic actions for Store and Print are:

```

//-----
// Store = Name Space Equ Sum
//          0     1     2     3
//-----
void store()
{ dictionary.put(rhs(0).text(), (Double)rhs(3).get()); }

```

```

//-----
// Print = Sum
//          0
//-----
void print()
{ System.out.println((Double)rhs(0).get()); }

```

The stored result is retrieved by semantic action for the fourth alternative of Factor:

```

//-----
// Factor = Name Space
//          0     1
//-----
void retrieve()
{
    String name = rhs(0).text();
    Double d = dictionary.get(name);
    if (d==null)
    {
        d = (Double)Double.NaN;
        System.out.println
            (rhs(0).where(0) + ": " + name + " is not defined.");
    }
    lhs().put(d);
}

```

You have to handle the case where the name is not in the dictionary. The example shown here returns in such a case a NaN (Not a Number) object. Using NaN in subsequent calculations produces inevitably a NaN result. A message printed to System.out explains where this NaN comes from. Note the call to rhs(0).where(0) that returns a text telling where to find the name in the input.

You find the modified grammar and semantics in example7. You find there also a file Calc.java containing invocation of the calculator. Copy these files to work directory, generate a new parser, and compile everything. You can now invoke the calculator as shown at the beginning of this section.

14 Get error handling right

Try the following with the calculator just constructed:

```
java Calc
> 2a
After '2': 'a' is not defined
NaN
> a =
At start: 'a' is not defined
NaN
After 'a =': expected ' ' or Sum
> a = 12(
After 'a = 12(': expected ' ' or Sum
> a
12.0
> 2a)
24.0
After '2a': expected [a-z] or ' ' or [*/] or Factor or AddOp or end of text
>
```

Note that the syntactically incorrect input "a =" resulted "not defined" message and "NaN" before the message about syntax error.

Note also that syntactically incorrect input "a = 12(" actually entered "a" in the dictionary, which is shown by response to the subsequent "a".

Another syntactically incorrect input "2a)" printed the result of "2a" before the message about syntax error.

What happens?

In the first case, Store failed on "a =" and backtracked. The other alternative, Print, succeeded in consuming "a ". Its sub-procedure Factor, through semantic action retrieve() printed the "not defined" message, and produced the result "NaN" that was eventually printed by print(). Then, Print returned success to Input, which failed by finding "=" instead end of text, backtracked, and issued the "expected" message.

In the second case, Store successfully consumed "a = 12", and its action store() entered "a" into the dictionary with 12 as value. Then, Store returned success to Input, which failed by finding "(" instead of end of text, backtracked, and issued the error message.

In the third case, Print successfully consumed "2a", printed its result, and returned success to Input, which failed by not finding end of text.

All these confusing effects are caused by premature printing of a message or call to semantic action in an uncertain situation that can be changed by backtracking. The moral of the story is that **you must not take any irreversible actions in a situation that can be reversed by backtracking**.

The problem of premature calls to semantic actions can be solved by moving the predicate "!" from Input to Store and Print:

```
Input  = Space (Store / Print) ;
Store  = Name Space Equ Sum !_ {store} ;
Print  = Sum !_ {print} ;
```

In this way, store() and print() are called only after the entire input has been successfully parsed.

Another irreversible action here is the printing of "not defined" message. As soon as Print (via its sub-procedures) encounters an undefined "a", it issues the message, even if it is going to fail later on. This can be solved using the "deferred action" feature, described in Section 20.5. A deferred action is any lambda-expression without parameters and with void result. It is a chunk of code that can be stored for later execution. In this case, it can be, for example:

```
()->{System.out.println("At start: 'a' is not defined")}
```

Using the helper method actAdd(), you can save it in the Phrase object of your parsing procedure. If your procedure is successful, it is passed on to the Phrase object of its caller, and further on, as long as the procedures

are successful. The actions that made it to the end of the parse are executed then. If any procedure in the chain fails, the deferred action is discarded together with the Phrase object containing it.

To exploit this feature, the semantic action `retrieve()` can be changed as follows:

```

//-----
// Factor = Name Space
//           0     1
//-----
void retrieve()
{
    String name = rhs(0).text();
    Double d = dictionary.get(name);
    if (d==null)
    {
        d = (Double)Double.NaN;
        String msg = rhs(0).where(0) + ": '" + name + "' is not defined";
        lhs().actAdd(()->{System.out.println(msg);});
    }
    lhs().put(d);
}

```

(Note. You may choose to handle undefined name in an entirely different way, as a syntax error. This could be done by defining `retrieve()` as a boolean action, a feature described in Section 20.4.)

Looking at smaller details, you may wonder why the message after `"a = 12("` says nothing about the expected end of text.

A non-backtracking parser stops after failing to find an expected character in the input text, and this failure is reported as *the* syntax error. A backtracking parser may instead backtrack and fail several times. It terminates and reports failure when no more alternatives are left. The strategy used by *Mouse* is to report only the failure that occurred farthest down in the input. If several different attempts failed at the same point, all such failures are reported.

In the example, the message says that `Lparen` expected blanks after `"("`, and that the third alternative of `Factor` expected `Sum` at that place. The failure to find end of text occurred earlier in the input and is not mentioned.

In the case of `"2a)"`, several procedures failed immediately after `"2a"`. The processing did not advance beyond that point, so all these failures are reported.

You may feel that information `"expected ' ' "` is uninteresting noise. It results from the iteration in `Space` terminated by not finding another blank. To suppress it, you can define the following semantic action for `Space`:

```

//-----
// Space = " "*
//-----
void space()
{ lhs().errClear(); }

```

When `space()` is invoked, the `lhs()` object seen by it contains the information that `Space()` ended by not finding another blank. The helper method `errClear()` erases all failure information from the `lhs()` object, so `Space` will never report any failure.

If you wish, you can make messages more readable by saying `"+"` or `"-"` instead of `"AddOp"` and `"*"` or `"/"` instead of `"[*/]"`. You specify such "diagnostic names" in pointed brackets at the end of a rule, like this:

```

AddOp   = [-+] Space <+ or -> ;
MultOp  = MulDiv? ;
MulDiv  = [*/] Space <* or /> ;
Lparen  = "(" Space <(> ;
Rparen  = ")" Space <)> ;
Equ     = "=" Space <=> ;

```

Note that it was necessary to introduce a name for `[*/]`.

A grammar and semantics with these modifications are found in `example8`. Copy them to the work directory, generate new parser (replacing the old), and compile both. A parser session may now appear like this:

```

java Calc
> 2a
NaN
After '2': 'a' is not defined
> a =
After 'a '=': expected Sum
> a = 12(
After 'a = 12(': expected Sum
> a
NaN
At start: 'a' is not defined
> 2a)
After '2a)': expected [a-z] or * or / or Factor or + or - or end of text
>

```

15 Input from file

In a more serious applications, like a compiler, the parser has to read input from a file. We are going to show the way of doing this by an adaptation of the calculator developed above.

Just suppose you want to perform a sequence of calculations such as this:

```

n = 0
f = 1
e = 1
e
n = n + 1
f = f * n
e = e + 1/f
e
n = n + 1
f = f * n
e = e + 1/f
e
...

```

that will print the consecutive stages of computing the number e from its series expansion. You want the calculator to read this input from a file. All you need is to enlarge your syntax by defining the input as a sequence of Lines:

```

Input = [\n\r]* Line+ EOF ;
Line = Space (Store / Print) ;
Store = Name Space Equ Sum EOL {store} ;
Print = Sum EOL {print} ;

... as before ...

EOL = [\n\r]+ / EOF <end of line> ;
EOF = !_ <end of file> ;

```

This grammar defines EOL to be either end of line or end of input. End of line is defined as one or more newlines and/or carriage returns, meaning that the syntax allows empty lines – really empty, without even space characters in them. (The "`\n`" and "`\r`" stand for newline and carriage return characters, respectively; see Section 20.1.) The `[\n\r]*` at the beginning permits the file to start with empty lines.

You do not need to change semantics.

Copy `myGrammar.txt` and sample file `e.txt` from `example9`, generate the parser and compile it in the usual way.

To run the parser on the sample file, type at command prompt:

```

java mouse.TryParser -P myParser -f e.txt

```

The option `-f e.txt` instructs the tool to take its input from the file `e.txt`. You should get this output:

```
e.txt
1.0
2.0
2.5
2.6666666666666665
2.7083333333333333
2.7166666666666663
2.7180555555555554
2.7182539682539684
2.71827876984127
2.7182815255731922
2.7182818011463845
--- ok.
```

The *Mouse* tool `TryParser` is provided to help you in developing the parser, but eventually you will need to invoke the parser from one of your own programs. You do it as described in Section 12, with the difference that the input file is wrapped in class `SourceFile` supplied by *Mouse*. The complete code may look like this:

```
import mouse.runtime.SourceFile;
class FileCalc
{
    public static void main(String argv[])
    {
        myParser parser = new myParser();           // Instantiate Parser+Semantics
        SourceFile src = new SourceFile(argv[0]);    // Wrap the file
        if (!src.created()) return;                 // Return if no file
        boolean ok = parser.parse(src);             // Apply parser to it
        System.out.println(ok? "ok":"failed");      // Tell if succeeded
    }
}
```

You find it as `FileCalc.java` in `example9`. Copy it into the work directory and compile. You can then invoke the calculation by typing:

```
java FileCalc e.txt
```

To see the kind of diagnostics produced for a file, damage `e.txt` by, for example, duplicating one of the operators or changing `"n"` to `"x"` in an expression, and run the parser on the damaged file.

Note that the "not defined" messages come directly after the line to which they apply. This is the effect of invoking helper method `actExec()` in `print()` and `store()` to execute deferred actions.

16 Error recovery

If you tried the parser on a faulty file, you have noticed that it stops after encountering the first faulty line. It perhaps does not make sense for this particular computation, but in many cases you would like the parser to just skip the faulty line and continue with the next. For example, if you use the parser to implement a compiler, you want it to continue and possibly discover more errors. Such behavior can be achieved by defining a faulty line as an additional alternative for `Line`, to be tried after `Store` and `Print` failed. With this addition, the first lines of your grammar would look like this:

```
Input  = [\n\r]* (!EOF Line)+ EOF ;
Line   = Space (Store / Print) ~{badLine}
       / BadLine ;
Store  = Name Space Equ Sum EOL {store} ;    // store result
Print  = Sum EOL {print} ;                  // print result
BadLine = (!EOL _)+ EOL ;                   // skip bad line
```

The `BadLine` is defined to be any sequence of characters up to the nearest end of line or end of input. After a failure of `(Store / Print)`, the parser backtracks and consumes all of `BadLine`.

The name in brackets preceded by a tilde: `~{badLine}` identifies a semantic action to be invoked in case of failure. Thus, `badLine` is invoked when the first alternative of `Line` fails. The method has an access to a `Phrase` object for `Line`, appearing as the left-hand side of the failing rule `Line = (Store / Print)`. There is no right-hand side, but `lhs()` contains error information. The function of `badLine` is to print out this information and erase it. This is done by calling helper methods `errMsg()` and `errClear()`:

```
//-----
// Line = Store / Print (failed)
//-----
void badLine()
{
    System.out.println(lhs().errMsg());
    lhs().errClear();
}
```

(The line will not be rescanned, so it is safe to print the message.)

You find the modified grammar and semantics in `example10`, together with the source for `FileCalc`. Generate the new parser in the usual way and compile everything; then damage the input file in several places, and run the parser on it using `FileCalc` to see the diagnostics.

17 Left Recursion

Looking at specifications of programming languages you probably noticed that syntax of arithmetic expressions is often defined in a different way. A sum of integers, such as in Section 2.2, may be defined like this:

```
Sum    = Sum "+" Number / Number ;
Number = [0-9]+ ;
```

But this does not work as PEG. Procedure `Sum` would invoke itself repeatedly on the same input, in an infinite descent. We say that `Sum` is *left-recursive*.

The above, on the other hand, makes perfect sense when read as EBNF, and defines `Sum` as one or more `Number` s separated by `"+"`.

Beginning with Version 2.0, *Mouse* accepts such "pseudo-PEG" expressions and treats them as EBNF syntax definitions. The way of defining semantics is unchanged. The left-recursive version of example from Section 5 may appear like this:

```
Compute = Sum !_ {print} ;
Sum     = Sum "+" Number {sum}
        / Number {pass} ;
Number  = [0-9]+ {number} ;
```

It was necessary to add `Compute` because you do not want to print the result after each occurrence of `Sum`. The semantic action `number()` is the same as in Section 5. Action `pass()` just passes the value from `Number` to `Sum`:

```
//-----
// Sum = Number
//      0
//-----
void pass()
{ lhs().put(rhs(0).get()); }
```

Action `sum()` does the actual computation and `print()` prints the result.

```
//-----  
// Sum = Sum "+" Number  
//      0  1  2  
//-----  
void sum()  
{ lhs().put((Integer)rhs(0).get() + (Integer)rhs(2).get()); }
```

You find the grammar and semantics file in `example2R`; copy it to the work directory. Generating the parser may look like this:

```
java mouse.Generate -P myParser -G myGrammar.txt -S mySemantics  
Parsing procedures:  
2 rules  
0 inner  
3 terminals  
4 procedures for 1 left-recursion class(es)
```

You may then try the generated parser:

```
java mouse.TryParser -P myParser  
> 17+4711  
4728  
> 12+2+3  
17  
>
```

Mouse does this trick by generating behind the scenes a Parsing Expression Grammar that is equivalent to your PEG and invokes the semantic actions in the correct context. Looking at the generated parser, you can see that left-recursive expressions have special procedures whose names begin with "\$".

The generated grammar is not left-recursive under condition that your PEG does not contain "cycles". A cycle means that procedure can call itself without consuming any input. Here is an example:

```
A = B "a"? / "a" ;  
B = A "b"? / "b" ;
```

as "a"? and "b"? can succeed without consuming anything, A can call B and then call A with nothing being consumed.

The support for left recursion is restricted to Choice and Sequence expressions. In addition, the first expression in recursive Sequence must never accept an empty string. The semantic action "on failure" (described later) and Boolean action (see 20.4) are not supported for left-recursive rules.

In directory `example4R` you find the grammar of Example 4 from Section 8 rewritten in left-recursive way:

```
Compute = Space Sum !_ {print} ;  
Sum      = Sum AddOp Number {sum}  
          / Number {pass} ;  
Number   = Sign Digits? "." Digits Space {fraction}  
          / Sign Digits Space {integer} ;  
Sign     = ("-" Space)? ;  
AddOp    = ["+" Space] ;  
Digits   = "[0-9]+" ;  
Space    = " " * ;
```

As in the preceding case, it was necessary to add `Compute`. `Sign` had to be moved to `Number` so that it applies only to the first term of the sum. You find semantic actions in the same directory.

Directory `example8R` contains left-recursive version of Example 8 from Section 14.

Directory `example11` contains examples of left-recursive grammars found in the literature.

18 Extensions to PEG

Mouse accepts the following four forms of expression that are not part of the standard PEG:

$e_1 * e_2$	Shorthand for $(!e_2 e_1) * e_2$: iterate e_1 until e_2 .
$e_1 ++ e_2$	Shorthand for $(!e_2 e_1) + e_2$: iterate e_1 at least once until e_2 .
$e_1 : e_2$	Invoke e_1 . If it succeeds, invoke e_2 on the string consumed by e_1 . Indicate success if it consumes exactly that string. Otherwise reset input as it was before the invocation of e_1 (do not consume anything) and indicate failure.
$e_1 ! e_2$	Invoke e_1 . If it succeeds, invoke e_2 on the string consumed by e_1 . Indicate success if it does not consume exactly that string. Otherwise reset input as it was before the invocation of e_1 (do not consume anything) and indicate failure.

Using `*+` and `++`, the definitions

```
Comment = "/*" (!"/" _) * "*/" ;
Input   = (!EOF Line)+ EOF ;
String  = "[" (!"[" Char)+ "]" ;
```

are rewritten as

```
Comment = "/*" _*+ "*/" ;
Input   = Line++ EOF ;
String  = "[" Char++ "]" ;
```

which is not only easier to read, but also has a better implementation.

The operators `:` and `!` are intended to simplify definition of reserved words.

Suppose you want to define `Identifier` as any sequence of letters other than `"int"` or `"float"`. Using pure PEG, you can do it as follows, with `INT` and `FLOAT` used to denote these keywords in your syntax:

```
Identifier = !Keyword Letter+ ;
Keyword    = ("int"/"float") !Letter ;
INT        = "int" !Letter ;
FLOAT      = "float" !Letter ;
```

The predicate `!Letter` is needed because `("int"/"float")` fails on strings like `internal` or `floater`, which are thus not recognized as identifiers, while their prefixes are falsely recognized as keywords.

Using the `colon` operators you define the same like this:

```
Word = Letter+ ;
Identifier = Word!("int"/"float") ;
INT     = Word:"int" ;
FLOAT   = Word:"float" ;
```

It is more clear and, more important, facilitates the inspection by *PEG Explorer*.

All four operators have precedence 4 (cf. table in Section 2.1).

To obtain the `right-hand side` described in Section 6, replace `*+` and `++` by, respectively, `*` and `+` in the construction outlined there.

The second expression of $e_1 : e_2$ and $e_1 ! e_2$ does not appear in the `right-hand side`. The expression together with the preceding operator should be removed in the described construction.

Note: reserved words can also be defined with the help of Boolean actions, described under 20.4, but their definition is not visible in the syntax and is beyond the reach of *PEG Explorer*. This is, however, the only way if the keywords depend on the preceding context like in, for example, languages C and C++.

19 Checking the grammar

Before generating the parser, *Mouse* performs a number of checks on the grammar, and issues messages about its findings. Using the *Mouse* tool `TestPEG`, you can perform the same checking and obtain the same diagnostics without generating the parser. By specifying option `-D` to the tool, you can also display the grammar.

As an example, below is the result of `TestPEG` for the grammar:

```
A = (!"a")* !_ / A "b"? ;
B = B "a" ;
```

```
C:\Users\Giraf\Work>java mouse.TestPEG -G bad.peg -DL
Error: B is void.
Error: B "a" is void.
Error: the grammar has cycle involving A.
Info: recursion class of B is not used.
Error: argument of (!"a")* is nullable.

  2 rules
  6 unnamed
  4 terminals
  4 left-recursive expressions in 2 classes

Rules
  A = (!"a")* !_ / A "b"? ; // 1 rec class A
  B = B "a" ; // v rec class B

Inner
  (!"a")* !_ // 1ef
  (!"a")* // 0
  !"a" // 0f
  A "b"? // 1 rec class A
  "b"? // 01
  B "a" // v rec class B

Terminals
  "a" // 1f
  !_ // 1ef
  "b" // 1f
  "a" // 1f
```

An expression being *void* means that it does not generate any string, in the EBNF sense.

Thus, `B = B "a"` is void because it can only produce an infinite chain of replacements.

The notion of *cycle* was explained in Section 17.

The symbols appearing as comments on the right have these meanings:

- 0 = may consume null string;
- 1 = may consume non-null string;
- f = may fail;
- e = ends processing;
- v = void.

Comment `"rec class A"` means that expression belongs to the *recursion class* containing A. Recursion classes are sets of left-recursive expressions that interact only among themselves. `TestPEG` with option `-L` displays information about recursion classes:

```
java mouse.TestPEG -G bad.peg -L
Recursion class A
members:
  A (entry) (exit)
  A "b"?
seeds:
  (!"a")* !_ (in A)

Recursion class B
members:
  B
  B "a"
seeds:
```

An *exit* of recursion class is a Choice expression with at least one alternative outside the class. This alternative is a *seed* that terminates the recursion. An *entry* to the class is the expression that starts left recursion.

You may also want to find out how your grammar is affected by limited backtracking.

An interactive tool, the *PEG Explorer* assists you in checking the condition (*) stated in Section 9. It is described in <http://mousepeg.sourceforge.net/explorer.htm>, and is invoked as `mouse.ExplorePEG`.

20 Miscellaneous features

This section describes some features not mentioned in the tutorial.

20.1 Escaped characters

None of the characters you write in String Literal, Character Class, and Character Range is allowed to be newline or carriage return. The same applies to diagnostic names in pointed brackets. It is also not advisable to use there any white-space characters other than blank. Other characters, namely ",], -, >, cannot be used because they will be mistaken for delimiters. Still other characters may be not available at your keyboard. You specify all such characters using the conventions of "escape".

Any character other than `n`, `r`, `t`, or `u`, preceded by a backslash `\` represents itself as part of the string *s* in String Literal and Character Class, or as *c*₁, *c*₂ in Character Range, and is not recognized as a delimiter. Thus, for example, `"\"` is a String Literal of one double quote `"`, and `[[\]]` is a Character Class consisting of the brackets `[` and `]`.

The special combinations `\n`, `\r`, and `\t` represent the newline, carriage return, and tab characters, respectively. The special combination `\u` must be followed by four hexadecimal digits representing a Unicode character. Thus, for example, `\u000B` stands for the vertical tab character.

20.2 Character encoding

Both your grammar file and the file processed by your parser consist of bytes that may be interpreted in different ways depending on the used *character encoding scheme*. For example, the byte with value 248 (hex `f8`) means the Norwegian character `ø` in the encoding ISO-8859-1, or the Czech character `ř` in the encoding ISO-8859-2; the command prompt window of Windows XP displays it as a degree sign `°`. The bytes with values in the range 32 through 126, known as the ASCII character set, have a fixed interpretation that does not depend of encoding.

Both the `Generate` tool and the resulting parser use the default character encoding set up for the Java Virtual Machine executing them. Your grammar uses, in principle, only the ASCII characters, but you can have other characters in values of terminals, in diagnostic names, or in comments. These characters are interpreted according to the default encoding. When they appear as literals in the generated Java file, they are converted to Unicode escapes that are composed of ASCII characters. It means that the compiled parser does not depend on the encoding used at the compile time.

The input file processed by the parser is also interpreted according to the default encoding. If you need, you may enforce a specific encoding for input files by modifying the wrapper `SourceFile`.

20.3 Reserved names

The names you give to expressions and syntactic actions are used as method and object names in the generated parser. Therefore they must be distinct from Java reserved words. They may also conflict with names used in the base classes, but no checks are made in the present version, and **you have, unfortunately, to watch yourself for possible conflicts!**

20.4 Boolean actions

You may specify your semantic action to return a boolean value. You do this by placing an ampersand in front of the action name. The action returning `false` causes the associated expression to fail.

For example, the action `identifier()` invoked by `Identifier` below may return `false` if the recognized text is one of reserved words, thus forcing `Identifier` to fail on reserved words:

```
Identifier = Letter (Letter / Digit)* {& identifier} ;
```

The failure will result in a message saying "expected Identifier".

In the following example, the action `identifier()` returning `false` causes the first alternative to fail, so that `Factor` will proceed to try `Number`:

```
Factor = Letter (Letter / Digit)* {& identifier}  
       / Number {number} ;
```

No error message is generated in this case; if you wish, you may generate one using the helper method `errAdd()`.

The actions invoked on failure cannot return boolean value.

20.5 Deferred actions

You may save in a `Phrase` object an action to be executed later on. Such *deferred action* is a lambda expression of type `Deferred`. You save the action using the helper method `actAdd()`, and may have several actions saved in the same `Phrase`. When procedure represented by the `Phrase` successfully terminates, all its saved actions are added to `Phrase` of the calling procedure. When procedure represented by the `Phrase` fails, all its saved actions are discarded together with the `Phrase` object.

You execute the saved actions at the moment of your choice using the helper method `actExec()`. It executes all actions saved in the `Phrase` and removes them.

You can remove saved actions using the helper method `actClear()`.

Any saved actions remaining at the end are executed automatically on exit from the parser.

The use of deferred actions is illustrated on in Section 14 on example 8.

20.6 Omitted action names

To avoid clobbering your grammar with names of semantic actions, you may only indicate the presence of an action with an empty pair of brackets `{}`. The semantic action thus specified has the same name as the rule. The actions for multiple alternatives of a choice expression will have numbers attached to them. For example, you can write:

```
Product = Factor (MultOp Factor)* {} ;  
Factor  = Digits? "." Digits Space {}  
        / Digits Space           {}  
        / Lparen Sum Rparen      {}  
        / Name Space             {} ;
```

The semantic actions defined here are `Product()` and `Factor_0()` through `Factor_3()`. These names will appear in the skeleton generated by `-s` option.

You specify a Boolean action by an ampersand in brackets: `{&}`.

You specify an error action by brackets preceded with a tilde: `~{}`.

The name of the action thus specified has suffix `"_fail"`.

20.7 Package name

In a real application, your parser and semantics classes will belong to some named package. You specify the name of that package using option `-p` on `mouse.Generate`.

20.8 Specifying output directory

By default, the `Generate` tool places the resulting parser (and semantics skeleton if requested) in the current work directory. This default has been used in all the examples. You can specify any output directory of your choice using the `-D` option on `mouse.Generate`. The name of this directory need not correspond to the specified package name.

20.9 Own diagnostic output

The error messages described in Section 14 are written to `System.out`. If this does not suit you, you may yourself take care of the message as shown in Section 16. You define a semantic action to be called on failure of the top procedure. In that action, you obtain the error message using `lhs().errMsg()`, and handle it in the way you need. You should then issue `lhs().errClear()` to prevent the message from being written to `System.out`.

The messages returned by `errMsg()` may contain unprintable characters appearing in your String Literals, Character Classes, and Character Ranges. In the `String` returned by `errMsg()`, characters outside the range `\u0020` through `\u00ff` are replaced by Java-like escapes such as `\n` for a newline character, or `\u03A9` for the Greek letter Ω . Characters in the range `\u007f` through `\u00ff` are not replaced; they may be displayed differently depending on the encoding used to view the message.

(The above applies to all messages from the parser.)

20.10 Trace

The base class `mouse.runtime.SemanticsBase` contains a `String` variable `trace`, normally initialized to empty string. Using the method `setTrace()` of your parser, or the option `-T` on `Mouse` tools `TryParser` and `TestParser`, you can assign a value to this variable in order to trigger special actions such as tracing.

Note that the same value is used to trigger traces in the test version of the parser; therefore, you should not use strings containing letters `e`, `i`, or `r` if you are going to use this feature together with the test version.

20.11 Initialization

The base class `mouse.runtime.SemanticsBase` contains an empty method `void Init()` that is invoked every time you start your parser by invoking its `parse` method. By overriding this method in your semantics class, you can provide your own initialization of that class.

21 Deploying the parser

To include the generated parser in your application, you need to include your parser and semantics class, in the package of your choice (specified by the `-p` option on `mouse.Generate`).

You also need the standard *Mouse* package `mouse.runtime`. It is included in the JAR file `Mouse-2.3.jar`, and is also provided as a separate JAR file `Mouse-2.3.runtime.jar`. It is sufficient to have any of them accessible via `CLASSPATH`. You may also extract the directory `mouse/runtime` from any of these JAR files and make it accessible via `CLASSPATH`.

Alternatively, you can integrate the runtime support into your package. You can use for this purpose the *Mouse* tool `MakeRuntime`. For example, the command:

```
mouse.MakeRuntime -r my.runtime
```

produces in the current work directory the source code of all runtime support classes, with declaration `"package my.runtime;"`. By specifying the option `-r my.runtime` on the invocation of `Generate`, your generated parser will be an extension of `my.runtime.ParserBase` rather than `mouse.runtime.ParserBase`. The generated skeleton of your semantics class will be an extension of `my.runtime.SemanticsBase`. (If you create the class on your own, you have to declare it so.)

The runtime support consists of eleven classes:

- `ParserBase`.
- `ParserMemo`.
- `ParserTest`.
- `SemanticsBase`.
- `Phrase`.
- `CurrentRule` - interface for accessing parser stack.
- `FuncVB` - functional interface void to boolean.
- `FuncVV` - functional interface void to void.
- `Source` - interface of input wrappers.
- `SourceFile` - wrapper for input from a file.
- `SourceString` - wrapper for input from a `String`.

Of these, you need `ParserMemo` only if you generate the memoizing (`-M`) or the instrumented (`-T`) version of the parser. You need `ParserTest` only for the instrumented version. Of `SourceFile` and `SourceString` you need only the one that you use.

`SourceFile` assumes that the input file uses default character encoding. You can change it by modifying `SourceFile.java` in the place indicated by a comment.

A Appendix: The grammar of *Mouse* PEG

```
Grammar = Space (Rule/Skip)** EOT {Grammar} ;
Rule    = Name EQUAL RuleRhs DiagName? SEMI {Rule} ~{Error} ;
Skip    = SEMI
        / _++ (SEMI/EOT) ;
RuleRhs = Sequence Actions (SLASH Sequence Actions)* {} <right-hand side> ;
Choice  = Sequence (SLASH Sequence)* {} ;
Sequence = Prefixed+ {} ;
Prefixed = (AND / NOT) Suffixed {Prefix}
         / Suffixed {Pass} ;
Suffixed = Primary (STARPLUS / PLUSPLUS / IS / ISNOT) Primary {Infix}
         / Primary (QUERY / STAR / PLUS) {Suffix}
         / Primary {Pass} ;
Primary  = Name {Resolve}
         / LPAREN Choice RPAREN {Pass2}
         / ANY {Any}
         / StringLit {Pass}
         / Range {Pass}
         / CharClass {Pass} ;
Actions  = OnSucc OnFail {} ;
OnSucc   = (LWING AND? Name? RWING)? {} ;
OnFail   = (TILDA LWING Name? RWING)? {} ;
Name     = Letter (Letter / Digit)* Space {} ;
DiagName = "<" Char++ ">" Space {} ;
StringLit = "[" Char++ "]" Space {} ;
CharClass = ("[" / "^[" Char++ "]" Space {} ;
Range     = "[" Char "-" Char "]" Space {} ;
Char      = Escape {Pass}
         / ^[\r\n\\] {Char} ;
Escape    = "\\u" HexDigit HexDigit HexDigit HexDigit {Unicode}
         / "\\t" {Tab}
         / "\\n" {Newline}
         / "\\r" {CarRet}
         / !"\\u" "\\_ {Escape} ;
Letter    = [a-z] / [A-Z] ;
Digit     = [0-9] ;
HexDigit  = [0-9] / [a-f] / [A-F] ;
AND       = [&] Space <&> ;
NOT       = [!] Space <!> ;
QUERY    = [?] Space <?> ;
STAR     = [*]![+] Space <*> ;
PLUS     = [+]![+] Space <+> ;
STARPLUS = ["*+" Space <*+> ;
PLUSPLUS = ["++" Space <++> ;
IS       = [:]![!] Space <:> ;
ISNOT    = [!] Space <!> ;
EQUAL    = [=] Space <=> ;
LPAREN   = [(] Space <<> ;
RPAREN   = [)] Space <>> ;
LWING    = [{] Space <<> ;
RWING    = [}] Space <>> ;
SEMI     = [;] Space <;> ;
SLASH    = [/] Space </> ;
TILDA    = [~] Space <~> ;
ANY      = [_] Space <> ;
Space    = ([ \r\n\t] / Comment)* ;
Comment  = "//" _** EOL ;
EOL      = [\r]? [\n] / !_ <end of line> ;
EOT      = !_ <end of text> ;
```

B Appendix: Helper methods

These methods provide access to the environment seen by a semantic action.

The following four methods are inherited from `mouse.runtime.Semantics`. They call back the parser to access the relevant `Phrase` objects.

`Phrase lhs()`

Returns the left-hand side `Phrase`.

`int rhsSize()`

Returns the number of right-hand side `Phrases`.

`Phrase rhs(int i)`

Returns the i -th `Phrase` of the right-hand side, $0 \leq i < \text{rhsSize}()$.

`String rhsText(int i, int j)`

Returns as one `String` the text represented by the right-hand side `Phrases` i through $j - 1$, where $0 \leq i < j \leq \text{rhsSize}()$.

The following fifteen methods can be applied to a `Phrase` object:

`void put(Object v)`

Inserts v as the semantic value of this `Phrase`.

`Object get()`

Returns the semantic value of this `Phrase`.

`String text()`

Returns the text represented by this `Phrase`.

`char charAt(int i)`

Returns the i -th character of the text represented by this `Phrase`.

`String rule()`

Returns name of the rule that created this `Phrase`.

`boolean isA(String name)`

Returns true if this `Phrase` was created by rule $name$.

`boolean isTerm()`

Returns true if this `Phrase` was created by a terminal.

`boolean isEmpty()`

Returns true if the text represented by this `Phrase` is empty.

`String where(int i)`

Returns a printable string describing where in the text being parsed find the i -th character of text represented by this `Phrase`.

`String errMsg()`

Returns the error message contained in this `Phrase`.

`void errClear()`

Erases the error message contained in this `Phrase`.

`void errAdd(String expr, int i)`

Updates error message contained in this `Phrase` with information that expression $expr$ failed at the i -th character of text represented by this `Phrase`.

`void actAdd(FuncVV act)`

Adds to this `Phrase` deferred action act : a lambda expression implementing the interface `FuncVV`.

`void actClear()`

Removes all deferred actions saved in this `Phrase`.

`void actExec()`

Executes and removes all deferred actions saved in this `Phrase`.

C Appendix: Your parser class

These are the methods you can apply to your generated parser.

"Parser" and "Semantics" are names of your parser and semantics classes, respectively.

Parser()

Parser constructor. Instantiates your parser and semantics, connects semantics object to the parser, and returns the resulting parser object.

boolean **parse**(Source src)

Parses input wrapped into a Source object *src*.

Returns true if the parse was successful, or false otherwise.

Semantics **semantics**()

Returns the semantics object associated with the parser.

void **setTrace**(String s)

Assigns *s* to the trace field in semantics object.

void **setMemo**(int n)

Sets the amount of memoization to n , $0 \leq n \leq 9$.

Can only be applied to a parser generated with option -M or -T (see mouse.Generate tool).

D Appendix: *Mouse* tools

mouse.Generate *arguments*

Generates parser.

The *arguments* are specified as options according to POSIX syntax:

-G *filename*

Identifies the file containing your grammar. Mandatory.

The *filename* need not be a complete path, just enough to identify the file in current environment. It should include file extension, if any.

-D *directory*

Identifies target directory to receive the generated file(s).

Optional. If omitted, files are generated in current work directory.

The *directory* need not be a complete path, just enough to identify the directory in current environment. The directory must exist.

-P *parser*

Specifies name of the parser to be generated. Mandatory.

Must be an unqualified class name. The tool generates a file named "*parser*.java" in the target directory. The file contains definition of Java class *parser*. If target directory already contains a file "*parser*.java", the file is replaced without a warning,

-S *semantics*

Indicates that semantic actions are methods in the Java class *semantics*.

Mandatory if your grammar specifies semantic actions. Must be an unqualified class name.

-p *package*

Generate parser as member of package *package*.

The semantics class, if specified, is assumed to belong to the same package.

Optional. If not specified, both classes belong to unnamed package.

The specified package need not correspond to the target directory.

-r *runtime-package*

Generate parser that will use runtime support package *runtime-package*.

Optional. If not specified, `mouse.runtime` is assumed.

-s Generate skeleton of semantics class. Optional.

The skeleton is generated as file "*semantics*.java" in the target directory, where *semantics* is the name specified by -S option. The option is ignored if -S is not specified. If target directory already contains a file "*semantics*.java", the tool is not executed, to prevent accidental destruction of your semantics class.

-M Generate memoizing version of the parser.

-T Generate instrumented test version of the parser.

(Options -M and -T are mutually exclusive.)

mouse.TestPEG arguments

Checks your grammar without generating parser.

The *arguments* are specified as options according to POSIX syntax:

-G *filename*

Identifies the file that contains your grammar. Mandatory.

The *filename* should include any extension.

Need not be a complete path, just enough to identify the file in your environment.

-D Display the grammar. Optional.

Shows the rules and subexpressions together with their attributes.

(0=may consume empty string, 1=may consume nonempty string, f= may fail, e= ends processing, v=void.)

-C Display the grammar in compact form: without duplicate subexpressions. Optional.

-R Display only the rules. Optional.

-L Display data about left recursion. Optional.

-d Display the dual grammar created for left-recursion processing. Optional.

-r Display only the rules of dual grammar. Optional.

mouse.TryParser arguments

Runs the generated parser.

The *arguments* are specified as options according to POSIX syntax:

-P *parser*

Identifies the parser. Mandatory.

parser is the class name, fully qualified with package name, if applicable. The class must reside in a directory corresponding to the package.

-f *filename*

Apply the parser to file *filename*. Optional.

The *filename* should include any extension.

Need not be a complete path, just enough to identify the file in your environment.

-F *list*

Apply the parser separately to each file in a list of files. Optional.

The *list* identifies a text file containing one fully qualified file name per line. The *list* itself need not be a complete path, just enough to identify the file in your environment.

-m *n*

Amount of memoization. Optional. Applicable only to a parser generated with option **-M**.

n is a digit from 1 through 9 specifying the number of results to be cached. Default is no memoization.

-T *string*

Tracing switches. Optional.

The *string* is assigned to the `trace` field in your semantics object, where you can use it to activate any trace you have programmed there.

-t Display timing information.

Note that the precision of time measurements corresponds to the granularity of the system's clock. Also, that Java's execution times may vary.

Options **-f** and **-F** are mutually exclusive.

If you do not specify **-f** or **-F**, the parser is executed interactively, prompting for input by printing `>`. It is invoked separately for each input line after you press `Enter`. You terminate the session by pressing `Enter` directly at the prompt.

mouse.TestParser *arguments*

Runs the instrumented parser (generated with option `-T`), and prints information about its operation, as explained in Sections 9 and 10. The *arguments* are specified as options according to POSIX syntax:

`-P parser`

Identifies the parser. Mandatory.

parser is the class name, fully qualified with package name, if applicable. The class must reside in a directory corresponding to the package.

`-f filename`

Apply the parser to file *filename*. Optional.

The *filename* should include any extension.

Need not be a complete path, just enough to identify the file in your environment.

`-F list`

Apply the parser separately to each file in a list of files. Optional.

The *list* identifies a text file containing one fully qualified file name per line. The *list* itself need not be a complete path, just enough to identify the file in your environment.

`-m n`

Amount of memoization. Optional.

n is a digit from 1 through 9 specifying the number of results to be retained. Default is no memoization.

`-T string`

Tracing switches. Optional,

The *string* is assigned to the `ttrace` field in your semantics object, where you can use it to trigger any trace you have programmed there. In addition, presence of certain letters in *string* activates traces in the parser:

`r` – trace execution of parsing procedures for rules.

`i` – trace execution of parsing procedures for inner expressions.

`e` – trace error information.

`-d` Show detailed statistics for procedures involved in backtracking - rescan - reuse. Optional.

`-D` Show detailed statistics for all invoked procedures. Optional.

`-C filename`

Write all statistics as comma-separated values (CSV) to file *filename*, rather than to `System.out`.

Optional; can only be specified with `-F`.

The *filename* should include any extension.

Need not be a complete path, just enough to identify the file in your environment.

If the file *filename* exists, it is overwritten. Otherwise the file is created.

`-t` Display timing information.

Options `-f` and `-F` are mutually exclusive, as well as `-d` and `-D`.

If you do not specify `-f` or `-F`, the parser is executed interactively, prompting for input by printing `>`. It is invoked separately for each input line after you press `Enter`. You terminate the session by pressing `Enter` directly at the prompt.

mouse.MakeRuntime *arguments*

Re-package *Mouse* runtime support: write out Java source files for runtime-support classes using specified package name instead of `mouse.runtime`.

The *arguments* are specified as options according to POSIX syntax:

`-r runtime-package`

Generate files with package name *runtime-package*. Mandatory.

`-D directory`

Identifies target directory to receive the files.

Optional. If omitted, files are written to the current work directory.

The *directory* need not be a complete path, just enough to identify the directory in current environment. The directory must exist.

The directory name need not correspond to the package name.

mouse.ExplorePEG `-G filename`

Invokes *PEG Explorer* to analyze the grammar contained in file *filename*. The *filename* need not be a complete path, just enough to identify the file in current environment. It should include file extension, if any.

E References

- [1] AHO, A. V., AND ULLMAN, J. D. *The Theory of Parsing, Translation and Compiling, Vol. I, Parsing*. Prentice Hall, 1972.
- [2] BIRMAN, A. *The TMG Recognition Schema*. PhD thesis, Princeton University, February 1970.
- [3] BIRMAN, A., AND ULLMAN, J. D. Parsing algorithms with backtrack. *Information and Control* 23 (1973), 1–34.
- [4] BROOKER, P., AND MORRIS, D. A general translation program for phrase structure languages. *J. ACM* 9, 1 (1962), 1–10.
- [5] FORD, B. The packrat parsing and Parsing Expression Grammars page. <http://bford.info/packrat>. Accessed 2015-07-15.
- [6] FORD, B. Packrat parsing: Simple, powerful, lazy, linear time. In *Proceedings of the Seventh ACM SIGPLAN International Conference on Functional Programming* (Pittsburgh, PA, USA, 2002), ACM, pp. 36–47. <http://www.bford.info/pub/lang/packrat-icfp02.pdf>.
- [7] FORD, B. Parsing expression grammars: A recognition-based syntactic foundation. In *Proceedings of the 31st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (Venice, Italy, 14–16 January 2004), N. D. Jones and X. Leroy, Eds., ACM, pp. 111–122. <http://pdos.csail.mit.edu/papers/parsing:popl04.pdf>.
- [8] GRIMM, R. Practical packrat parsing. Tech. Rep. TR2004-854, Dept. of Computer Science, New York University, March 2004.
- [9] HILL, O. Support for Left-Recursive PEGs, 2010. <https://github.com/orlandohill/peg-left-recursion>.
- [10] HOPGOOD, F. R. A. *Compiling Techniques*. MacDonalds, 1969.
- [11] LUCAS, P. The structure of formula-translators. *ALGOL Bulletin Supplement 16* (September 1961), 1–27.
- [12] MASCARENHAS, F., MEDEIROS, S., AND IERUSALIMSCHY, R. On the relation between context-free grammars and Parsing Expression Grammars. *Science of Computer Programming* 89 (2014), 235–250.
- [13] MCCLURE, R. M. TMG – a syntax directed compiler. In *Proceedings of the 20th ACM National Conference* (24–26 August 1965), L. Winner, Ed., ACM, pp. 262–274.
- [14] MEDEIROS, S., MASCARENHAS, F., AND IERUSALIMSCHY, R. Left recursion in Parsing Expression Grammars. *Science of Computer Programming* 96 (2014), 177–190.
- [15] REDZIEJOWSKI, R. R. Parsing Expression Grammar as a primitive recursive-descent parser with backtracking. *Fundamenta Informaticae* 79, 3–4 (2007), 513–524. <http://www.romanredz.se/papers/FI2007.pdf>.
- [16] REDZIEJOWSKI, R. R. Some aspects of Parsing Expression Grammar. *Fundamenta Informaticae* 85, 1–4 (2008), 441–454.
- [17] REDZIEJOWSKI, R. R. More about converting BNF to PEG. *Fundamenta Informaticae* 133, 2-3 (2014), 257–270. <http://www.romanredz.se/papers/FI2014.pdf>.
- [18] TRATT, L. Direct left-recursive parsing expression grammars. Tech. Rep. EIS-10-01, School of Engineering and Information Sciences, Middlesex University, Oct. 2010.
- [19] WARTH, A., DOUGLASS, J. R., AND MILLSTEIN, T. D. Packrat parsers can support left recursion. In *Proceedings of the 2008 ACM SIGPLAN Symposium on Partial Evaluation and Semantics-based Program Manipulation, PEPM 2008, San Francisco, California, USA, January 7-8, 2008* (2008), pp. 103–110.
- [20] WIKIPEDIA. Comparison of parser generators. https://en.wikipedia.org/wiki/Comparison_of_parser_generators. Accessed 2015-07-15.
- [21] WIKIPEDIA. Parsing expression grammar. http://en.wikipedia.org/wiki/Parsing_expression_grammar. Accessed 2015-07-15.