Apuntes de la Asignatura Procesadores de Lenguajes

Casiano R. León [1]

5 de abril de 2013

[1]DEIOC Universidad de La Laguna

Índice general

Índice de figuras

Índice de cuadros

A Juana

For it is in teaching that we learn
And it is in understanding that we are understood

Agradecimientos/Acknowledgments

I'd like to thank

A mis alumnos de Procesadores de Lenguajes del Grado de Informática de la Escuela Superior de Informática en la Universidad de La Laguna

Parte I

PRIMERA PARTE: APUNTES DE JAVASCRIPT

Capítulo 1

Introducción

Capítulo 2

Estructura Léxica

Capítulo 3

Tipos, Valores y Variables

Capítulo 4

Expresiones y Operadores

Capítulo 5

Sentencias

Capítulo 6

Objetos

6.1. Tutoriales de OOP en JavaScript en la Web

1. The Basics of Object-Oriented JavaScript Leigh Kaszick on Nov 11th 2009 en NetTuts+
2. Understanding JavaScript OOP por Quildreen Motta.
3. Object.defineProperty(obj, prop, descriptor)
4. What is the 'new' keyword in JavaScript?
5. Constructors considered mildly confusing
6. Google I/O 2011: Learning to Love JavaScript por Alex Russell, Mayo 2011. ¡Excelente!

6.2. Ejercicios

1. ¿Cual es la salida?

```
> z
{ x: 3, y: 1 }
> Object.keys(z)
?????
> Object.keys(z).forEach(function(i) { console.log(i+" -> "+z[i]); })
?????
```

2. ¿Que queda finalmente en `z`?

```
> z
{ x: 2, y: 1 }
> Object.defineProperty(z, 'y', {writable : false})
{ x: 2, y: 1 }
> z.x = 3
??????
> z.y = 2
??????
> z
```

3. ¿Cuales son las salidas?

```
> obj = { x : 1, y : 2}
{ x: 1, y: 2 }
> bValue = 5
5
> Object.defineProperty(obj, "b", {get: function(){ return bValue; },
                                    set: function(y){ bValue = y; }})
{ x: 1, y: 2 }
> obj.b = "hello"

> bValue

> obj.b

> bValue = "world"

> obj.b

>
```

4. ¿Cuales son las salidas?

```
> var o = {};
undefined
> Object.defineProperty(o, "a", { value : 1, enumerable:true });
{ a: 1 }
> Object.defineProperty(o, "b", { value : 2, enumerable:false });
{ a: 1 }
> Object.defineProperty(o, "c", { value : 3 }); // enumerable defaults to false
{ a: 1 }
> o.d = 4; // enumerable defaults to true when creating a property by setting it
4
>
undefined
> for (var i in o) {
...   console.log(i);
... }
???????
> Object.keys(o);
???????
> o.propertyIsEnumerable('a');
????
> o.propertyIsEnumerable('b');
????
> o.propertyIsEnumerable('c');
????
> o.b
????
> o["b"]
????
```

5.
```
> function foo(a, b){return a * b;}
undefined
> f = function foo(a, b){return a * b;}
```

```
[Function: foo]
> foo.length
2
> foo.constructor
[Function: Function]
> foo.prototype
{}
> typeof foo.prototype
'object'
> [1, 2].constructor
[Function: Array]
```

6. ¿Cuales son las salidas?

```
> Object.getPrototypeOf({ a: 1})

> Object.getPrototypeOf([1,2,3])

> Object.getPrototypeOf([1,2,3]) == Array.prototype

> Object.getPrototypeOf({ a:1 }) === Object.prototype

> Object.getPrototypeOf(Array.prototype)

> Object.getPrototypeOf(Object.prototype)

> Object.getPrototypeOf(function() {})

> Object.getPrototypeOf(Object.getPrototypeOf(function() {}))

> [1,2,3].__proto__

> [1,2,3].__proto__ == Array.prototype
```

```
var b = new Foo(20);
var c = new Foo(30);
```

6.3. Comprobando Propiedades

```
> o = { x: 1}
{ x: 1 }
> "x" in o
true
> "y" in o
false
> "toString" in o
true
> o.hasOwnProperty('x')
true
> o.hasOwnProperty('toString')
false
```

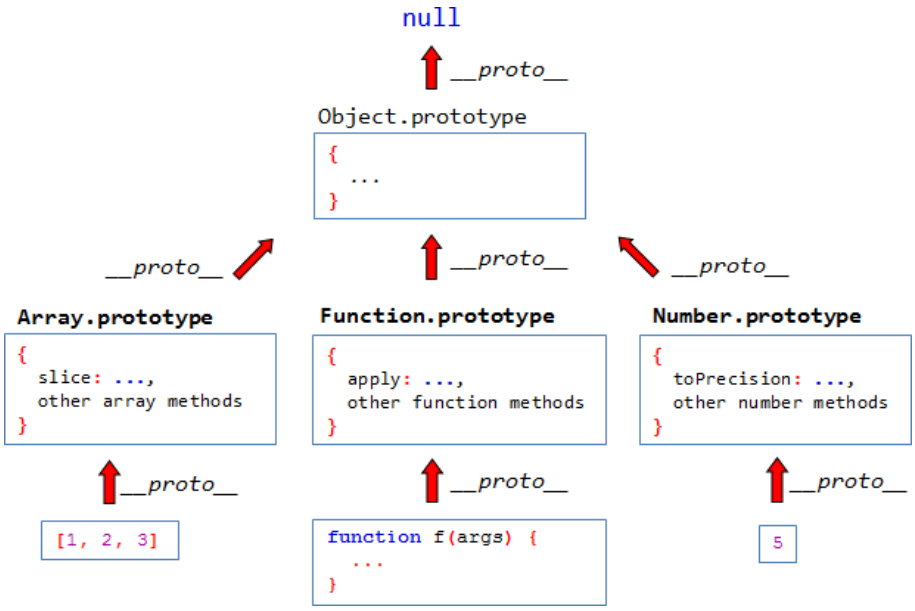


Figura 6.1: Jerarquía de Prototipos Nativos

6.4. Enumeración de Propiedades

```
> o = { x: 1 }
{ x: 1 }
> b = Object.create(o)
{}
> b.y = 2
2
> b.propertyIsEnumerable('x')
false
> b.propertyIsEnumerable('y')
true
> Object.prototype.propertyIsEnumerable('toString')
false
> a = {x:1, y:2 }
{ x: 1, y: 2 }
> b = Object.create(a)
{}
> b.z = 3
3
> for(i in b) console.log(b[i])
3
1
2
undefined
> for(i in b) console.log(i)
z
x
y
undefined
> b.propertyIsEnumerable("toString")
false
```

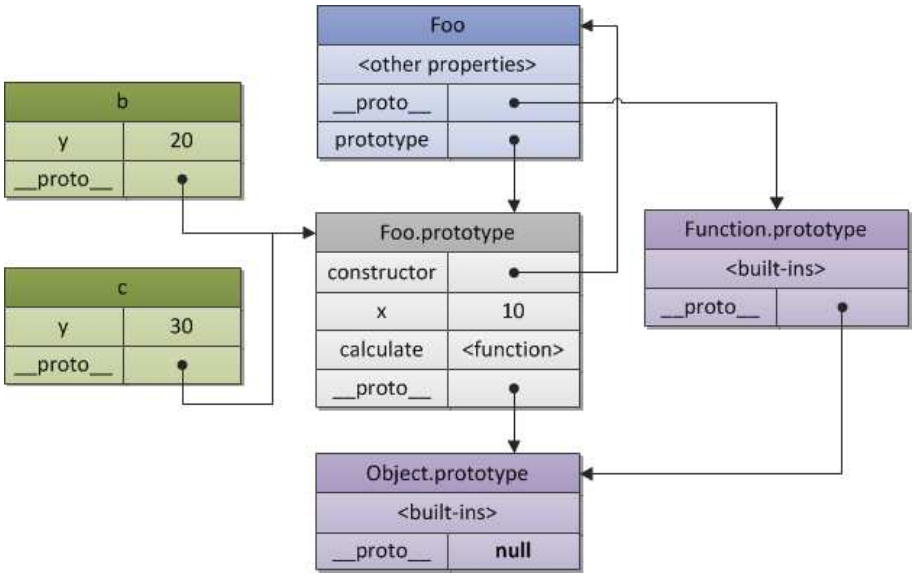


Figura 6.2: __proto__ and prototypes

Capítulo 7

Arrays

Capítulo 8

Funciones

8.1. Definiendo Funciones

8.2. Invocando Funciones

8.3. Argumentos y Parámetros

8.4. Funciones como Valores

8.5. Funciones como Espacios de Nombres

8.6. Clausuras

8.7. Propiedades, Métodos y Constructor

8.7.1. La propiedad `length`

8.7.2. La Propiedad `property`

8.7.3. Los Métodos `call` y `apply`

Los métodos `call` y `apply` nos permiten invocar una función como si fuera un método de algún otro objeto.

```
[~/Dropbox/src/javascript/learning]$ cat call.js
var Bob = {
  name: "Bob",
  greet: function() {
    console.log("Hi, I'm " + this.name);
  }
}

var Alice = {
  name: "Alice",
};

Bob.greet.call(Alice);

[~/Dropbox/src/javascript/learning]$ node call.js
Hi, I'm Alice
```

1. Function.apply and Function.call in JavaScript

```
> function f() { console.log(this.x); }
undefined
> f.toString()
'function f() { console.log(this.x); }'
> z = { x : 99 }
{ x: 99 }
> f.call(z)
99
undefined
>
```

2. ```
   > o  = { x : 15 }
   { x: 15 }
   > function f(m) { console.log(m+" "+this.x); }
   undefined
   > f("invoking f")
   invoking f 10
   undefined
   > f.call(o, "invoking f via call");
   invoking f via call 15
   undefined
   ```

8.8. Programación Funcional
   ```

Capítulo 9

Clases y Módulos

Véase el libro *Learning JavaScript Design Pattern* [1]

9.1. Herencia

```
[~/Dropbox/src/javascript/inheritance]$ cat inh1.js
//Shape - superclass
function Shape() {
  this.x = 0;
  this.y = 0;
}

Shape.prototype.toString = function() {
  return "("+this.x+", "+this.y+")";
}

Shape.prototype.move = function(x, y) {
    this.x += x;
    this.y += y;
    console.info("Shape moved to "+this);
};

// Rectangle - subclass
function Rectangle() {
  Shape.call(this); //call super constructor.
}

// Rectangle inherits from Shape
Rectangle.prototype = Object.create(Shape.prototype);

var rect = new Rectangle();

console.log("x = "+rect);
console.log("rect is an instance of Rectangle? "+(rect instanceof Rectangle)) //true.
console.log("rect is an instance of Shape? "+(rect instanceof Shape))         //true.

rect.move(1, 2); //Outputs, "Shape moved."

[~/Dropbox/src/javascript/inheritance]$ node inh1.js
x = (0, 0)
rect is an instance of Rectangle? true
```

```
rect is an instance of Shape? true
Shape moved to (1, 2)
```

9.2. Ejercicios

1. ¿Cual es la salida?

```
> String.prototype.repeat = function(times) {
...   return new Array(times+1).join(this)
... }
[Function]
>
undefined
> "hello".repeat(3)

>
```

2. ¿Cuales son los resultados?

```
> z = Object.create({x:1, y:2})
{}
> z.x

> z.y

> z.__proto__

> z.__proto__.__proto__

> z.__proto__.__proto__.__proto__
```

3. Describa las salidas:

```
> obj = {x : 'something' }
{ x: 'something' }
> w = Object.create(obj)
{}
> w.x
'something'
> w.x = "another thing"
'another thing'
> w.__proto__

> obj == w.__proto__
```

4. Explique la salida:

```
> obj = {x : { y : 1} }
{ x: { y: 1 } }
> w = Object.create(obj)
{}
```

```
> w.x == obj.x
true
> w.x.y = 2
2
> obj
{ x: { y: 2 } }
>
```

5. Explique las salidas:

```
> inherit = Object.create
[Function: create]
> o = {}
{}
> o.x = 1
1
> p = inherit(o)
{}
> p.x
1
> p.y = 2
2
> p
{ y: 2 }
> o
{ x: 1 }
> o.y
undefined
> q = inherit(p)
{}
> q.z = 3
3
> q
{ z: 3 }
> s = q.toString()
'[object Object]'
> q.x+q.y+q.z
6
> o.x
1
> o.x = 4
4
> p.x
4
> q.x
4
> q.x = 5
5
> p.x
4
> o.x
4
```

Capítulo 10

Subconjuntos y Extensiones de JavaScript

Capítulo 11

JavaScript en el Lado del Servidor

Node es un intérprete JavScript escrito en C++ y con una API ligada a Unix para trabajar con procesos, fichero, sockets, etc. Los programas Node por defecto nunca se bloquean. Node utiliza manejadores de eventos que a menudo se implementan haciendo uso de funciones anidadas y clausuras.

11.1. Instalar Node.js

1. http://nodejs.org/download/
2. Should I install node.js on Ubuntu using package manager or from source?
3. The Node Beginner Book

11.2. Primeros Pasos. Un Ejemplo Simple

```
[~/src/javascript/node.js/hector_correa_introduction_to_node(master)]$ cat -n hello_world.js
     1  console.log("Hello world!");
     2  a = [ 'batman', 'robin'];
     3  a.push("superman");
     4  console.log(a);
     5  h = { name: 'jane rodriguez-leon', department: 'IT' };
     6  console.log(h);
     7  console.log(h['name']);

[~/src/javascript/node.js/hector_correa_introduction_to_node(master)]$ node hello_world.js
Hello world!
[ 'batman', 'robin', 'superman' ]
{ name: 'jane rodriguez-leon', department: 'IT' }
jane rodriguez-leon

[~/Dropbox/academica/ETSII/grado/LPP/LPPbook]$ node
> .help
.break  Sometimes you get stuck, this gets you out
.clear  Alias for .break
.exit Exit the repl
.help Show repl options
.load Load JS from a file into the REPL session
.save Save all evaluated commands in this REPL session to a file

> console.log("Hello world!")
Hello world!
undefined
```

```
> a = [ 'batman', 'robin']
[ 'batman', 'robin' ]
> a.push("superman")
3
> a
[ 'batman', 'robin', 'superman' ]
> h = { name: 'jane rodriguez-leon', department: 'IT' }
{ name: 'jane rodriguez-leon',
  department: 'IT' }
> h['name']
'jane rodriguez-leon'
> 4+2
6
> _    # ultimo valor evaluado
6
> _+1
7
>
> a = [1,2,3]
[ 1, 2, 3 ]
> a.forEach(function(e) { console.log(e); })
1
2
3
> a.forEach(function(v) {
... console.log(v
..... .break
> a
[ 1, 2, 3 ]
> .exit # también CTRL-D en Unix
[~/Dropbox/academica/ETSII/grado/LPP/LPPbook]$
```

11.3. Usando REPL desde un programa

Es posible crear un bucle REPL en cualquier punto de nuestro programa - quizá para depurarlo. Para ello usamos la función `repl.start` . Esta función retorna una instancia REPLServer. Acepta como argumento un objeto `options` que toma los siguientes valores:

1. `prompt` - the prompt and stream for all I/O. Defaults to ¿.
2. `input` - the readable stream to listen to. Defaults to process.stdin.
3. `output` - the writable stream to write readline data to. Defaults to process.stdout.
4. `terminal` - pass true if the stream should be treated like a TTY, and have ANSI/VT100 escape codes written to it. Defaults to checking isTTY on the output stream upon instantiation.
5. `eval` - function that will be used to eval each given line. Defaults to an async wrapper for eval().
6. `useColors` - a boolean which specifies whether or not the writer function should output colors. If a different writer function is set then this does nothing. Defaults to the repl's terminal value.
7. `useGlobal` - if set to true, then the repl will use the global object, instead of running scripts in a separate context. Defaults to false.

8. `ignoreUndefined` - if set to true, then the repl will not output the return value of command if it's undefined. Defaults to false.

9. `writer` - the function to invoke for each command that gets evaluated which returns the formatting (including coloring) to display. Defaults to util.inspect.

```
[~/Dropbox/src/javascript/node.js/repl(master)]$ cat repl.js
var repl = require("repl");

connections = 0;

repl.start({
  prompt: "node via stdin> ",
  input: process.stdin,
  output: process.stdout
});

[~/Dropbox/src/javascript/node.js/repl(master)]$ node repl.js
node via stdin> 2+3
5
node via stdin> .exit
```

el bucle REPL proporciona acceso a las variables de ámbito global. Es posible hacer explícitamente visible una variable al REPL asignándosela al `context` asociado con el REPLServer. Por ejemplo:

```
[~/Dropbox/src/javascript/node.js/repl(master)]$ cat repl2.js
var repl = require("repl");

z = 4
repl.start({
  prompt: "node via stdin> ",
  input: process.stdin,
  output: process.stdout
}).context.m = "message";
```

Las variables en el objeto `context` se ven como locales al REPL:

```
[~/Dropbox/src/javascript/node.js/repl(master)]$ node repl2.js
node via stdin> z
4
node via stdin> m
'message'
```

11.4. Usando REPL via un socket TCP

```
[~/Dropbox/src/javascript/node.js/repl(master)]$ cat repl_server.js
var net = require("net"),
    repl = require("repl");

connections = 0;

net.createServer(function (socket) {
  connections += 1;
```

```
  repl.start({
    prompt: "node via TCP socket> ",
    input: socket,
    output: socket
  }).on('exit', function() {
    socket.end();
  });
}).listen(5001);

[~/Dropbox/src/javascript/node.js/repl(master)]$ node repl_server.js
```

Podemos ahora usar *netcat* para comunicar con el servidor:

```
[~/Dropbox/src/javascript/node.js/hector_correa_introduction_to_node(master)]$ nc -v localhost
nc: connect to localhost port 5001 (tcp) failed: Connection refused
Connection to localhost 5001 port [tcp/commplex-link] succeeded!
node via TCP socket> a = 2+3
5
node via TCP socket> a
5
node via TCP socket> .exit
[~/Dropbox/src/javascript/node.js/hector_correa_introduction_to_node(master)]$
```

11.5. Referencias sobre REPL

1. Véase Node.js v0.8.18 Manual & Documentation
2. Véase How do I use node's REPL? en http://docs.nodejitsu.com/.

11.6. Entrada Salida en Node.js

1. How To Read User Input With NodeJS por Nikolay V. Nemshilov

11.7. Debugger

1. Node.js debugger

11.8. Modulos

11.8.1. Introducción

```
[~/javascript/node.js/creating_modules(master)]$ cat foo.js
var circle = require('./circle.js');
console.log( 'The area of a circle of radius 4 is '
           + circle.area(4));

[~/javascript/node.js/creating_modules(master)]$ cat circle.js
var PI = Math.PI;

exports.area = function (r) {
  return PI * r * r;
};
```

```
exports.circumference = function (r) {
  return 2 * PI * r;
};

[~/javascript/node.js/creating_modules(master)]$ node foo.js The area of a circle of radius 4
```

El módulo `circle.js` exporta las funciones `area()` y `circumference()`. Para exportar un objeto lo añadimos al objeto expecial `exports`.

Las variables locales al módulo serán privadas. En este ejemplo la variable `PI` es privada a `circle.js`.

```
[~/javascript/node.js/creating_modules(master)]$ node debug foo.js
< debugger listening on port 5858
connecting... ok
break in foo.js:1
  1 var circle = require('./circle.js');
  2 console.log( 'The area of a circle of radius 4 is '
  3            + circle.area(4));
debug> n
break in foo.js:2
  1 var circle = require('./circle.js');
  2 console.log( 'The area of a circle of radius 4 is '
  3            + circle.area(4));
  4
debug> repl
Press Ctrl + C to leave debug repl
> circle
{ circumference: [Function],
  area: [Function] }
> circle.area(2)
12.566370614359172
> PI
ReferenceError: PI is not defined
>
```

11.8.2. Ciclos

```
[~/javascript/node.js/creating_modules/cycles(master)]$ cat a.js
console.log('a starting');
exports.done = false;
var b = require('./b.js');
console.log('in a, b.done = %j', b.done);
exports.done = true;
console.log('a done');

[~/javascript/node.js/creating_modules/cycles(master)]$ cat b.js
console.log('b starting');
exports.done = false;
var a = require('./a.js');
console.log('in b, a.done = %j', a.done);
exports.done = true;
console.log('b done');

[~/javascript/node.js/creating_modules/cycles(master)]$ cat main.js
console.log('main starting');
```

```
var a = require('./a.js');
var b = require('./b.js');
console.log('in main, a.done=%j, b.done=%j', a.done, b.done);
```

When `main.js` loads `a.js`, then `a.js` in turn loads `b.js`. At that point, `b.js` tries to load `a.js`. In order to prevent an infinite loop an unfinished copy of the `a.js` exports object is returned to the `b.js` module. `b.js` then finishes loading, and its exports object is provided to the `a.js` module.

By the time `main.js` has loaded both modules, they're both finished. The output of this program would thus be:

```
[~/javascript/node.js/creating_modules/cycles(master)]$ node main.js
main starting
a starting
b starting
in b, a.done = false
b done
in a, b.done = true
a done
in main, a.done=true, b.done=true
```

11.8.3. Especificación de Ficheros Conteniendo Módulos

1. If the exact filename is not found, then node will attempt to load the required filename with the added extension of `.js`, `.json`, and then `.node`.

2. `.js` files are interpreted as JavaScript text files, and `.json` files are parsed as JSON text files. `.node` files are interpreted as compiled addon modules

3. A module prefixed with `'/'` is an absolute path to the file. For example, `require('/home/marco/foo.js')` will load the file at `/home/marco/foo.js`.

4. A module prefixed with `'./'` is relative to the file calling require().

5. Without a leading verb—'/'— or `'./'` to indicate a file, the module is either a *core module* or is loaded from a `node_modules` folder.

6. If the given path does not exist, `require()` will throw an Error with its code property set to `MODULE_NOT_FOUND`.

11.8.4. Carga desde Carpetas `node_modules`

1. If the module identifier passed to `require()` is not a native module, and does not begin with `'/'`, `'../'`, or `'./'`, then node starts at the parent directory of the current module, and adds `/node_mo1dules`, and attempts to load the module from that location.

2. If it is not found there, then it moves to the parent directory, and so on, until the root of the tree is reached.

For example, if the file at `'/home/ry/projects/foo.js'` called require(`'bar.js'`), then node would look in the following locations, in this order:

```
/home/ry/projects/node_modules/bar.js
/home/ry/node_modules/bar.js
/home/node_modules/bar.js
/node_modules/bar.js
```

This allows programs to localize their dependencies, so that they do not clash.

11.8.5. Las Carpetas Usadas Como Módulos

It is convenient to organize programs and libraries into self-contained directories, and then provide a single entry point to that library.

There are a few ways in which a folder may be passed to `require()` as an argument.

1. The first is to create a `package.json` file in the root of the folder, which specifies a `main` module. An example `package.json` file might look like this:

   ```
   { "name" : "some-library",
     "main" : "./lib/some-library.js" }
   ```

   If this was in a folder at `./some-library`, then `require('./some-library')` would attempt to load `./some-library/lib/some-library.js`.

   This is the extent of Node's awareness of `package.json` files.

2. If there is no `package.json` file present in the directory, then node will attempt to load an `index.js` or `index.node` file out of that directory.

   For example, if there was no `package.json` file in the above example, then `require('./some-library')` would attempt to load:

   ```
   ./some-library/index.js
   ./some-library/index.node
   ```

11.8.6. Caching

1. Modules are cached after the first time they are loaded. This means (among other things) that every call to `require('foo')` will get exactly the same object returned, if it would resolve to the same file.

2. Multiple calls to `require('foo')` may not cause the module code to be executed multiple times. This is an important feature. With it, *partially done* objects can be returned, thus allowing transitive dependencies to be loaded even when they would cause cycles.

3. If you want to have a module execute code multiple times, then export a function, and call that function.

4. Modules are cached based on their **resolved filename**. Since modules may resolve to a different filename based on the location of the calling module (loading from `node_modules folders`), it is not a guarantee that `require('foo')` will always return the exact same object, if it would resolve to different files.

11.8.7. El Objeto `module` y `module.exports`

1. In each module, the `module` free variable is a reference to the object representing the current module.

2. In particular `module.exports` is the same as the exports object.

3. `module` isn't actually a global but rather local to each module.

4. The exports object is created by the `Module` system. Sometimes this is not acceptable, many want their module to be an instance of some class. To do this assign the desired export object to `module.exports`.

- [~/javascript/node.js/creating_modules/module_exports(master)]$ cat a.js
```
var EventEmitter = require('events').EventEmitter;

module.exports = new EventEmitter();

// Do some work, and after some time emit
// the 'ready' event from the module itself.
setTimeout(function() {
  module.exports.emit('ready');
}, 1000);
```
- $ cat main.js
```
var a = require('./a');
a.on('ready', function() {
  console.log('module a is ready');
});
```
- $ node main.js
```
module a is ready
```

La asignación a `module.exports` debe hacerse inmediatamente. No puede hacerse en un callback.

11.8.8. Algoritmo de Búsqueda Ejecutado por `require`

require(X) from module at path Y

1. If X is a core module,
   a) return the core module
   b) STOP
2. If X begins with './' or '/' or '../'
   a) `LOAD_AS_FILE(Y + X)`
   b) `LOAD_AS_DIRECTORY(Y + X)`
3. `LOAD_NODE_MODULES(X, dirname(Y))`
4. THROW "not found"

LOAD_AS_FILE(X)

1. If X is a file, load X as JavaScript text. STOP
2. If X.js is a file, load X.js as JavaScript text. STOP
3. If X.node is a file, load X.node as binary addon. STOP

LOAD_AS_DIRECTORY(X)

1. If X/package.json is a file, a. Parse X/package.json, and look for "main"field. b. let M = X + (json main field) c. `LOAD_AS_FILE(M)`
2. If X/index.js is a file, load X/index.js as JavaScript text. STOP
3. If X/index.node is a file, load X/index.node as binary addon. STOP

LOAD_NODE_MODULES(X, START)

1. let `DIRS=NODE_MODULES_PATHS(START)`
2. for each DIR in DIRS: a. `LOAD_AS_FILE(DIR/X)` b. `LOAD_AS_DIRECTORY(DIR/X)`

NODE_MODULES_PATHS(START)

1. let PARTS = path split(START)
2. let ROOT = index of first instance of "node_modulesïn PARTS, or 0
3. let I = count of PARTS - 1
4. let DIRS = []
5. while I ¿ROOT, a. if PARTS[I] = "node_modulesÇONTINUE c. DIR = path join(PARTS[0 .. I] + "node_modules") b. DIRS = DIRS + DIR c. let I = I - 1
6. return DIRS

11.9. Como Crear tu Propio Módulo en Node.js

11.9.1. Introducción

Cuando Node carga nuestro fichero JavaScript crea un nuevo ámbito. Cuando estamos en nuestro módulo, no podemos ver el ámbito externo lo que evita las colisiones de nombres.

11.9.2. Un Fichero `package.json`

Creamos un fichero en la raíz de nuestro proyecto con nombre `package.json`. Este fichero describe nuestro proyecto. Es esencial si vamos a publicar nuestro proyecto con `npm`.

Podemos especificar en este fichero:

1. Name, version, description, and keywords to describe your program.
2. A homepage where users can learn more about it.
3. Other packages that yours depends on.

Si hemos instalado `npm` podemos usar el comando `npm init` para empezar. Véase `npm help json` para obtener información sobre este fichero:

La cosa mas importante a especificar cuando estamos escribiendo un programa para su uso por otros, es el módulo `main`. Este consituirá el punto de entrada a nuestro programa.

Es esencial documentar las dependencias.

El siguiente es un ejemplo de fichero `package.json` tomado del proyecto `ebnf-parser`:

```
[~/javascript/PLgrado/ebnf-parser(master)]$ cat -n package.json
   1  {
   2    "name": "ebnf-parser",
   3    "version": "0.1.1",
   4    "description": "A parser for BNF and EBNF grammars used by jison",
   5    "main": "ebnf-parser.js",
   6    "scripts": {
   7      "test": "make test"
   8    },
   9    "repository": "",
  10    "keywords": [
```

```
  11      "bnf",
  12      "ebnf",
  13      "grammar",
  14      "parser",
  15      "jison"
  16    ],
  17    "author": "Zach Carter",
  18    "license": "MIT",
  19    "devDependencies": {
  20      "jison": "0.4.x",
  21      "lex-parser": "0.1.0",
  22      "test": "*"
  23    }
  24  }
```

11.9.3. README y otros documentos

Pon la información basica acerca del módulo en la raíz del proyecto. Como ejemplo veamos el `README.md` (observa que está en formato `markdown`) del proyecto `ebnf-parser`:

```
$ cat README.md

# ebnf-parser

A parser for BNF and EBNF grammars used by jison.

## install

    npm install ebnf-parser


## build

To build the parser yourself, clone the git repo then run:

    make

This will generate ‘parser.js‘, which is required by ‘ebnf-parser.js‘.

## usage

The parser translates a string grammar or JSON grammar into a JSON grammar that jison can use

    var ebnfParser = require('ebnf-parser');

    // parse a bnf or ebnf string grammar
    ebnfParser.parse("%start ... %");

    // transform an ebnf JSON gramamr
    ebnfParser.transform({"ebnf": ...});


## example grammar
```

The parser can parse its own BNF grammar, shown below:

```
%start spec

/* grammar for parsing jison grammar files */

%{
var transform = require('./ebnf-transform').transform;
var ebnf = false;
%}

%%

spec
    : declaration_list '%%' grammar optional_end_block EOF
        {$$ = $1; return extend($$, $3);}
    | declaration_list '%%' grammar '%%' CODE EOF
        {$$ = $1; yy.addDeclaration($$,{include:$5}); return extend($$, $3);}
    ;

optional_end_block
    :
    | '%%'
    ;

declaration_list
    : declaration_list declaration
        {$$ = $1; yy.addDeclaration($$, $2);}
    |
        {$$ = {};}
    ;

declaration
    : START id
        {$$ = {start: $2};}
    | LEX_BLOCK
        {$$ = {lex: $1};}
    | operator
        {$$ = {operator: $1};}
    | ACTION
        {$$ = {include: $1};}
    ;

operator
    : associativity token_list
        {$$ = [$1]; $$.push.apply($$, $2);}
    ;

associativity
    : LEFT
        {$$ = 'left';}
    | RIGHT
        {$$ = 'right';}
```

```
    | NONASSOC
        {$$ = 'nonassoc';}
    ;

token_list
    : token_list symbol
        {$$ = $1; $$.push($2);}
    | symbol
        {$$ = [$1];}
    ;

grammar
    : production_list
        {$$ = $1;}
    ;

production_list
    : production_list production
        {$$ = $1;
          if($2[0] in $$) $$[$2[0]] = $$[$2[0]].concat($2[1]);
          else  $$[$2[0]] = $2[1];}
    | production
        {$$ = {}; $$[$1[0]] = $1[1];}
    ;

production
    : id ':' handle_list ';'
        {$$ = [$1, $3];}
    ;

handle_list
    : handle_list '|' handle_action
        {$$ = $1; $$.push($3);}
    | handle_action
        {$$ = [$1];}
    ;

handle_action
    : handle prec action
        {$$ = [($1.length ? $1.join(' ') : '')];
            if($3) $$.push($3);
            if($2) $$.push($2);
            if ($$.length === 1) $$ = $$[0];
        }
    ;

handle
    : handle expression_suffix
        {$$ = $1; $$.push($2)}
    |
        {$$ = [];}
    ;
```

```
handle_sublist
    : handle_sublist '|' handle
        {$$ = $1; $$.push($3.join(' '));}
    | handle
        {$$ = [$1.join(' ')];}
    ;

expression_suffix
    : expression suffix
        {$$ = $expression + $suffix; }
    ;

expression
    : ID
        {$$ = $1; }
    | STRING
        {$$ = ebnf ? "'"+$1+"'" : $1; }
    | '(' handle_sublist ')'
        {$$ = '(' + $handle_sublist.join(' | ') + ')'; }
    ;

suffix
    : {$$ = ''}
    | '*'
    | '?'
    | '+'
    ;

prec
    : PREC symbol
        {$$ = {prec: $2};}
    |
        {$$ = null;}
    ;

symbol
    : id
        {$$ = $1;}
    | STRING
        {$$ = yytext;}
    ;

id
    : ID
        {$$ = yytext;}
    ;

action
    : '{' action_body '}'
        {$$ = $2;}
    | ACTION
        {$$ = $1;}
    | ARROW_ACTION
```

```
            {$$ = '$$ ='+$1+';';}
        |
            {$$ = '';}
        ;

    action_body
        :
            {$$ = '';}
        | ACTION_BODY
            {$$ = yytext;}
        | action_body '{' action_body '}' ACTION_BODY
            {$$ = $1+$2+$3+$4+$5;}
        | action_body '{' action_body '}'
            {$$ = $1+$2+$3+$4;}
        ;

    %%

    // transform ebnf to bnf if necessary
    function extend (json, grammar) {
        json.bnf = ebnf ? transform(grammar) : grammar;
        return json;
    }

## license

MIT
```

En general se anima a que uses el formato markdown. Salva el fichero como `README.md`.

La documentación adicional se pone en un directorio `./docs`. Los ficheros markdown teminan en `.md` y los html en `.html`.

11.9.4. Véase También

- How To: Create Your Own Node.js Module por Isaac Z. Schlueter autor de npm. Véase también este gist en GitHub
- Creating Custom Modules
- How to Build a Nodejs Npm Package From Scratch. May 2012 Decodize

11.10. Mas sobre Node

1. Véase el libro 'Learning Node' de S. Powers [2].
2. Node.js
3. docs.nodejitsu.com: collection of node.js how-to articles. These articles range from basic to advanced. They provide relevant code samples and insights into the design and philosophy of node itself
4. http://howtonode.org/ contiene un número creciente de tutoriales
5. El manual de node.js puede encontrarse en formato pdf en el proyecto https://github.com/zeMirco/nodejs-pd en GitHub. En concreto en este enlace

6. Guías de node.js de Felix Geisendörfer
7. Introduction to Node.js por Hector Correa
8. El Libro para Principiantes en Node.js por Manuel Kiessling y Herman A. Junge

Para aprender JavaScript podemos usar el libro eloquent JavaScript de Marijn Haverbeke [3].

Capítulo 12

JavaScript en los Navegadores

Capítulo 13

El Objeto Window

Capítulo 14

Manejo de Documentos en JavaScript

Capítulo 15

Manejo de CSS

Capítulo 16

Manejo de Eventos

Capítulo 17

HTTP

Capítulo 18

La Librería JQuery

Capítulo 19

Almacenamiento en el Cliente

Capítulo 20

Multimedia y Gráficos

Capítulo 21

HTML5

Capítulo 22

XML

Véase el libro Sams Teach Yourself XML in 21 Days en Google o en informit [4]

Capítulo 23

Backbone

1. Developing Backbone.js Applications

Capítulo 24

Closure Tools

24.1. Véase También

1. Google I/O 2011: JavaScript Programming in the Large with Closure Tools (YouTube)
2. The Closure Tools project is an effort by Google to open source the tools used in many of Google's sites ar
3. Herramientas:
   a) Closure Compiler en Google-Code
   b) Closure Library en Google-Code
   c) Closure Template en Google-Code
   d) Closure Linter en Google-Code
4. Getting Started with the Closure Library (Hello World!)

Capítulo 25

Pruebas

25.1. Testing en JavaScript: Fácil y Rápido

1. Quick Tip: Quick and Easy JavaScript Testing with "Assert" por Jeffrey Way

25.2. Unit Testing, TDD y BDD con Jasmine

1. Jasmine: BDD Style JavaScript Testing Hello World por Chris McNabb (YouTube Sep. 2012)
2. Download Jasmine
3. Testing Your JavaScript with Jasmine Andrew Burgess on Aug 4th 2011
4. Unit Testing in JavaScript via Jasmine (Youtube)
5. Jasmine en GitHub
6. Behavior Driven Testing with Jasmine (YouTube, Davis Frank de Pivotal Labs, Contiene una introducción a BDD)
7. Jasmine Wiki
8. Testem tutorial en net.tutplus (trabaja sobre Jasmine y sobre Coffee)
9. Jasmine Matchers: Class jasmine.Matchers

Capítulo 26

Buenas Prácticas y Patrones

26.1. Véase También

1. Google I/O 2011: Learning to Love JavaScript por Alex Russell, Mayo 2011. ¡Excelente!
2. traceur compiler
3. Google I/O 2011: JavaScript Programming in the Large with Closure Tools

Capítulo 27

Herramientas

27.1. npm

27.2. n

`n` es una herramienta parecida a `rvm` para Node.js:

```
$sudo npm install n -g
```

y

```
[~/Dropbox/src/javascript/node.js/creating_modules(master)]$ n help

  Usage: n [options] [COMMAND] [config]

  Commands:

    n                           Output versions installed
    n latest [config ...]       Install or activate the latest node release
    n stable [config ...]       Install or activate the latest stable node release
    n <version> [config ...]    Install and/or use node <version>
    n use <version> [args ...]  Execute node <version> with [args ...]
    n bin <version>             Output bin path for <version>
    n rm <version ...>          Remove the given version(s)
    n prev                      Revert to the previously activated version
    n --latest                  Output the latest node version available
    n --stable                  Output the latest stable node version available
    n ls                        Output the versions of node available

  Options:

    -V, --version   Output current version of n
    -h, --help      Display help information

  Aliases:

    which   bin
    use     as
    list    ls
    -       rm
```

```
$ sudo n latest
```

27.3. Google Chrome y Javascript

1. Building Browser Apps with Google Chrome

Enlaces Relacionados

1. Chrome Developer Tools Tutorial: Elements (Part 1/2)

27.4. Plugins, Editores, IDEs

- jslint lint plugin para vim
- vim plugins for HTML and CSS hi-speed coding. disponible en http://www.vim.org y en github https://github.com/mattn/zencoding-vim/
- Vim Essential Plugin: Sparkup parecido a zenconding. El tutorial es de 2011.

27.5. Grunt

1. Grunt - The Basics Youtube
2. Grunt home page

27.6. Beautifiers, Pretty-Printers

1. beautifier de Rickeyski

27.7. Modulos

1. NODE.JS Modules

Parte II

SEGUNDA PARTE: APUNTES DE PROCESADORES DE LENGUAJES

Capítulo 28

Expresiones Regulares y Análisis Léxico en JavaScript

28.1. Mozilla Developer Network: Documentación

1. RegExp Objects
2. exec
3. search
4. match
5. replace

28.2. Práctica: Conversor de Temperaturas

index.html

```
<html>
  <head>
      <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
      <title>JavaScript Temperature Converter</title>
      <link href="global.css" rel="stylesheet" type="text/css">

     <script type="text/javascript" src="temperature.js"></script>
  </head>
  <body>
    <h1>Temperature Converter</h1>
    <table>
      <tr>
        <th>Enter  Temperature (examples: 32F, 45C, -2.5f):</th>
        <td><input id="original" onchange="calculate();"></td>
      </tr>
      <tr>
        <th>Converted Temperature:</th>
        <td><span class="output" id="converted"></span></td>
      </tr>
    </table>
  </body>
</html>
```

global.css

```
th, td      { vertical-align: top; text-align: right; font-size:large; }
#converted  { color: red; font-weight: bold; font-size:large;          }
input       { text-align: right;  border: none; font-size:large;       }
body
{
 background-color:#b0c4de;  /* blue */
 font-size:large;
}
```

temperature.js

```
"use strict"; // Use ECMAScript 5 strict mode in browsers that support it
function calculate() {
  var result;
  var original       = document.getElementById("........");
  var temp = original.value;
  var regexp = /............................../;

  var m = temp.match(......);

  if (m) {
    var num = ....;
    var type = ....;
    num = parseFloat(num);
    if (type == 'c' || type == 'C') {
      result = (num * 9/5)+32;
      result = ..............................
    }
    else {
      result = (num - 32)*5/9;
      result = ............................
    }
    converted.innerHTML = result;
  }
  else {
    converted.innerHTML = "ERROR! Try something like '-4.2C' instead";
  }
}
```

28.3. Práctica: Comma Separated Values

Introducción al formato CSV

Véase Comma Separated Values en la Wikipedia:

A comma-separated values (CSV) file stores tabular data (numbers and text) in plain-text form. A CSV file consists of any number of records, separated by line breaks of some kind; each record consists of fields, separated by a comma. All records have an identical sequence of fields.

Ejemplo de ejecución

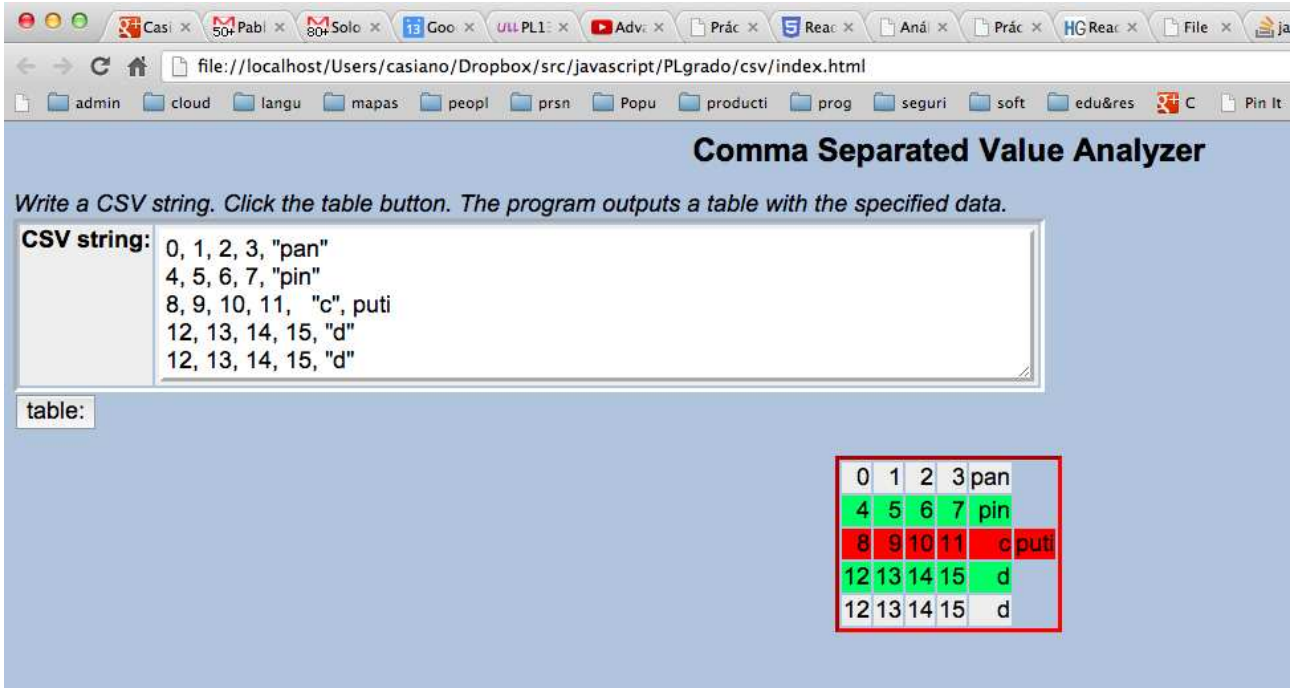


Figura 28.1: Ejemplo de pantalla de La aplicación para el Análisis de Datos en Formato CSV

Aproximación al análisis mediante expresiones regulares de CSV Una primera aproximación sería hacer `split` por las comas:

```
> x = '"earth",1,"moon",9.374'
'"earth",1,"moon",9.374'
> y = x.split(/,/)
[ '"earth"', '1', '"moon"', '9.374' ]
```

Esta solución deja las comillas dobles en los campos entrecomillados. Peor aún, los campos entrecomillados pueden contener comas, en cuyo caso la división proporcionada por `split` sería errónea:

```
> x = '"earth, mars",1,"moon, fobos",9.374'
'"earth, mars",1,"moon, fobos",9.374'
> y = x.split(/,/)
[ '"earth', ' mars"', '1', '"moon', ' fobos"', '9.374' ]
```

La siguiente expresión regular reconoce cadenas de comillas dobles con secuencias de escape seguidas opcionalmente de una coma:

```
> x = '"earth, mars",1,"moon, fobos",9.374'
'"earth, mars",1,"moon, fobos",9.374'
> r = /"((?:[^"\\]|\\.)*)"\s*,?/g
/"((?:[^"\\]|\\.)*)"\s*,?/g
> w = x.match(r)
[ '"earth, mars",', '"moon, fobos",' ]
```

Esta otra expresión regular `/([^,]+),?|\s*,/` actúa de forma parecida al `split`. Reconoce secuencias no vacías de caracteres que no contienen comas seguidas opcionalmente de una coma o bien una sóla coma (precedida opcionalmente de blancos):

```
> x = '"earth, mars",1,"moon, fobos",9.374'
'"earth, mars",1,"moon, fobos",9.374'
> r = /([^,]+),?|\s*,/g
/([^,]+),?|\s*,/g
> w = x.match(r)
[ '"earth,', ' mars",', '1,', '"moon,', ' fobos",', '9.374' ]
```

index.html

```
<html>
  <head>
     <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
     <title>CSV Analyzer</title>
     <link href="global.css" rel="stylesheet" type="text/css">

     <script type="text/javascript" src="../../underscore/underscore.js"></script>
     <script type="text/javascript" src="../../jquery/starterkit/jquery.js"></script>
     <script type="text/javascript" src="csv.js"></script>
  </head>
  <body>
    <h1>Comma Separated Value Analyzer</h1>
    <div>
      <i>Write a CSV string. Click the table button. The program outputs a table with the spec
    </div>
    <table>
      <tr>
        <th>CSV string:</th> <!-- autofocus attribute is HTML5 -->
        <td><textarea autofocus cols = "80" rows = "5" id="original"></textarea></td>
      </tr>
    </table>
    <button type="button">table:</button><br>
    <span class="output" id="finaltable"></span>
  </body>
</html>
```

1. jQuery (Descarga la librería)
2. Underscore
3. autofocus. Véase también [5]
4. textarea
5. button

global.css

```
html *
{
   font-size: large;
   /* The !important ensures that nothing can override what you've set in this style (unless i
   font-family: Arial;
```

```
}

h1            { text-align: center; font-size: x-large; }
th, td        { vertical-align: top; text-align: right; }
/* #finaltable  * { color: white; background-color: black; }   */

/* #finaltable table { border-collapse:collapse; } */
/* #finaltable table, td { border:1px solid white; } */
#finaltable:hover td { background-color: blue; }
tr:nth-child(odd)    { background-color:#eee; }
tr:nth-child(even)    { background-color:#00FF66; }
input        { text-align: right;  border: none;       }     /* Align input to the right  */
textarea     { border: outset; border-color: white;       }
table        { border: inset; border-color: white; }
table.center { margin-left:auto; margin-right:auto; }
#result      { border-color: red; }
tr.error       { background-color: red; }
body
{
 background-color:#b0c4de;  /* blue */
}
```

1. Introducción a las pseudo clases de CSS3
2. CSS pattern matching
3. CSS class selectors
4. CSS3: nth-child() selector
5. CSS: border
6.

csv.js

```
// See http://en.wikipedia.org/wiki/Comma-separated_values
"use strict"; // Use ECMAScript 5 strict mode in browsers that support it

$(document).ready(function() {
   $("button").click(function() {
     calculate();
   });
 });

function calculate() {
  var result;
  var original       = document.getElementById("original");
  var temp = original.value;
  var regexp = /_____________________________________________/g;
  var lines = temp.split(/\n+\s*/);
  var commonLength = NaN;
  var r = [];
  // Template using underscore
```

```
  var row = "<%% _.each(items, function(name) { %>"     +
            "                      <td><%%= name %></td>" +
            "               <%% }); %>";

  if (window.localStorage) localStorage.original  = temp;

  for(var t in lines) {
    var temp = lines[t];
    var m = temp.match(regexp);
    var result = [];
    var error = false;

    if (m) {
      if (commonLength && (commonLength != m.length)) {
        //alert('ERROR! row <'+temp+'> has '+m.length+' items!');
        error = true;
      }
      else {
        commonLength = m.length;
        error = false;
      }
      for(var i in m) {
        var removecomma = m[i].replace(/,\s*$/,'');
        var remove1stquote = removecomma.replace(/^\s*"/,'');
        var removelastquote = remove1stquote.replace(/"\s*$/,'');
        var removeescapedquotes = removelastquote.replace(/\"/,'"');
        result.push(removeescapedquotes);
      }
      var tr = error? '<tr class="error">' : '<tr>';
      r.push(tr+_.template(row, {items : result})+"</tr>");
    }
    else {
      alert('ERROR! row '+temp+' does not look as legal CSV');
      error = true;
    }
  }
  r.unshift('<p>\n<table class="center" id="result">');
  r.push('</table>');
  //alert(r.join('\n')); // debug
  finaltable.innerHTML = r.join('\n');
}

window.onload = function() {
  // If the browser supports localStorage and we have some stored data
  if (window.localStorage && localStorage.original) {
    document.getElementById("original").value = localStorage.original;
  }
};
```

1. Tutorials:Getting Started with jQuery
2. Underscore: template

3. Stackoverflow::how to use Underscore template
4. Como usar localstorage
5. HTML5 Web Storage

28.4. Ejercicios

1. Paréntesis:

```
> str = "John Smith"
'John Smith'
> newstr = str.replace(re, "$2, $1")
'Smith, John'
```

2. El método `exec`.

   If your regular expression uses the `g` flag, you can use the `exec` method multiple times to find successive matches in the same string. When you do so, the search starts at the substring of str specified by the regular expression's `lastIndex` property.

```
> re = /d(b+)(d)/ig
/d(b+)(d)/gi
> z = "dBdxdbbdzdbd"
'dBdxdbbdzdbd'
> result = re.exec(z)
[ 'dBd', 'B', 'd', index: 0, input: 'dBdxdbbdzdbd' ]
> re.lastIndex
3
> result = re.exec(z)
[ 'dbbd', 'bb', 'd', index: 4, input: 'dBdxdbbdzdbd' ]
> re.lastIndex
8
> result = re.exec(z)
[ 'dbd', 'b', 'd', index: 9, input: 'dBdxdbbdzdbd' ]
> re.lastIndex
12
> z.length
12
> result = re.exec(z)
null
```

3. JavaScript tiene lookaheads:

```
> x = "hello"
'hello'
> r = /l(?=o)/
/l(?=o)/
> z = r.exec(x)
[ 'l', index: 3, input: 'hello' ]
```

4. JavaScript no tiene lookbehinds:

```
> x = "hello"
'hello'
> r = /(?<=l)l/
SyntaxError: Invalid regular expression: /(?<=l)l/: Invalid group
> .exit

[~/Dropbox/src/javascript/PLgrado/csv(master)]$ irb
ruby-1.9.2-head :001 > x = "hello"
 => "hello"
ruby-1.9.2-head :002 > r = /(?<=l)l/
 => ll
ruby-1.9.2-head :008 > x =~ r
 => 3
ruby-1.9.2-head :009 > $&
 => "l"
```

5. El siguiente ejemplo comprueba la validez de números de teléfono:

```
$ cat phone.html
<!DOCTYPE html>
<html>
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
    <meta http-equiv="Content-Script-Type" content="text/javascript">
    <script type="text/javascript">
      var re = /\(?\d{3}\)?([-\/\.])\d{3}\1\d{4}/;
      function testInfo(phoneInput){
        var OK = re.exec(phoneInput.value);
        if (!OK)
          window.alert(RegExp.input + " isn't a phone number with area code!");
        else
          window.alert("Thanks, your phone number is " + OK[0]);
      }
    </script>
  </head>
  <body>
    <p>Enter your phone number (with area code) and then click "Check".
        <br>The expected format is like ###-###-####.</p>
    <form action="#">
      <input id="phone"><button onclick="testInfo(document.getElementById('phone'));">Che
    </form>
  </body>
</html>
```

6. ¿Con que cadenas casa la expresión regular `/^(11+)\1+$/`?

```
> '1111'.match(/^(11+)\1+$/) # 4 unos
[ '1111',
  '11',
  index: 0,
  input: '1111' ]
> '111'.match(/^(11+)\1+$/) # 3 unos
null
```

```
> '11111'.match(/^(11+)\1+$/) # 5 unos
null
> '111111'.match(/^(11+)\1+$/) # 6 unos
[ '111111',
  '111',
  index: 0,
  input: '111111' ]
> '11111111'.match(/^(11+)\1+$/) # 8 unos
[ '11111111',
  '1111',
  index: 0,
  input: '11111111' ]
> '1111111'.match(/^(11+)\1+$/)
null
>
```

Busque una solución al siguiente ejercicio (véase 'Regex to add space after punctuation sign' en PerlMonks) Se quiere poner un espacio en blanco después de la aparición de cada coma:

7.
```
> x = "a,b,c,1,2,d, e,f"
'a,b,c,1,2,d, e,f'
> x.replace(/,/g,", ")
'a, b, c, 1, 2, d,  e, f'
```

pero se quiere que la sustitución no tenga lugar si la coma esta incrustada entre dos dígitos. Además se pide que si hay ya un espacio después de la coma, no se duplique.

a) La siguiente solución logra el segundo objetivo, pero estropea los números:

```
> x = "a,b,c,1,2,d, e,f"
'a,b,c,1,2,d, e,f'
> x.replace(/,(\S)/g,", $1")
'a, b, c, 1, 2, d, e, f'
```

b) Esta otra funciona bien con los números pero no con los espacios ya existentes:

```
> x = "a,b,c,1,2,d, e,f"
'a,b,c,1,2,d, e,f'
> x.replace(/,(\D)/g,", $1")
'a, b, c,1,2, d,  e, f'
```

c) Explique cuando casa esta expresión regular:

```
> r = /(\d[,.]\d)|(,(?=\S))/g
/(\d[,.]\d)|(,(?=\S))/g
```

Aproveche que el método `replace` puede recibir como segundo argumento una función (vea replace):

```
> z = "a,b,1,2,d, 3,4,e"
'a,b,1,2,d, 3,4,e'
> f = function(match, p1, p2, offset, string) { return (p1 || p2 + " "); }
[Function]
> z.replace(r, f)
'a, b, 1,2, d, 3,4, e'
```

28.5. Práctica: Palabras Repetidas

Se trata de producir una salida en las que las palabras repetidas consecutivas sean reducidas a una sola aparición. Rellena las partes que faltan.

Ejemplo de ejecución

Figura 28.2: Ejemplo de pantalla de La aplicación para el Análisis de Datos en Formato CSV

Estructura

index.html

```
[~/Dropbox/src/javascript/PLgrado/repeatedwords(master)]$ cat index.html
<html>
  <head>
     <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
     <title>File Input</title>
     <link href="global.css" rel="stylesheet" type="text/css">

     <script type="text/javascript" src="../../underscore/underscore.js"></script>
     <script type="text/javascript" src="../../jquery/starterkit/jquery.js"></script>
     <script type="text/javascript" src="repeated_words.js"></script>
  </head>
  <body>
    <h1>File Input</h1>
    <input type="file" id="fileinput" />
    <div id="out" class="hidden">
    <table>
      <tr><th>Original</th><th>Transformed</th></tr>
      <tr>
        <td>
          <pre class="input" id="initialinput"></pre>
        </td>
        <td>
          <pre class="output" id="finaloutput"></pre>
        </td>
      </tr>
    </table>
    </div>
  </body>
</html>
```

1. Tag input

global.css

Rellena los estilos para `hidden` y `unhidden`:

```
[~/Dropbox/src/javascript/PLgrado/repeatedwords(master)]$ cat global.css
html *
{
   font-size: large;
   /* The !important ensures that nothing can override what you've set in this style (unless i
   font-family: Arial;
}

.thumb {
```

```
    height: 75px;
    border: 1px solid #000;
    margin: 10px 5px 0 0;
  }

h1            { text-align: center; font-size: x-large; }
th, td        { vertical-align: top; text-align: right; }
/* #finaltable  * { color: white; background-color: black; }   */

/* #finaltable table { border-collapse:collapse; } */
/* #finaltable table, td { border:1px solid white; } */
#finaltable:hover td { background-color: blue; }
tr:nth-child(odd)    { background-color:#eee; }
tr:nth-child(even)    { background-color:#00FF66; }
input        { text-align: right;  border: none;       }     /* Align input to the right  */
textarea     { border: outset; border-color: white;       }
table        { border: inset; border-color: white; }
.hidden      { display: ____; }
.unhidden    { display: _____; }
table.center { margin-left:auto; margin-right:auto; }
#result      { border-color: red; }
tr.error       { background-color: red; }
pre.output   { background-color: white; }
span.repeated { background-color: red }

body
{
 background-color:#b0c4de;  /* blue */
```

1. CSS display Property
2. Diferencias entre "Displayz "Visibility"

repeated_words.js

Rellena las expresiones regulares que faltan:

```
[~/srcPLgrado/repeatedwords(master)]$ cat repeated_words.js
"use strict"; // Use ECMAScript 5 strict mode in browsers that support it

$(document).ready(function() {
   $("#fileinput").change(calculate);
});

function generateOutput(contents) {
  return contents.replace(/____________________/__,'_____________________________________');
}

function calculate(evt) {
  var f = evt.target.files[0];
  var contents = '';

  if (f) {
    var r = new FileReader();
    r.onload = function(e) {
```

```
      contents = e.target.result;
      var escaped  = escapeHtml(contents);
      var outdiv = document.getElementById("out");
      outdiv.className = 'unhidden';
      finaloutput.innerHTML = generateOutput(escaped);
      initialinput.innerHTML = escaped;

    }
    r.readAsText(f);
  } else {
    alert("Failed to load file");
  }
}

var entityMap = {
    "&": "&amp;",
    "<": "&lt;",
    ">": "&gt;",
    '"': '&quot;',
    "'": '&#39;',
    "/": '&#x2F;'
  };

function escapeHtml(string) {
  return String(string).replace(/_________/g, function (s) {
    return ____________;
  });
```

1. jQuery event.target
2. HTML 5 File API
3. HTML 5 File API: FileReader
4. HTML 5 File API: FileReader
5. element.className
6. HTML Entities
7. Tutorials:Getting Started with jQuery
8. Underscore: template

Ficheros de Entrada

```
[~/Dropbox/src/javascript/PLgrado/repeatedwords(master)]$ cat input2.txt
habia una vez
vez un viejo viejo
hidalgo que vivia
vivia
[~/Dropbox/src/javascript/PLgrado/repeatedwords(master)]$ cat input.txt
one one
nothing rep
is two three
three four
```

```
[~/Dropbox/src/javascript/PLgrado/repeatedwords(master)]$ cat inputhtml1.txt
habia => una vez
vez & un viejo viejo <puchum>
hidalgo  & <pacham> que vivia
vivia </que se yo>
```

28.6. Ejercicios

El formato *INI* es un formato estandar para la escritura de ficheros de configuración. Su estrucutra básica se compone de "seccionesz "propiedades". Véase la entrada de la wikipedia INI.

```
; last modified 1 April 2001 by John Doe
[owner]
name=John Doe
organization=Acme Widgets Inc.

[database]
; use IP address in case network name resolution is not working
server=192.0.2.62
port=143
file = "payroll.dat"
```

1. Escriba un programa javascript que obtenga las cabeceras de sección de un fichero INI
2. Escriba un programa javascript que case con los bloques de un fichero INI (cabecera mas lista de pares `parámetro=valor`)
3. Se quieren obtener todos los pares nombre-valor, usando paréntesis con memoria para capturar cada parte.
4. ¿Que casa con cada paréntesis en esta regexp para los pares nombre-valor?

   ```
   > x = "h     = 4"
   > r = /([^=]*)(\s*)=(\s*)(.*)/
   > r.exec(x)
   >
   ```

28.7. Ejercicios

1. Escriba una expresión regular que reconozca las cadenas de doble comillas. Debe permitir la presencia de comillas y caracteres escapados.
2. ¿Cual es la salida?

   ```
   > "bb".match(/b|bb/)

   > "bb".match(/bb|b/)
   ```

28.8. Práctica: Ficheros INI

index.html

```
<html>
  <head>
     <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
     <title>INI files</title>
     <link href="global.css" rel="stylesheet" type="text/css">

     <script type="text/javascript" src="../../underscore/underscore.js"></script>
     <script type="text/javascript" src="../../jquery/starterkit/jquery.js"></script>
     <script type="text/javascript" src="______"></script>
  </head>
  <body>
    <h1>INI files</h1>
    <input type="file" id="_________" />
    <div id="out" class="hidden">
    <table>
      <tr><th>Original</th><th>Tokens</th></tr>
      <tr>
        <td>
          <pre class="input" id="____________"></pre>
        </td>
        <td>
          <pre class="output" id="____________"></pre>
        </td>
      </tr>
    </table>
    </div>
  </body>
</html>
```

global.css

```
[~/Dropbox/src/javascript/PLgrado/ini(master)]$ cat global.css
html *
{
   font-size: large;
   /* The !important ensures that nothing can override what you've set in this style (unless i
   font-family: Arial;
}

.thumb {
    height: 75px;
    border: 1px solid #000;
    margin: 10px 5px 0 0;
  }

h1            { text-align: center; font-size: x-large; }
th, td        { vertical-align: top; text-align: left; }
/* #finaltable  * { color: white; background-color: black; }   */

/* #finaltable table { border-collapse:collapse; } */
/* #finaltable table, td { border:1px solid white; } */
#finaltable:hover td { background-color: blue; }
tr:nth-child(odd)    { background-color:#eee; }
tr:nth-child(even)    { background-color:#00FF66; }
```

```
input        { text-align: right;  border: none;       }     /* Align input to the right  */
textarea     { border: outset; border-color: white;       }
table        { border: inset; border-color: white; }
.hidden      { display: ____; }
.unhidden    { display: _____; }
table.center { margin-left:auto; margin-right:auto; }
#result      { border-color: red; }
tr.error       { background-color: red; }
pre.output   { background-color: white; }
span.repeated { background-color: red }
span.header { _________________: ____ }
span.comments { _________________: ______ }
span.blanks { _________________: _____ }
span.nameEqualValue { _________________: ____ }
span.error { _________________: ___ }

body
{
 background-color:#b0c4de;  /* blue */
}
```

Ficheros de Prueba

```
~/Dropbox/src/javascript/PLgrado/ini(master)]$ cat input.ini
; last modified 1 April 2001 by John Doe
[owner]
name=John Doe
organization=Acme Widgets Inc.

[database]
; use IP address in case network name resolution is not working
server=192.0.2.62
port=143
file = "payroll.dat"

$ cat input2.ini
[special_fields]
required = "EmailAddr,FirstName,LastName,Mesg"
csvfile = "contacts.csv"
csvcolumns = "EmailAddr,FirstName,LastName,Mesg,Date,Time"

[email_addresses]
sales = "jack@yahoo.com,mary@my-sales-force.com,president@my-company.com"

$ cat inputerror.ini
[owner]
name=John Doe
organization$Acme Widgets Inc.

[database
; use IP address in case network name resolution is not working
server=192.0.2.62
port=143
file = "payroll.dat"
```

ini.js

```
[~/Dropbox/src/javascript/PLgrado/ini(master)]$ cat ini.js
"use ______"; // Use ECMAScript 5 strict mode in browsers that support it

$(document)._____(function() {
   $("#fileinput")._____(calculate);
});

function calculate(evt) {
  var f = evt.target.files[0];

  if (f) {
    var r = new __________();
    r.onload = function(e) {
      var contents = e.target._____;

      var tokens = lexer(contents);
      var pretty = tokensToString(tokens);

      out.className = 'unhidden';
      initialinput._________ = contents;
      finaloutput._________ = pretty;
    }
    r.__________(f); // Leer como texto
  } else {
    alert("Failed to load file");
  }
}

var temp = '<li> <span class = "<%= _____ %>"> <%= _ %> </span>\n';

function tokensToString(tokens) {
   var r = '';
   for(var i in tokens) {
     var t = tokens[i];
     var s = JSON.stringify(t, undefined, 2);
     s = _.template(temp, {t: t, s: s});
     r += s;
   }
   return '<ol>\n'+r+'</ol>';
}

function lexer(input) {
  var blanks         = /^___/;
  var iniheader      = /^_____________/;
  var comments       = /^________/;
  var nameEqualValue = /^________________________/;
  var any            = /^_______/;

  var out = [];
  var m = null;

  while (input != '') {
```

```
    if (m = blanks.____(input)) {
      input = input.substr(m.index+___________);
      out.push({ type : ________, match: _ });
    }
    else if (m = iniheader.exec(input)) {
      input = input.substr(__________________);
      ____________________________________ // avanzemos en input
    }
    else if (m = comments.exec(input)) {
      input = input.substr(__________________);
      ________________________________________
    }
    else if (m = nameEqualValue.exec(input)) {
      input = input.substr(__________________);
      ____________________________________________
    }
    else if (m = any.exec(input)) {
      ____________________________________
      input = '';
    }
    else {
      alert("Fatal Error!"+substr(input,0,20));
      input = '';
    }
  }
  return out;
}
```

1. JSON.stringify
2. www.json.org
3. JSON in JavaScript
4. Underscore: template
5. Stackoverflow::how to use Underscore template

28.9. Práctica: Analizador Léxico para Un Subconjunto de JavaScript

1. Clone el proyecto ull-etsii-grado-pl-minijavascript en github. Rellene las partes que faltan en cada fichero.

28.10. Definición y Diseño de Lenguajes

28.11. Expresiones Regulares Posix en C

Las Expresiones Regulares 'a la Perl' no forman parte de ANSI C. La forma mas sencilla de usar regexps en C es utilizando la versión POSIX de la librería que viene con la mayoría de los Unix. Sigue un ejemplo que usa regex POSIX en C:

```
[~/src/C/regexp]$ cat use_posix.c
#include <stdio.h>
```

```
#include <stdlib.h>
#include <regex.h>

int main() {
       regex_t regex;
       int reti;
       char msgbuf[100];

/* Compile regular expression */
        reti = regcomp(&regex, "^a[[:alnum:]]", 0);
        if( reti ){ fprintf(stderr, "Could not compile regex\n"); exit(1); }

/* Execute regular expression */
        reti = regexec(&regex, "abc", 0, NULL, 0);
        if( !reti ){
                puts("Match");
        }
        else if( reti == REG_NOMATCH ){
                puts("No match");
        }
        else{
                regerror(reti, &regex, msgbuf, sizeof(msgbuf));
                fprintf(stderr, "Regex match failed: %s\n", msgbuf);
                exit(1);
        }

/* Free compiled regular expression if you want to use the regex_t again */
    regfree(&regex);
}
```

Compilación y ejecución:

```
[~/src/C/regexp]$ cc use_posix.c  -o use_posix
[~/src/C/regexp]$ ./use_posix
Match
```

Enlaces Relacionados

1. Regular Expression Matching in GNU C
2. Pattern Matching in GNU C
3. PCRE
4. PCRE doc

Capítulo 29

Expresiones Regulares en Flex

Puede encontrar los ejemplos de este capítulo en https://github.com/crguezl/flex-examples.

Un lenguaje regular es aquel que puede ser descrito mediante expresiones regulares como las que se utilizan en `ex`, `vi`, `sed`, `perl` y en tantas otras utilidades UNIX. Dado un lenguaje regular, un analizador léxico es un programa capaz de reconocer las entradas que pertenecen a dicho lenguaje y realizar las acciones semánticas que se hayan asociado con los estados de aceptación. Un generador de analizadores léxicos es una herramienta que facilita la construcción de un analizador léxico. Un generador de analizadores léxicos parte, por tanto, de un lenguaje adecuado para la descripción de lenguajes regulares (y de su semántica) y produce como salida una función (en C, por ejemplo) que materializa el correspondiente analizador léxico. La mayor parte de los generadores producen a partir del conjunto de expresiones regulares los correspondientes tablas de los autómatas finitos deterministas. Utilizando dichas tablas y un algoritmo de simulación genérico del autómata finito determinista se obtiene el analizador léxico. Una vez obtenido el estado de aceptación a partir de la entrada es posible, mediante una sentencia `switch` ejecutar la acción semántica asociada con la correspondiente expresión regular.

29.1. Estructura de un programa LEX

Estructura de un programa

`LEX` y `FLEX` son ejemplos de generadores léxicos. `Flex` lee desde la entrada estándar si no se especifica explícitamente un fichero de entrada. El fichero de entrada `reglen.l` (se suele usar el tipo `l`) debe tener la forma:

```
%{
declaration C1
.
.
.

declaration CM
%}
macro_name1 regular_definition1
.
.
.

macro_nameR regular_definitionR

%x exclusive_state
%s inclusive_state
%%
```

```
regular_expression1 { action1(); }
.
.
.

regular_expressionN { actionN(); }

%%
support_routine1() {
}
.
.
.

support_routineS() {
}
```

Como vemos, un programa `LEX` consta de 3 secciones, separadas por `%%`. La primera sección se denomina *sección de definiciones*, la segunda *sección de reglas* y la tercera *sección de código*. La primera y la última son opcionales, así el programa legal `LEX` mas simple es:

```
%%
```

que genera un analizador que copia su entrada en `stdout`.

Compilación

Una vez compilado el fichero de entrada `regleng.l` mediante la correspondiente orden:

```
flex reglen.l
```

obtenemos un fichero denominado `lex.yy.c`. Este fichero contiene la rutina `yylex()` que realiza el análisis léxico del lenguaje descrito en `regleng.l`. Supuesto que una de las `support_routines` es una función `main()` que llama a la función `yylex()`, podemos compilar el fichero generado con un compilador C para obtener un ejecutable `a.out`:

```
cc lex.yy.c -lfl
```

La inclusión de la opción `-fl` enlaza con la librería de `flex`, que contiene dos funciones: `main` y `yywrap()`.

Ejecución

Cuando ejecutamos el programa `a.out`, la función `yylex()` analiza las entradas, buscando la secuencia mas larga que casa con alguna de las expresiones regulares (`regular_expressionK`) y ejecuta la correspondiente acción (`actionK()`). Si no se encuentra ningun emparejamiento se ejecuta la *regla por defecto*, que es:

```
(.|\n)  { printf("%s",yytext); }
```

Si encuentran dos expresiones regulares con las que la cadena mas larga casa, elige la que figura primera en el programa `lex`. Una vez que `yylex()` ha encontrado el *token*, esto es, el patrón que casa con la cadena mas larga, dicha cadena queda disponible a través del puntero global `yytext` , y su longitud queda en la variable entera global `yyleng` .

Una vez que se ha ejecutado la correspondiente acción, `yylex()` continúa con el resto de la entrada, buscando por subsiguientes emparejamientos. Asi continúa hasta encontrar un `end of file`, en cuyo caso termina, retornando un cero o bien hasta que una de las acciones explicitamente ejecuta una sentencia `return`.

Sección de definiciones

La primera sección contiene, si las hubiera, las definiciones regulares y las declaraciones de los estados de arranque.

Las definiciones tiene la forma:

name regular_definition

donde *name* puede ser descrito mediante la expresión regular:

`[a-zA-Z_][a-zA-Z_0-9-]*`

La *regular_definition* comienza en el primer carácter no blanco que sigue a *name* y termina al final de la línea. La definición es una expresión regular extendida. Las subsiguientes definiciones pueden "llamar" a la macro `{name}` escribiéndola entre llaves. La macro se expande entonces a `(regular_definition)` en `flex` y a `regular_definition` en `lex`.

El código entre los delimitadores `%{` y `%}` se copia verbatim al fichero de salida, situándose en la parte de declaraciones globales. Los delimitadores deben aparecer (sólos) al comienzo de la línea.

El Lenguaje de las Expresiones Regulares Flex

La sintáxis que puede utilizarse para la descripción de las expresiones regulares es la que se conoce como "extendida":

- `x` Casa con 'x'
- `.` Cualquier carácter, excepto `\n`.
- `[xyz]` Una "clase"; en este caso una de las letras `x`, `y`, `z`
- `[abj-oZ]` Una "clase" con un rango; casa con `a`, `b`, cualquier letra desde `j` hasta `o`, o una `Z`
- `[^A-Z]` Una "Clase complementada" esto es, todos los caracteres que no están en la clase. Cualquier carácter, excepto las letras mayúsculas. Obsérvese que el retorno de carro `\n` casa con esta expresion. Así es posible que, ¡un patrón como `[^"]+` pueda casar con todo el fichero!.
- `[^A-Z\n]` Cualquier carácter, excepto las letras mayúsculas o un `\n`.
- `[[:alnum:]]` Casa con cualquier caracter alfanumérico. Aqui `[:alnum:]` se refiere a una de las clases predefinidas. Las otras clases son: `[:alpha:]` `[:blank:]` `[:cntrl:]` `[:digit:]` `[:graph:]` `[:lower:]` `[:print:]` `[:punct:]` `[:space:]` `[:upper:]` `[:xdigit:]`. Estas clases designan el mismo conjunto de caracteres que la correspondiente función C `isXXXX`.
- `r*` Cero o mas `r`.
- `r+` Una o mas `r`.
- `r?` Cero o una `r`.
- `r{2,5}` Entre 2 y 5 `r`.
- `r{2,}` 2 o mas `r`. `r{4}` Exactamente 4 `r`.
- `{macro_name}` La expansión de `macro_name` por su *regular_definition*
- `"[xyz]\"foo"` Exactamente la cadena: [xyz]"foo
- `\X` Si `X` is an `a`, `b`, `f`, `n`, `r`, `t`, o `v`, entonces, la interpretación ANSI-C de `\x`. En cualquier otro caso `X`.
- `\0` El carácter NUL (ASCII 0).
- `\123` El carácter cuyo código octal es 123.
- `\x2a` El carácter cuyo código hexadecimal es 2a.

- `(r)` Los paréntesis son utilizados para cambiar la precedencia.
- `rs` Concatenation
- `r|s` Casa con `r` o `s`
- `r/s` Una `r` pero sólo si va seguida de una `s`. El texto casado con `s` se incluye a la hora de decidir cual es el emparejamiento mas largo, pero se devuelve a la entrada cuando se ejecuta la acción. La acción sólo ve el texto asociado con `r`. Este tipo de patrón se denomina *trailing context* o *lookahead* positivo.
- `^r` Casa con `r`, al comienzo de una línea. Un `^` que no aparece al comienzo de la línea o un `$` que no aparece al final de la línea, pierde su naturaleza de "ancla" y es tratado como un carácter ordinario. Asi: `foo|(bar$)` se empareja con `bar$`. Si lo que se quería es la otra interpretación, es posible escribir `foo|(bar\n)`, o bien:

  ```
  foo |
  bar$   { /* action */ }
  ```

- `r$` Casa con `r`, al final de una línea. Este es también un operador de *trailing context*. Una regla no puede tener mas de un operador de *trailing context*. Por ejemplo, la expresión `foo/bar$` es incorrecta.
- `<s>r` Casa con `r`, pero sólo si se está en el estado `s`.
- `<s1,s2,s3>r` Idem, si se esta en alguno de los estados `s1`, `s2`, or `s3`
- `<*>r` Casa con `r` cualquiera que sea el estado, incluso si este es exclusivo.
- `<<EOF>>` Un final de fichero.
- `<s1,s2><<EOF>>` Un final de fichero, si los estados son `s1` o `s2`

Los operadores han sido listados en orden de precedencia, de la mas alta a la mas baja. Por ejemplo `foo|bar+` es lo mismo que `(foo)|(ba(r)+)`.

Las Acciones Semánticas

Cada patrón regular tiene su correspondiente acción asociada. El patrón termina en el primer espacio en blanco (sin contar aquellos que están entre comillas dobles o prefijados de secuencias de escape). Si la acción comienza con `{`, entonces se puede extender a través de multiples líneas, hasta la correspondiente `}`. El programa `flex` no hace un análisis del código C dentro de la acción. Existen tres directivas que pueden insertarse dentro de las acciones: `BEGIN`, `ECHO` y `REJECT`. Su uso se muestra en los subsiguientes ejemplos.

La sección de código se copia verbatim en `lex.yy.c`. Es utilizada para proveer las funciones de apoyo que se requieran para la descripción de las acciones asociadas con los patrones que parecen en la sección de reglas.

29.2. Versión Utilizada

Todos los ejemplos que aparecen en este documento fueron preparados con la versión 2.5.4 de *flex* en un entorno Linux

```
$ uname -a
Linux nereida.deioc.ull.es 2.2.12-20 #10 Mon May 8 19:40:16 WEST 2000 i686 unknown
$ flex --version
flex version 2.5.4
```

y con la versión 2.5.2 en un entorno Solaris

```
> uname -a
SunOS fonil 5.7 Generic_106541-04 sun4u sparc SUNW,Ultra-5_10
> flex --version
flex version 2.5.2
```

29.3. Espacios en blanco dentro de la expresión regular

La expresión regular va desde el comienzo de la línea hasta el primer espacio en blanco no escapado. Todos los espacios en blanco que formen parte de la expresión regular deben ser escapados o protegidos entre comillas. Así, el siguiente programa produce un error en tiempo de compilación C:

```
> cat spaces.l
%%
one two { printf("spaces\n"; }
%%
nereida:~/public_html/regexpr/lex/src> flex spaces.l
nereida:~/public_html/regexpr/lex/src> gcc lex.yy.c
spaces.l: In function ‘yylex’:
spaces.l:2: ‘two’ undeclared (first use in this function)
spaces.l:2: (Each undeclared identifier is reported only once
spaces.l:2: for each function it appears in.)
spaces.l:2: parse error before ‘{’
spaces.l:4: case label not within a switch statement
lex.yy.c:632: case label not within a switch statement
lex.yy.c:635: case label not within a switch statement
lex.yy.c:757: default label not within a switch statement
lex.yy.c: At top level:
lex.yy.c:762: parse error before ‘}’
```

Deberíamos escapar el blanco entre `one` y `two` o bien proteger la cadena poniéndola entre comillas: `"one two"`.

29.4. Ejemplo Simple

Este primer ejemplo sustituye las apariciones de la palabra *username* por el *login* del usuario:

```
$ cat subst.l
%option main
%{
#include <unistd.h>
%}
%%
username    printf( "%s",  getlogin());
%%
$ flex -osubst.c subst.l
$ gcc -o subst subst.c
$ subst
Dear username:
Dear pl:
```

He presionado CTRL-D para finalizar la entrada.
Observe el uso de la opción `%option main` en el fichero *subst.l* para hacer que *flex* genere una función *main*. También merece especial atención el uso de la opción `-osubst` para cambiar el nombre del fichero de salida, que por defecto será *lex.yy.c*.

29.5. Suprimir

Al igual que en *sed* y *awk*, es muy sencillo suprimir las apariciones de una expresión regular.

```
$ cat delete.l
/* delete all entries of zap me */
%%
"zap me"
$ flex delete.l ; gcc lex.yy.c -lfl; a.out
this is zap me a first zap me phrase
this is  a first  phrase
```

29.6. Declaración de yytext

En la sección de definiciones es posible utilizar las directivas `%pointer` o `%array`. Estas directivas hacen que `yytext` se declare como un puntero o un *array* respectivamente. La opción por defecto es declararlo como un puntero, salvo que se haya usado la opción `-l` en la línea de comandos, para garantizar una mayor compatibilidad con `LEX`. Sin embargo, y aunque la opción `%pointer` es la mas eficiente (el análisis es mas rápido y se evitan los *buffer overflow*), limita la posible manipulación de *yytext* y de las llamadas a `unput()`.

```
$ cat yytextp.l
%%
hello {
        strcat(yytext, " world");
        printf("\n%d: %s\n",strlen(yytext),yytext);
      }
$ flex yytextp.l ; gcc lex.yy.c -lfl ; a.out
hello

11: hello world

fatal flex scanner internal error--end of buffer missed
```

Este error no aparece si se utiliza la opción `%array`:

```
$ cat yytext.l
%array
%%
hello {
        strcat(yytext, " world");
        printf("\n%d: %s\n",strlen(yytext),yytext);
      }
$ flex yytext.l; gcc lex.yy.c -lfl; a.out
hello

11: hello world
```

Además, algunos programs `LEX` modifican directamente `yytext`, utilizando la declaración:
`extern char yytext[]`
que es incompatible con la directiva `%pointer` (pero correcta con `%array`). La directiva `%array` define `yytext` como un *array* de tamaño `YYLMAX`. Si deseamos trabajar con un mayor tamaño, basta con redefinir `YYLMAX`.

29.7. Declaración de yylex()

Por defecto la función *yylex()* que realiza el análisis léxico es declarada como `int yylex()`. Es posible cambiar la declaración por defecto utilizando la macro *YY_DECL*. En el siguiente ejemplo la definición:

```
#define YY_DECL char *scanner(int *numcount, int *idcount)
```

hace que la rutina de análisis léxico pase a llamarse *scanner* y tenga dos parametros de entrada, retornando un valor de tipo `char *`.

```
$ cat decl.l
%{
#define YY_DECL char *scanner(int *numcount, int *idcount)
%}

num [0-9]+
id [a-z]+
%%
{num} {(*numcount)++;}
halt  {return ((char *) strdup(yytext));}
{id}  {(*idcount)++;}
%%
main() {
  int a,b;
  char *t;

  a = 0; b = 0;
  t = scanner(&a, &b);
  printf("numcount = %d, idcount = %d, yytext = %s\n",a,b,t);
  t = scanner(&a, &b);
  printf("numcount = %d, idcount = %d, yytext = %s\n",a,b,t);
}

int yywrap() {
  return 1;
}
```

La ejecución del programa anterior produce la siguiente salida:

```
$ decl
a b 1 2 3 halt
     numcount = 3, idcount = 2, yytext = halt

e 4 5 f

numcount = 5, idcount = 4, yytext = (null)
$ decl
a b 1 2 3 halt
     numcount = 3, idcount = 2, yytext = halt

e 4 f 5 halt
    numcount = 5, idcount = 4, yytext = halt
```

29.8. yywrap()

Cuando el analizador léxico alcanza el final del fichero, el comportamiento en las subsiguientes llamadas a *yylex* resulta indefinido. En el momento en que *yylex()* alcanza el final del fichero llama a la función *yywrap*, la cual retorna un valor de 0 o 1 según haya mas entrada o no. Si el valor es 0, la función *yylex* asume que la propia *yywrap* se ha encargado de abrir el nuevo fichero y asignarselo a `yyin` . Otra manera de continuar es haciendo uso de la función *yyrestart(FILE *file)*. El siguiente ejemplo cuenta el número de líneas, palabras y caracteres en una lista de ficheros proporcionados como entrada.

```
%{
unsigned long charCount = 0, wordCount = 0, lineCount = 0;
%}

word [^ \t\n]+
eol \n

%%
{word} { wordCount++; charCount += yyleng; }
{eol} { charCount++; lineCount++; }
. charCount++;

%%

char **fileList;
unsigned nFiles;
unsigned currentFile = 0;
unsigned long totalCC = 0;
unsigned long totalWC = 0;
unsigned long totalLC = 0;

main ( int argc, char **argv) {
  FILE *file;

  fileList = argv + 1; nFiles = argc - 1;

  if (nFiles == 0) {
    fprintf(stderr,"Usage is:\n%s file1 file2 file3 ...\n",argv[0]);
    exit(1);
  }
  file = fopen (fileList[0], "r");
  if (!file) {
      fprintf (stderr, "could not open %s\n", argv[1]);
      exit (1);
  }
  currentFile = 1; yyrestart(file);
  yylex ();
  printf ("%8lu %8lu %8lu %s\n", lineCount, wordCount,
    charCount, fileList[currentFile - 1]);
  if (argc > 2) {
      totalCC += charCount; totalWC += wordCount; totalLC += lineCount;
      printf ("%8lu %8lu %8lu total\n", totalLC, totalWC, totalCC);
  }
  return 0;
```

```
}

int yywrap () {
  FILE *file;

  if (currentFile < nFiles) {
     printf ("%8lu %8lu %8lu %s\n", lineCount, wordCount,
       charCount, fileList[currentFile - 1]);
     totalCC += charCount; totalWC += wordCount; totalLC += lineCount;
     charCount = wordCount = lineCount = 0;
     fclose (yyin);

     while (fileList[currentFile] != (char *) 0) {
       file = fopen (fileList[currentFile++], "r");
       if (file != NULL) { yyrestart(file); break; }
  fprintf (stderr, "could not open %s\n", fileList[currentFile - 1]);
     }
     return (file ? 0 : 1);
    }
    return 1;
}
```

La figura muestra el proceso de compilación y la ejecución:

```
$ flex countlwc.l;gcc lex.yy.c; a.out *.l
      58      179     1067 ape-05.l
      88      249     1759 countlwc.l
      11       21      126 magic.l
       9       17      139 mgrep.l
       9       16      135 mlg.l
       5       15      181 ml.l
       7       12       87 subst.l
     187      509     3494 total
```

La diferencia esencial entre asignar *yyin* o llamar a la función *yyrestart* es que esta última puede ser utilizada para conmutar entre ficheros en medio de un análisis léxico. El funcionamiento del programa anterior no se modifica si se se intercambian asignaciones a *yyin* (*yyin = file*) y llamadas a *yyrestart(file)*.

29.9. unput()

La función *unput(c)* coloca el carácter *c* en el flujo de entrada, de manera que será el primer carácter leído en próxima ocasión.

```
$ cat unput2.l
%array
%%
[a-z] {unput(toupper(yytext[0]));}
[A-Z] ECHO;
%%
$ flex unput2.l ; gcc lex.yy.c -lfl;a.out
abcd
ABCD
```

Un problema importante con *unput* es que, cuando se utiliza la opción `%pointer`, las llamadas a *unput* destruyen los contenidos de *yytext*. Es por eso que, en el siguiente ejemplo se hace una copia de *yytext*. La otra alternativa es, por supuesto, usar la opción `%array`.

```
$ cat unput.l
%%
[0-9]+ {
  int i;
  char *yycopy = (char *) strdup(yytext);

  unput(')');
  for(i=strlen(yycopy)-1; i>=0; --i)
    unput(yycopy[i]);
  unput('(');
  free(yycopy);
}
\([0-9]+\) printf("Num inside parenthesis: %s\n",yytext);
.|\n
$ flex unput.l ; gcc lex.yy.c -lfl ; a.out
32
Num inside parenthesis: (32)
(43)
Num inside parenthesis: (43)
```

29.10. input()

La función *input()* lee desde el flujo de entrada el siguiente carácter. Normalmente la utilizaremos si queremos tomar "personalmente el control" del análisis. El ejemplo permite "engullir" los comentarios (no anidados):

```
$ cat input.l
%%
"/*" {
        int c;
        for(;;) {
          while ((c=input()) != '*' && c != EOF)
            ;
          if (c == '*')  {
            while ((c = input()) == '*')
              ;
            if (c == '/') break;
          }
          if (c == EOF) {
            fprintf(stderr,"Error: EOF in comment");
            yyterminate();
          }
        }
      }
```

La función `yyterminate()` termina la rutina de análisis léxico y devuelve un cero indicándole a la rutina que llama que todo se ha acabado. Por defecto, `yyterminate()` es llamada cuando se encuentra un final de fichero. Es una macro y puede ser redefinida.

```
$ flex input.l ; gcc lex.yy.c -lfl ; a.out
hello /* world */
```

```
hello
unfinished /* comment
unfinished Error: EOF in comment
```

He presionado CTRL-D después de entrar la palabra *comment*.

29.11. REJECT

La directiva `REJECT` le indica al analizador que proceda con la siguiente regla que casa con un prefijo de la entrada. Como es habitual en *flex*, se elige la siguiente regla que casa con la cadena mas larga. Consideremos el siguiente ejemplo:

```
$ cat reject.l
%%
a    |
ab   |
abc  |
abcd ECHO; REJECT; printf("Never seen\n");
.|\n
```

La salida es:

```
$ gcc lex.yy.c -lfl;a.out
abcd
abcdabcaba
```

Observe que `REJECT` supone un cambio en el flujo de control: El código que figura después de `REJECT` no es ejecutado.

29.12. yymore()

La función `yymore()` hace que, en vez de vaciar *yytext* para el siguiente *matching*, el valor actual se mantenga, concatenando el valor actual de *yytext* con el siguiente:

```
$ cat yymore.l
%%
mega- ECHO; yymore();
kludge ECHO;

$ flex yymore.l ; gcc lex.yy.c -lfl ; a.out
mega-kludge
mega-mega-kludge
```

La variable `yyleng` no debería ser modificada si se hace uso de la función `yymore()`.

29.13. yyless()

La función `yyless(n)` permite retrasar el puntero de lectura de manera que apunta al carácter n de *yytext*. Veamos un ejemplo:

```
$ cat yyless.l
%%
foobar ECHO; yyless(4);
[a-z]+ ECHO;
```

```
$ flex yyless.l; gcc lex.yy.c -lfl; a.out
foobar
foobarar
```

Veamos un ejemplo mas "real". supongamos que tenemos que reconocer las cadenas entre comillas dobles, pero que pueden aparecer en las mismas secuencias de escape `\"`. La estrategia general del algoritmo es utilizar la expresión regular `\"[^"]+\"` y examinar si los dos últimos carácteres en `yytext` son `\"`. En tal caso, se concatena la cadena actual (sin la `"` final) como prefijo para el próximo emparejamiento (utilizando `yymore`). La eliminación de la `"` se hace a través de la ejecución de `yyless(yyleng-1)`, que al mismo tiempo garantiza que el próximo emparejamiento tendrá lugar con este mismo patrón `\"[^"]+\"`.

```
$ cat quotes.l
%%
\"[^"]+\" {
            printf("Processing string. %d: %s\n",yyleng,yytext);
            if (yytext[yyleng-2] =='\\') {
             yyless(yyleng-1); /* so that it will match next time */
             yymore(); /* concatenate  with current yytext */
             printf("After yyless. %d: %s\n",yyleng,yytext);
            } else {
              printf("Finished. The string is: %s\n",yytext);
            }
          }
```

El ejemplo no puede entenderse si no se tiene en cuenta que `yyless(yyleng-1)` actualiza los valores de *yyleng* y *yytext*, como muestra la salida.
¿Qué ocurre si intercambiamos las posiciones de `yymore()` e `yyless(yyleng-1)` en el código? ¿Cambiara la salida? La respuesta es que no. Parece que la concatenación se hace con el valor final de *yytext* y no con el valor que este tenía en el momento de la llamada a *yymore*.
Otra observación a tener en cuenta es que *yyless()* es una macro y que, por tanto, sólo puede ser utilizada dentro del fichero *lex* y no en otros ficheros fuentes.

En general, el uso de estas funciones nos puede resolver el problema de reconocer límites que de otra forma serían difíciles de expresar con una expresión regular.

```
$ flex quotes.l ; gcc lex.yy.c -lfl ; a.out
"Hello \"Peter\", nice to meet you"
Procesing string. 9: "Hello \"
After yyless. 8: "Hello \
Procesing string. 16: "Hello \"Peter\"
After yyless. 15: "Hello \"Peter\
Procesing string. 35: "Hello \"Peter\", nice to meet you"
Finished. The string is: "Hello \"Peter\", nice to meet you"
```

29.14. Estados

Las expresiones regulares pueden ser prefijadas mediante *estados*. Los estados o condiciones de arranque, se denotan mediante un identificador entre ángulos y se declaran en la parte de las definiciones. Las declaraciones se hacen mediante `%s` para los estados "inclusivos" o bien `%x` para los estados "exclusivos", seguidos de los nombres de los estados. No pueden haber caracteres en blanco antes de la declaración. Un *estado* se activa mediante la acción BEGIN *estado*. A partir de ese momento, las reglas que esten prefijadas con el estado pasan a estar activas. En el caso de que el estado sea inclusivo, las reglas no prefijadas también permanecen activas. Los estados exclusivos son especialmente útiles para especificar "sub analizadores" que analizan porciones de la entrada cuya estructura "sintáctica" es diferente de la del resto de la entrada.

El ejemplo "absorbe" los comentarios, conservando el numero de líneas del fichero en la variable `linenum`

```
$ cat comments.l
%option noyywrap
%{
  int linenum = 0;
%}
%x comment
%%

"/*" BEGIN(comment); printf("comment=%d, YY_START = %d, YYSTATE = %d",comment,YY_START,YYSTATE
<comment>[^*\n]* /* eat anything that is not a star * /
<comment>"*"+[^*/\n]* /* eat up starts not followed by / */
<comment>\n ++linenum; /* update number of lines */
<comment>"*"+"/" BEGIN(0);

\n ECHO; linenum++;
.  ECHO;
%%
main() {
  yylex();
  printf("\n%d lines\n",linenum);
}
```

La opción `noyywrap` hace que `yylex()` no llame a la función `yywrap()` al final del fichero y que asuma que no hay mas entrada por procesar.

Los estados se traducen por enteros, pudiendo ser manipulados como tales. La macro `INITIAL` puede utilizarse para referirse al estado 0. Las macros `YY_START` y `YYSTATE` contienen el valor del estado actual.

```
$ flex comments.l ; gcc lex.yy.c ; a.out < hello.c
main() <%
int a<:1:>; comment=1, YY_START = 1, YYSTATE = 1
  a<:0:> = 4; comment=1, YY_START = 1, YYSTATE = 1
  printf("hello world! a(0) is %d\n",a<:0:>);
%>

6 lines
$ cat hello.c
main() <%
int a<:1:>; /* a comment */
  a<:0:> = 4; /* a comment in
                 two lines */
  printf("hello world! a(0) is %d\n",a<:0:>);
%>
```

En *flex* es posible asociar un ámbito con los estados o condiciones iniciales. Basta con colocar entre llaves las parejas *patrón acción* gobernadas por ese estado. El siguiente ejemplo procesa las cadenas C:

```
$ cat ststring.l
%option main
%x str
%{
```

```
#define MAX_STR_CONST 256

  char string_buffer[MAX_STR_CONST];
  char *string_buf_ptr;
%}

%%
\"  string_buf_ptr  = string_buffer; BEGIN(str);
<str>{
\"              {BEGIN (INITIAL); *string_buf_ptr = '\0'; printf("%s",string_buffer); }
\n              { printf("Error: non terminated string\n"); exit(1); }
\\[0-7]{1,3}    { int result; /* octal escape sequence */
                         (void) sscanf(yytext+1,"%o",&result);
                          if (result > 0xff) {printf("Error: constant out of bounds\n"); exit(2
                          *string_buf_ptr++ = result;
                     }
\\[0-9]+        { printf("Error: bad escape sequence\n"); exit(2); }
\\n             {*string_buf_ptr++ = '\n';}
\\t             {*string_buf_ptr++ = '\t';}
\\b             {*string_buf_ptr++ = '\b';}
\\r             {*string_buf_ptr++ = '\r';}
\\f             {*string_buf_ptr++ = '\f';}
\\(.|\n)        {*string_buf_ptr++ = yytext[1];}
[^\\\n\"]+      {char *yptr = yytext; while(*yptr) *string_buf_ptr++ = *yptr++; }
}
(.|\n)
%%
$ flex ststring.l ; gcc lex.yy.c ; a.out < hello.c
        hello
world! a(0) is %d
$ cat hello.c
main() <%
int a<:1:>; /* a comment */
  a<:0:> = 4; /* a comment in
                 two lines */
  printf("\thell\157\nworld! a(0) is %d\n",a<:0:>);
%>
```

Obsérve la conducta del programa ante las siguientes entradas:

- Entrada:

  ```
  "hello \
  dolly"
  ```

  ¿Cuál será la salida? ¿Que patrón del programa anterior es el que casa aqui?

- Entrada: `"hello\ndolly"`. ¿Cuál será la salida? ¿Que patrón del programa anterior es el que casa aqui?

- `|`
  `"hello`

  Donde hay un retorno del carro después de `hello`. ¿Cuál será la salida?

29.15. La pila de estados

Mediante el uso de la opción
`%option stack`
tendremos acceso a una pila de estados y a tres rutinas para manipularla:

- `void yy_push_state(int new_state)`
  Empuja el estado actual y bifurca a `new_state`.

- `void yy_pop_state()`
  Saca el estado en el *top* de la pila y bifurca a el mismo.

- `int yy_top_state()`
  Nos devuelve el estado en el *top* de la pila, sin alterar los contenidos de la misma.

29.15.1. Ejemplo

El siguiente programa `flex` utiliza las funciones de la pila de estados para reconocer el lenguaje (no regular) $\{a^n b^n \ / \ n \in N\}$

```
%option main
%option noyywrap
%option stack
%{
#include <stdio.h>
#include  <stdlib.h>
%}
%x estado_a
%%
^a { yy_push_state(estado_a);}
<estado_a>{
a      { yy_push_state(estado_a); }
b      { yy_pop_state();  }
b[^b\n]+    {   printf ("Error\n");
        while (YYSTATE != INITIAL)
          yy_pop_state();
        while (input() != '\n') ;
      }
(.|\n)  {  printf ("Error\n");
        while (YYSTATE != INITIAL)
          yy_pop_state();
        while (input() != '\n') ;
      }
}
.      {  printf ("Error\n");
        while (input() != '\n') ;
      }
\n      { printf("Aceptar\n");}
%%
```

29.16. Final de Fichero

El patrón `<<EOF>>` permite asociar acciones que se deban ejecutar cuando se ha encontrado un *end of file* y la macro `yywrap()` ha devuelto un valor no nulo.

Cualquiera que sea, la acción asociada deberá de optar por una de estas cuatro alternativas:

- Asignar `yyin` a un nuevo fichero de entrada
- Ejecutar `return`
- Ejecutar `yyterminate()` (véase la sección 29.10)
- Cambiar de *buffer* de entrada utilizando la función `yy_switch_buffer` (véase la sección 29.21).

El patrón `<<EOF>>` no puede usarse con otras expresiones regulares. Sin embargo, es correcto prefijarlo con estados.

Si `<<EOF>>` aparece sin condiciones de arranque, la regla se aplica a todos los estados que no tienen una regla `<<EOF>>` específica. Si lo que se quiere es que la regla se restringa al ámbito del estado inicial se deberá escribir:

`<INITIAL><<EOF>>`

Sigue un programa que reconoce los comentarios anidados en C. Para detectar comentarios incacabados usaremos `<<EOF>>`.

```
%option stack
%x comment
%%
"/*"   { yy_push_state(comment); }
(.|\n) ECHO;
<comment>"/*"     { yy_push_state(comment); }
<comment>"*/"     { yy_pop_state(); }
<comment>(.|\n)   ;
<comment><<EOF>>  { fprintf(stderr,"Error\n"); exit(1); }
%%

$ cat hello.c
main() {
int a[1]; /* a /* nested comment */. */
  a[0] = 4; /* a  /* nested comment in
                 /* two */ lines  */ *****/
}
$ flex nestedcom.l ; gcc lex.yy.c -lfl ; a.out < hello.c
main() {
int a[1];
  a[0] = 4;
}
$ cat hello4.c
main() {
int a[1]; /* a /* nested comment */. */
  a[0] = 4; /* an  /* incorrectly nested comment in
                 /* two  lines  */ *****/
}
$ a.out < hello4.c
main() {
int a[1];
Error
  a[0] = 4;
```

29.17. Uso de Dos Analizadores

La opción `-Pprefix` de flex cambia el prefijo por defecto `yy` para todas las variables globales y funciones. Por ejemplo `-Pfoo` cambia el nombre de `yytext` `footext`. También cambia el nombre del fichero de salida de `lex.yy.c` a `lex.foo.c`. Sigue la lista de identificadores afectados:

```
yy_create_buffer
yy_delete_buffer
yy_flex_debug
yy_init_buffer
yy_flush_buffer
yy_load_buffer_state
yy_switch_to_buffer
yyin
yyleng
yylex
yylineno
yyout
yyrestart
yytext
yywrap
```

Desde dentro del analizador léxico puedes referirte a las variables globales y funciones por cualquiera de los nombres, pero externamente tienen el nombre cambiado. Esta opción nos permite enlazar diferentes programas flex en un mismo ejecutable.

Sigue un ejemplo de uso de dos analizadores léxicos dentro del mismo programa:

```
$ cat one.l
%%
one {printf("1\n"); return 1;}
.   {printf("First analyzer: %s\n",yytext);}
%%

int onewrap(void) {
  return 1;
}

$ cat two.l
%%
two {printf("2\n"); return 2;}
.   {printf("Second analyzer: %s\n",yytext);}
%%
int twowrap(void) {
  return 1;
}
$ cat onetwo.c
main() {
  onelex();
  twolex();
}
```

Como hemos mencionado, la compilación *flex* se debe realizar con el opción `-P`, que cambia el prefijo por defecto `yy` de las funciones y variables accesibles por el usuario. El mismo efecto puede conseguirse utilizando la opción `prefix`, escribiendo `%option prefix="one"` y `%option prefix="two"` en los respectivos programas `one.l` y `two.l`.

```
$ flex -Pone one.l
$ flex -Ptwo two.l
$ ls -ltr | tail -2
-rw-rw----   1 pl       casiano     36537 Nov  7 09:52 lex.one.c
-rw-rw----   1 pl       casiano     36524 Nov  7 09:52 lex.two.c
$ gcc onetwo.c lex.one.c lex.two.c
$ a.out
two
First analyzer: t
First analyzer: w
First analyzer: o

one
1
one
Second analyzer: o
Second analyzer: n
Second analyzer: e

two
2
$
```

29.18. La Opción `outfile`

Es posible utilizar la opción `-ooutput.c` para escribir el analizador léxico en el fichero `output.c` en vez de en `lex.yy.c`. El mismo efecto puede obtenerse usando la opción `outfile="output.c"` dentro del programa `lex`.

29.19. Leer desde una Cadena: YY_INPUT

En general, la rutina que hace el análisis léxico, *yylex()*, lee su entrada a través de la macro `YY_INPUT`. Esta macro es llamada con tres parámetros

```
YY_INPUT(buf,result,max)
```

el primero, *buf* es utilizado para guardar la entrada. el tercero *max* indica el número de caracteres que *yylex()* pretende leer de la entrada. El segundo *result* contendrá el número de caracteres realmente leídos. Para poder leer desde una cadena (*string*) basta con modificar `YY_INPUT` para que copie los datos de la cadena en el *buffer* pasado como parámetro a `YY_INPUT`. Sigue un ejemplo:

```
$ cat string.l
%{
#undef YY_INPUT
#define YY_INPUT(b,r,m) (r = yystringinput(b,m))
#define min(a,b) ((a<b)?(a):(b))
%}

%%
[0-9]+ printf("Num-");
[a-zA-Z][a-zA-Z_0-9]* printf("Id-");
[ \t]+
. printf("%c-",yytext[0]);
%%
```

```
extern char string[];
extern char *yyinputptr;
extern char *yyinputlim;

int yystringinput(char *buf, int maxsize) {
  int n = min(maxsize, yyinputlim-yyinputptr);

  if (n > 0) {
    memcpy(buf, yyinputptr, n);
    yyinputptr += n;
  }
  return n;
}

int yywrap() { return 1; }
```

Este es el fichero conteniendo la función *main*:

```
$ cat stringmain.c
char string[] = "one=1;two=2";
char *yyinputptr;
char *yyinputlim;

main() {
  yyinputptr = string;
  yyinputlim = string + strlen(string);
  yylex();
  printf("\n");
}
```

Y esta es la salida:

```
$ a.out
Id-=-Num-;-Id-=-Num-
```

La cadena `string = "one=1;two=2"` definida en la línea 2 ha sido utilizada como entrada para el análisis léxico.

29.20. El operador de "trailing context" o "lookahead" positivo

En el lenguaje FORTRAN original los "blancos" no eran significativos y no se distinguía entre mayúsculas y minúsculas. Así pues la cadena `do i = 1, 10` es equivalente a la cadena `DOI=1,10`. Un conocido conflicto ocurre entre una cadena con la estructura `do i = 1.10` (esto es `DOI=1.10`) y la cadena anterior. En la primera `DO` e `I` son dos "tokens" diferentes, el primero correspondiendo a la palabra reservada que indica un bucle. En la segunda, `DOI` constituye un único "token" y la sentencia se refiere a una asignación. El conflicto puede resolverse utilizando el operador de "trailing" `r/s`. Como se mencionó, el operador de "trailing"`r/s` permite reconocer una `r` pero sólo si va seguida de una `s`. El texto casado con `s` se incluye a la hora de decidir cual es el emparejamiento mas largo, pero se devuelve a la entrada cuando se ejecuta la acción. La acción sólo ve el texto asociado con `r`. El fichero `fortran4.l` ilustra una posible solución:

```
 cat fortran4.l
%array
%{
#include <string.h>
```

```
#undef YY_INPUT
#define YY_INPUT(buf,result,max) (result = my_input(buf,max))
%}
number [0-9]+
integer [+-]?{number}
float ({integer}\.{number}?|\.{number})(E{integer})?
label [A-Z0-9]+
id   [A-Z]{label}*
%%
DO/{label}={number}\, { printf("do loop\n"); }
{id} { printf("Identifier %s\n",yytext); }
{number} { printf("Num %d\n",atoi(yytext)); }
{float} { printf("Float %f\n",atof(yytext)); }
(.|\n)
%%

int my_input(char *buf, int max)
{
  char *q1, *q2, *p = (char *) malloc(max);
  int i;
  if ('\0' != fgets(p,max,yyin)) {
    for(i=0, q1=buf, q2=p;(*q2 != '\0');q2++) {
      if (*q2 != ' ') { *q1++ = toupper(*q2); i++; };
    }
    free(p);
    return i;
  }
  else exit(1);
}
```

La función

`char *fgets(char *s, int size, FILE *stream)`

lee a lo mas uno menos que `size` caracteres desde `stream` y los almacena en el *buffer* apuntado por `s`. La lectura termina después de un `EOF` o un retorno de carro. Si se lee un `\n`, se almacena en el *buffer*. La función pone un carácter nulo `\0` como último carácter en el *buffer*.

A continuación, puedes ver los detalles de una ejecución:

```
$ flex fortran4.l; gcc lex.yy.c -lfl; a.out
do j = 1 . 10
Identifier DOJ
Float 1.100000
do k = 1, 5
do loop
Identifier K
Num 1
Num 5
```

29.21. Manejo de directivas `include`

El analisis léxico de algunos lenguajes requiere que, durante la ejecución, se realice la lectura desde diferentes ficheros de entrada. El ejemplo típico es el manejo de las directivas *include file* existentes en la mayoría de los lenguajes de programación.

¿Donde está el problema? La dificultad reside en que los analizadores generados por *flex* proveen almacenamiento intermedio (*buffers*) para aumentar el rendimiento. No basta con reescribir nuestro

propio *YY_INPUT* de manera que tenga en cuenta con que fichero se esta trabajando. El analizador sólo llama a *YY_INPUT* cuando alcanza el final de su *buffer*, lo cual puede ocurrir bastante después de haber encontrado la sentencia *include* que requiere el cambio de fichero de entrada.

```
$ cat include.l
%x incl
%{
#define yywrap() 1
#define MAX_INCLUDE_DEPTH 10
YY_BUFFER_STATE include_stack[MAX_INCLUDE_DEPTH];
int include_stack_ptr = 0;
%}
%%
include          BEGIN(incl);
.                ECHO;
<incl>[ \t]*
<incl>[^ \t\n]+  { /* got the file name */
                   if (include_stack_ptr >= MAX_INCLUDE_DEPTH) {
                     fprintf(stderr,"Includes nested too deeply\n");
                     exit(1);
                   }
                   include_stack[include_stack_ptr++] = YY_CURRENT_BUFFER;
                   yyin = fopen(yytext,"r");
                   if (!yyin) {
                     fprintf(stderr,"File %s not found\n",yytext);
                     exit(1);
                   }
                   yy_switch_to_buffer(yy_create_buffer(yyin, YY_BUF_SIZE));
                   BEGIN(INITIAL);
                 }
<<EOF>> {
          if ( --include_stack_ptr < 0) {
            yyterminate();
          } else {
            yy_delete_buffer(YY_CURRENT_BUFFER);
            yy_switch_to_buffer(include_stack[include_stack_ptr]);
          }
        }
%%
main(int argc, char ** argv) {

  yyin = fopen(argv[1],"r");
  yylex();
}
```

La función `yy_create_buffer(yyin, YY_BUF_SIZE));` crea un *buffer* lo suficientemente grande para mantener `YY_BUF_SIZE` caracteres. Devuelve un `YY_BUFFER_STATE`, que puede ser pasado a otras rutinas. `YY_BUFFER_STATE` es un puntero a una estructura de datos opaca (`struct yy_buffer_state`) que contiene la información para la manipulación del *buffer*. Es posible por tanto inicializar un puntero `YY_BUFFER_STATE` usando la expresión `((YY_BUFFER_STATE) 0)`.

La función `yy_switch_to_buffer(YY_BUFFER_STATE new_buffer);` conmuta la entrada del analizador léxico. La función `void yy_delete_buffer( YY_BUFFER_STATE buffer )` se usa para recuperar la memoria consumida por un *buffer*. También se pueden limpiar los contenidos actuales de un buffer llamando a: `void yy_flush_buffer( YY_BUFFER_STATE buffer )`

La regla especial `<<EOF>>` indica la acción a ejecutar cuando se ha encontrado un final de fichero e `yywrap()` retorna un valor distinto de cero. Cualquiera que sea la acción asociada, esta debe terminar con uno de estos cuatro supuestos:

1. Asignar `yyin` a un nuevo fichero de entrada.
2. Ejecutar `return`.
3. Ejecutar `yyterminate()`.
4. Cambiar a un nuevo buffer usando `yy_switch_to_buffer()`.

   La regla `<<EOF>>` no se puede mezclar con otros patrones.

Este es el resultado de una ejecución del programa:

```
$ cat hello.c
#include hello2.c
main() <%
int a<:1:>; /* a comment */
  a<:0:> = 4; /* a comment in
                 two lines */
  printf("\thell\157\nworld! a(0) is %d\n",a<:0:>);
%>
$ cat hello2.c
#include hello3.c
/* file hello2.c  */
$ cat hello3.c
/*
third file
*/
$ flex include.l ; gcc lex.yy.c ; a.out hello.c
##/*
third file
*/

/* file hello2.c  */

main() <%
int a<:1:>; /* a comment */
  a<:0:> = 4; /* a comment in
                 two lines */
  printf("\thell\157\nworld! a(0) is %d\n",a<:0:>);
%>
```

Una alternativa a usar el patrón `<<EOF>>` es dejar la responsabilidad de recuperar el *buffer* anterior a `yywrap()`. En tal caso suprimiríamos esta parajea patrón-acción y reescribiríamos `yywrap()`:

```
%x incl
%{
#define MAX_INCLUDE_DEPTH 10
YY_BUFFER_STATE include_stack[MAX_INCLUDE_DEPTH];
int include_stack_ptr = 0;
%}
%%
include          BEGIN(incl);
.                ECHO;
```

```
<incl>[ \t]*
<incl>[^ \t\n]+  { /* got the file name */
                  if (include_stack_ptr >= MAX_INCLUDE_DEPTH) {
     fprintf(stderr,"Includes nested too deeply\n");
     exit(1);
   }
   include_stack[include_stack_ptr++] = YY_CURRENT_BUFFER;
   yyin = fopen(yytext,"r");
   if (!yyin) {
     fprintf(stderr,"File %s not found\n",yytext);
     exit(1);
   }
   yy_switch_to_buffer(yy_create_buffer(yyin, YY_BUF_SIZE));
   BEGIN(INITIAL);
 }
%%
main(int argc, char ** argv) {

  yyin = fopen(argv[1],"r");
  yylex();
}

int yywrap() {
  if ( --include_stack_ptr < 0) {
    return 1;
  } else {
    yy_delete_buffer(YY_CURRENT_BUFFER);
    yy_switch_to_buffer(include_stack[include_stack_ptr]);
    return 0;
  }
}
```

29.22. Análisis Léxico desde una Cadena: yy_scan_string

El objetivo de este ejercicio es mostrar como realizar un análisis léxico de los argumentos pasados en la línea de comandos. Para ello *flex* provee la función `yy_scan_string(const char * str)`. Esta rutina crea un nuevo *buffer* de entrada y devuelve el correspondiente manejador `YY_BUFFER_STATE` asociado con la cadena `str`. Esta cadena debe estar terminada por un carácter `\0`. Podemos liberar la memoria asociada con dicho *buffer* utilizando `yy_delete_buffer(BUFFER)`. La siguiente llamada a `yylex()` realizará el análisis léxico de la cadena `str`.

```
$ cat scan_str.l
%%
[0-9]+     printf("num\n");
[a-zA-Z]+  printf("Id\n");
%%
main(int argc, char ** argv) {
int i;

  for(i=1;i<argc;i++) {
    yy_scan_string(argv[i]);
    yylex();
    yy_delete_buffer(YY_CURRENT_BUFFER);
```

```
  }
}

int yywrap() { return 1; }
$ flex scan_str.l ; gcc lex.yy.c ; a.out Hello World! 1234
Id
Id
!num
```

Alternativamente, la función `main()` podría haber sido escrita asi:

```
main(int argc, char ** argv) {
int i;
YY_BUFFER_STATE p;

  for(i=1;i<argc;i++) {
    p = yy_scan_string(argv[i]);
    yylex();
    yy_delete_buffer(p);
  }
}
```

La función `yy_scan_bytes(const char * bytes, int len)` hace lo mismo que `yy_scan_string` pero en vez de una cadena terminada en el carácter nulo, se usa la longitud `len`. Ambas funciones `yy_scan_string(const char * str)` y `yy_scan_bytes(const char * bytes, int len)` hacen una copia de la cadena pasada como argumento.

Estas dos funciones crean una copia de la cadena original. Es mejor que sea asi, ya que `yylex()` modifica los contenidos del *buffer* de trabajo. Si queremos evitar la copia, podemos usar

`yy_scan_buffer(char *base, yy_size_t size)`,

la cual trabaja directamente con el *buffer* que comienza en `base`, de tamaño `size` *bytes*, los últimos dos de los cuáles deben ser `YY_END_OF_BUFFER_CHAR` (ASCII NUL). Estos dos últimos *bytes* no son "escaneados". El área de rastreo va desde `base[0]` a `base[size-2]`, inclusive. Si nos olvidamos de hacerlo de este modo y no establecemos los dos *bytes* finales, la función `yy_scan_buffer()` devuelve un puntero nulo y no llega a crear el nuevo buffer de entrada. El tipo `yy_size_t` es un tipo entero. Como cabe esperar, `size` se refiere al tamaño del *buffer*.

29.23. Análisis de la Línea de Comandos y 2 Analizadores

El objetivo de este ejercicio es mostrar como realizar un análisis léxico de los argumentos pasados en la línea de comandos. Para ello diseñaremos una librería que proporcionará un función `yylexarg(argc,argv)` que hace el análisis de la línea de acuerdo con la especificación *flex* correspondiente. En el ejemplo, esta descripción del analizador léxico es proporcionada en el fichero *fl.l*. Para complicar un poco mas las cosas, supondremos que queremos hacer el análisis léxico de un fichero (especificado en la línea de comandos) según se especifica en un segundo analizador léxico *trivial.l*. El siguiente ejemplo de ejecución muestra la conducta del programa:

```
$ fl -v -V -f tokens.h
verbose mode is on
version 1.0
File name is: tokens.h
Analyzing tokens.h
#-id-blanks-id-blanks-int-blanks-#-id-blanks-id-blanks-int-blanks-#-id-blanks-id-blanks
-int-blanks-#-id-blanks-id-blanks-int-blanks-#-id-blanks-id-blanks-int-blanks-
```

Los contenidos del fichero *Makefile* definen las dependencias y la estructura de la aplicación:

```
$ cat Makefile
LIBS=-lflarg
CC=gcc -g
LIBPATH=-L. -L~/lib
INCLUDES=-I. -I~/include

fl: main.c lex.arg.c lex.yy.c libflarg.a tokens.h
        $(CC) $(LIBPATH) $(INCLUDES) main.c lex.arg.c lex.yy.c $(LIBS) -o fl
lex.arg.c: fl.l
        flex -Parg fl.l
lex.yy.c: trivial.l tokens.h
        flex trivial.l
libflarg.a: flarg.o
        ar r libflarg.a flarg.o
flarg.o: flarg.c
        $(CC) -c flarg.c
clean:
$ make clean;make
rm lex.arg.c lex.yy.c *.o  fl
flex -Parg fl.l
flex trivial.l
gcc -g -c flarg.c
ar r libflarg.a flarg.o
gcc -g -L. -L~/lib -I. -I~/include main.c lex.arg.c lex.yy.c -lflarg -o fl
```

Observa el uso de la opción `-Parg` en la traducción del fichero *fl.l*. Así no solo el fichero generado por *flex*, sino todas las variables y rutinas accesibles estarán prefijadas por *arg* en vez de *yy*. La librería la denominamos *libflarg.a*. (*flarg* por flex arguments). El correspondiente fichero cabecera será *flarg.h*. Los fuentes de las rutinas que compondrán la librería se mantendrán en el fichero *flarg.c*.

Lo que haremos será redefinir `YY_INPUT(buf, result, max)` para que lea su entrada desde la línea de argumentos.

```
$ cat flarg.h
int yyarglex(int argc, char **argv);
int YY_input_from_argv(char *buf, int max);
int argwrap(void);

#undef YY_INPUT
#define YY_INPUT(buf,result,max) (result = YY_input_from_argv(buf,max))
```

La función `int YY_input_from_argv(char *buf, int max)` utiliza los punteros `char **YY_targv` y `char **YY_arglim` para moverse a través de la familia de argumentos. Mientras que el primero es utilizado para el recorrido, el segundo marca el límite final. Su inicialización ocurre en

```
yyarglex(int argc, char **argv)
```

con las asignaciones:

```
  YY_targv = argv+1;
  YY_arglim = argv+argc;
```

despues, de lo cual, se llama al analizador léxico generado, *arglex* .

```
$ cat flarg.c
char **YY_targv;
char **YY_arglim;
```

```
int YY_input_from_argv(char *buf, int max)
{
  static unsigned offset = 0;

  int len, copylen;

    if (YY_targv >= YY_arglim) return 0;        /* EOF */
    len = strlen(*YY_targv)-offset;     /* amount of current arg */
    if(len >= max) {copylen = max-1; offset += copylen; }
    else copylen = len;
    if(len > 0) memcpy(buf, YY_targv[0]+offset, copylen);
    if(YY_targv[0][offset+copylen] == '\0') {   /* end of arg */
      buf[copylen] = ' '; copylen++; offset = 0; YY_targv++;
    }
    return copylen;
}

int yyarglex(int argc, char **argv) {
  YY_targv = argv+1;
  YY_arglim = argv+argc;
  return arglex();
}

int argwrap(void) {
  return 1;
}
```

El fichero *fl.l* contiene el analizador léxico de la línea de comandos:

```
$ cat fl.l
%{
unsigned verbose;
unsigned thereisfile;
char *progName;
char fileName[256];
#include "flarg.h"
#include "tokens.h"
%}

%%
-h      |
"-?"    |
-help   { printf("usage is: %s [-help | -h | -? ] [-verbose | -v]"
          " [-Version | -V]"
          " [-f filename]\n", progName);
        }

-v      |
-verbose { printf("verbose mode is on\n"); verbose = 1; }

-V      |
-version { printf("version 1.0\n"); }
```

```
-f[[:blank:]]+[^ \t\n]+ {
              strcpy(fileName,argtext+3);
              printf("File name is: %s\n",fileName);
              thereisfile = 1;
            }
.

\n
```

Observe el uso de la clase `[:blank:]` para reconocer los blancos. Las clases son las mismas que las introducidas en *gawk*.

El análisis léxico del fichero que se lee después de procesar la línea de comandos es descrito en *trivial.l*. Partiendo de *trivial.l*, la ejecución del *Makefile* da lugar a la construcción por parte de *flex* del fichero *lex.yy.c* conteniendo la rutina *yylex*.

```
$ cat trivial.l
%{
#include "tokens.h"
%}
digit [0-9]
id [a-zA-Z][a-zA-Z0-9]+
blanks [ \t\n]+
operator [+*/-]
%%
{digit}+ {return INTTOKEN; }
{digit}+"."{digit}+ {return FLOATTOKEN; }
{id} {return IDTOKEN;}
{operator} {return OPERATORTOKEN;}
{blanks} {return BLANKTOKEN;}
. {return (int) yytext[0];}
%%
int yywrap() {
  return 1;
}
```

El fichero *tokens.h* contiene la definición de los *tokens* y es compartido con *main.c*.

```
$ cat tokens.h
#define INTTOKEN  256
#define FLOATTOKEN 257
#define IDTOKEN 258
#define OPERATORTOKEN 259
#define BLANKTOKEN 260
```

Nos queda por presentar el fichero *main.c*:

```
$ cat main.c
#include <stdio.h>
#include "flarg.h"
#include "tokens.h"
extern unsigned verbose;
extern unsigned thereisfile;
extern char *progName;
extern char fileName[256];
extern FILE * yyin;
```

```
main(int argc, char **argv) {
  unsigned lookahead;
  FILE * file;

  progName = *argv;
  yyarglex(argc,argv);
  if (thereisfile) {
    if (verbose) printf("Analyzing %s\n",fileName);
    file = (fopen(fileName,"r"));
    if (file == NULL) exit(1);
    yyin = file;
    while (lookahead = yylex()) {
      switch (lookahead) {
        case INTTOKEN:
            printf("int-");
            break;
          case FLOATTOKEN:
            printf("float-");
            break;
          case IDTOKEN:
            printf("id-");
            break;
          case OPERATORTOKEN:
            printf("operator-");
            break;
          case BLANKTOKEN:
            printf("blanks-");
            break;
          default: printf("%c-",lookahead);
       }
    } /* while */
    printf("\n");
  } /* if */
}
```

29.24. Declaraciones pointer y array

Como se comentó, las opciones `%pointer` y `%array` controlan la definición que *flex* hace de *yytext*. en el caso en que eligamos la opción `%array` la variable `YYLMAX` controla el tamaño del *array*. Supongamos que en el fichero *trivial.l* del ejemplo anterior introducimos las siguientes modificaciones:

```
$ cat trivial.l
%array
%{
#undef YYLMAX
#define YYLMAX 4
#include "tokens.h"
%}
digit [0-9]
id [a-zA-Z][a-zA-Z0-9]+
blanks [ \t\n]+
operator [+*/-]
%%
{digit}+ {return INTTOKEN; }
```

```
{digit}+"."{digit}+ {return FLOATTOKEN; }
{id} {return IDTOKEN;}
{operator} {return OPERATORTOKEN;}
{blanks} {return BLANKTOKEN;}
. {return (int) yytext[0];}
%%
int yywrap() {
  return 1;
  }
```

En tal caso, la definición excesivamente pequeña de YYLMAX provoca un error en tiempo de ejecución:

```
$ fl -V -f tokens.h
version 1.0
File name is: tokens.h
token too large, exceeds YYLMAX
```

29.25. Las Macros YY_USER_ACTION, yy_act e YY_NUM_RULES

La macro YY_USER_ACTION permite ejecutar una acción inmediatamente después del "emparejamiento" y antes de la ejecución de la acción asociada. cuando se la invoca, la variable yy_act contiene el número de la regla que ha emparejado (las reglas se numeran a partir de uno). La macro YY_NUM_RULES contiene el número de reglas, incluyendo la regla por defecto.

El siguiente programa aprovecha dichas macros para mostrar las frecuencias de uso de las reglas.

```
$ cat user_action.l
%array
%{
#include <string.h>

int ctrl[YY_NUM_RULES];
#undef YY_USER_ACTION
#define YY_USER_ACTION { ++ctrl[yy_act]; }
%}
number [0-9]+
id   [a-zA-Z_]+[a-zA-Z0-9_]*
whites [ \t\n]+
%%
{id}
{number}
{whites}
.
%%

int yywrap() {
  int i;

  for(i=1;i<YY_NUM_RULES;i++)
    printf("Rule %d: %d occurrences\n",i,ctrl[i]);
}

$ flex user_action.l ; gcc lex.yy.c -lfl ; a.out
a=b+2*(c-4)
Rule 1: 3 occurrences
```

```
Rule 2: 2 occurrences
Rule 3: 1 occurrences
Rule 4: 6 occurrences
```

29.26. Las opciones `interactive`

La opción `option always-interactive` hace que `flex` genere un analizador que considera que su entrada es "interactiva". Concretamente, el analizador para cada nuevo fichero de entrada, intenta determinar si se trata de un a entrada interactiva o desde fichero haciendo una llamada a la función `isatty()`. Vea un ejemplo de uso de esta función:

```
$ cat isatty.c
#include <unistd.h>
#include <stdio.h>
main() {

  if (isatty(0))
    printf("interactive\n");
  else
    printf("non interactive\n");
}
$ gcc isatty.c; a.out
interactive
$ a.out < isatty.c
non interactive
$
```

cuando se usa la opción `option always-interactive`, se elimina esta llamada.

29.27. La macro `YY_BREAK`

Las acciones asociadas con los patrones se agrupan en la rutina de análisis léxico `yylex()` en una sentencia `switch` y se separan mediante llamadas a la macro `YY_BREAK`. Asi, al compilar con `flex` el siguiente fichero `.l`

```
$ cat interactive.l
%%
. printf("::%c",yytext[0]);
\n printf("::%c",yytext[0]);
```

tenemos el fichero de salida `lex.yy.c` que aparece a continuación (hemos omitido las líneas de código en las que estamos menos interesados, sustituyendolas por puntos suspensivos)

```
/* A lexical scanner generated by flex */
....
#define YY_NUM_RULES 3
#line 1 "interactive.l"
#define INITIAL 0
#line 363 "lex.yy.c"
....
YY_DECL {
 ....
#line 1 "interactive.l"
#line 516 "lex.yy.c"
```

```
 ....
 if ( yy_init ) {
   yy_init = 0;
#ifdef YY_USER_INIT
   YY_USER_INIT;
#endif
   if ( ! yy_start ) yy_start = 1;  /* first start state */
   if ( ! yyin ) yyin = stdin;
   if ( ! yyout ) yyout = stdout;
   if ( ! yy_current_buffer ) yy_current_buffer = yy_create_buffer( yyin, YY_BUF_SIZE );
   yy_load_buffer_state();
 }
 while ( 1 )    /* loops until end-of-file is reached */ {
   ..........................
yy_match:
   do {
    .....
   }
   ............
yy_find_action:
   ............
   YY_DO_BEFORE_ACTION;
   do_action:  /* This label is used only to access EOF actions. */
     switch ( yy_act ) { /* beginning of action switch */
       case 0:
         ...................
         goto yy_find_action;
     case 1:
     YY_RULE_SETUP
     #line 2 "interactive.l"
     printf("::%c",yytext[0]);
       YY_BREAK
     case 2:
     YY_RULE_SETUP
     #line 3 "interactive.l"
     printf("::%c",yytext[0]);
       YY_BREAK
     case 3:
     YY_RULE_SETUP
     #line 4 "interactive.l"
     ECHO;
       YY_BREAK
     #line 614 "lex.yy.c"
     case YY_STATE_EOF(INITIAL):
       yyterminate();
       case YY_END_OF_BUFFER:
         { .....  }
     default:
       YY_FATAL_ERROR("fatal flex scanner internal error--no action found");
  } /* end of action switch */
 } /* end of scanning one token */
} /* end of yylex */
```

```
#if YY_MAIN
int main()
  { yylex(); return 0; }
#endif
#line 4 "interactive.l"
```

Por defecto, la macro `YY_BREAK` es simplemente un `break`. Si cada acción de usuario termina en un `return`, puedes encontrarte con que el compilador genera un buen número de `warning! unreachable code`. Puedes entonces redefinir `YY_BREAK` a vacío y evitar estos mensajes.

Capítulo 30

Expresiones Regulares en sed

· El editor `sed` es un editor no interactivo que actúa ejecutando los comandos (similares a los de `ex`) que figuran en el guión `sed` (*script sed*) sobre los ficheros de entrada y escribiendo el resultado en la salida estandar. Normalmente se llama en una de estas dos formas:

```
sed [opciones] 'comando' fichero(s)
sed [opciones] -f guion fichero(s)
```

Si no se especifican ficheros de entrada, `sed` lee su entrada desde la entrada estandar.

Todos los comandos en el guión son ejecutados sobre todas las líneas de la entrada, salvo que las condiciones en el ámbito del comando indiquen lo contrario.

```
nereida:~/sed> cat b.without.sed
#example of succesive replacements
s/fish/cow/
s/cow/horse/
nereida:~/sed> cat b.test
fish
cow
nereida:~/sed> sed -f b.without.sed b.test
horse
horse
```

como se ve en el ejemplo, si un comando cambia la entrada, los siguientes comandos se aplican a la línea modificada (denominada *pattern space*).

Los comandos `sed` tienen el formato:

```
[direccion1][,direccion2][!]comando[argumentos]
```

30.1. Transferencia de Control

La orden *b* tiene la sintáxis:
[address1][,address2]b[label]
Transfiere el control a la etiqueta especificada. Si no se especifica la etiqueta, el control se transfiere al final del *script*, por lo que no se aplica ningún otro comando a la línea actual. Los dos siguientes ejemplos utilizan *b*:

```
$ cat b.sed
s/fish/cow/
b
s/cow/horse/
$ sed -f b.sed b.test
cow
cow
```

Utilizando *b* con una etiqueta:

```
$ cat blabel.sed
s/fish/cow/
b another
s/cow/horse/
:another
s/cow/cat/
$ sed -f blabel.sed b.test
cat
cat
```

30.2. Inserción de Texto

El uso de llaves {, } permite ejecutar los comandos en la lista entre llaves a las líneas seleccionadas. La llave que cierra debe estar en su propia línea aparte. Las llaves nos permiten, como se ve en el ejemplo, anidar selecciones y expresar condiciones del tipo "si esta entre estos dos patrones y además está entre estos otros dos . . . ".

Los comandos *a* e *i* tienen una sintaxis parecida:

[address]a	*[address]i*
text	*text*

a añade (*i* inserta) el texto en cada línea que casa con la dirección especificada. El *text* no queda disponible en el "*pattern space*", de manera que los subsiguientes comandos no le afectan. El siguiente ejemplo convierte un fichero *ascii* a *html*:

```
$ cat aandi.sed
1{
i\
<html>\
<head>\
<title>
p
i\
</title>\
</head>\
<body bgcolor=white>
}
$a\
</pre>\
</body>\
</html>
$ cat aandi.test
hello.world!
$ sed -f aandi.sed aandi.test
<html>
<head>
<title>
hello.world!
</title>
</head>
<body bgcolor=white>
hello.world!
</pre>
```

```
</body>
</html>
```

30.3. Trasferencia de Control Condicional

La sintaxis de la orden *t* es:

[address1][,address2]t[label]

Si la sustitución tiene éxito, se bifurca a *label*, Si la etiqueta no se especifica, se salta al final del *script*. Consideremos el siguiente fichero de entrada:

```
$ cat t.test
name: "fulano de tal" address: "Leganitos,4" phone: "342255"
name: "fulano de cual"
name: "fulano de alli" address: "Legitos,4"
name: "zutano de tal"  address: "Leg,8"  phone: "342255"
```

Se asume que siempre que figura el teléfono se ha puesto la dirección y que si está la dirección se ha escrito el nombre. Se pretenden rellenar las líneas con un campo por defecto:

```
$ cat t.sed
/name:/{
s/.*name:[ ]*".*"[ ]*address:[ ]*".*"[ ].*phone:[ ]*".*".*/&/
t
s/.*name:[ ]*".*"[ ]*address:[ ]*".*"/& phone: "????"/
t
s/.*name:[ ]*".*"/& address: "???? ????" phone: "????"/
}
```

Esta es la llamada al script y la salida generada:

```
$ sed -f t.sed t.test
name: "fulano de tal" address: "Leganitos,4" phone: "342255"
name: "fulano de cual" address: "???? ????" phone: "????"
name: "fulano de alli" address: "Legitos,4" phone: "????"
name: "zutano de tal"  address: "Leg,8"  phone: "342255"
$
```

Ejemplo:

El fichero de entrada es un *folder* de `pine` en el que los mensajes recibidos tienen el formato:

...

```
01_NAME: XXX
02_SURNAME: XXX
03_TITLE: SISTEMAS
04_OtherTitle:
05_BIRTHDAY: XXX
06_BIRTHMONTH: XXX
07_BIRTHYEAR: XXX
08_ADDRESSFAMILY: XXX
09_ADDRESSACTUAL: XXX
10_POSTALCODE: XXX
11_EMAIL: XXX@csi.ull.es
12_TELEPHONE: XXX
13_FAX: XXX
14_LABGROUP: xxx
15_COMMENTS:
```

...

Se trata de escribir un script que produzca como salida los *emails* de los diferentes alumnos. Esta es una solución (suponemos que el *script* se llama con la opción `-n`):

```
#!/bin/sed -f
/^11_EMAIL:/{
s/11_EMAIL: *\([a-zA-Z0-9]*@[a-zA-Z0-9.]*\)/\1/
t print
s/11_EMAIL: *\([a-zA-Z0-9]*\)/\1@csi.ull.es/
:print
p
}
```

Una característica no comentada y observada en algunos `sed`, incluyendo la versión Linux, es que, si existen varias sustituciones consecutivas antes de la sentencia de transferencia de control, basta con que una tenga éxito para que el salto se realice.

30.4. Rangos

Cuando se usan dos direcciones separadas por una coma, el rango que representan se extiende desde la primera línea que casa con el patrón hasta la siguiente línea que casa con el segundo patrón. El siguiente *script* muestra las tablas que aparecen en un fichero LaTeX:

```
$ cat tables.sed
#Imprime las tablas en un fichero tex
/\\begin{tabular}/,/end{tabular}/p
$ sed -n -f tables.sed *.tex
\begin{tabular}{c|c}
        [address]a$\backslash$ & [address]i$\backslash$\\
        text                   & text\\
\end{tabular}\\
\begin{tabular}{c|c}
        [address]a$\backslash$ & [address]i$\backslash$\\
        text                   & text\\
\end{tabular}\\
```

La siguiente selección para un rango comienza después de la última línea del rango previo; esto es, si el rango es /A/,/B/, el primer conjunto de líneas que casa va desde la primera aparicion de /A/ hasta la siguiente aparición de /B/. Asi, el número mínimo de líneas seleccionadas (si no es cero) es dos. Sólo en el caso en que la primera dirección es una expresión regular y la segunda un número de línea que viene antes de la línea que casa, es que sólo se selecciona una línea de emparejamiento. No hay por tanto solapes entre dos casamientos en un misma rango de direcciones. El siguiente ejemplo ilustra la forma en la que `sed` interpreta los rangos de direcciones. Es una aproximación al problema de escribir los comentarios de un programa *C*.

```
$ cat scope.sed
#execute: sed -n -f scope.sed scope.test
/\/\*/,/\*\//p
```

Este *script* puede funcionar con programas cuyos comentarios (no anidados) empiecen y terminen en líneas diferentes, como el del ejemplo:

```
$ cat scope2.test
#file scope2.test
#include <stdio.h>
```

```
fact(int n) { /* recursive function
if(n == 0) return 1;
else(return n*fact(n-1));*/
}

void toto(id) {
}

main() {
  toto();
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
}
```

La ejecución del script selecciona los dos grupos de líneas que contienen comentarios:

```
$ sed -n -f scope.sed scope2.test
fact(int n) { /* recursive function
if(n == 0) return 1;
else(return n*fact(n-1));*/
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
```

Sin embargo, el script fallará en este otro ejemplo:

```
$ cat scope.test
#include <stdio.h>

fact(int n) { /* recursive function */
if(n == 0) return 1;
else(return n*fact(n-1));
}

void toto(id) {
}

main() {
  toto();
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
}
```

La ejecución del programa da como resultado:

```
$ sed -n -f scope.sed scope.test
fact(int n) { /* recursive function */
if(n == 0) return 1;
else(return n*fact(n-1));
}

void toto(id) {
}

main() {
  toto();
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
```

30.5. Uso de la negación

El ejemplo muestra como convertir los carácteres españoles en un fichero LaTeX escrito usando un teclado español a sus secuencias LaTeX. puesto que en modo matemático los acentos tienen un significado distinto, se aprovecha la disponibilidad de las teclas españolas para simplificar la labor de tecleado. Obsérvese el uso de la exclamación ! para indicar la negación del rango de direcciones seleccionadas.

```
> cat sp2en.sed
/begin{verbatim}/,/end{verbatim}/!{
/begin{math}/,/end{math}/!{
s/á/\\'a/g
s/é/\\'e/g
s/í/\\'{\\i}/g
s/ó/\\'o/g
s/ú/\\'u/g
s/â/\\^a/g
s/ç/\\c{c}/g
s/ñ/\\~n/g
s/Á/\\'A/g
s/É/\\'E/g
s/Í/\\'I/g
s/Ó/\\'O/g
s/Ú/\\'U/g
s/Ñ/\\~N/g
s/¿/>/g
s/¡/</g
}
}
/begin{math}/,/end{math}/{
s/â/\\hat{a}/g
s/á/\\acute{a}/g
s/à/\\grave{a}/g
s/ä/\\ddot{a}/g
s/ê/\\hat{e}/g
s/é/\\acute{e}/g
s/è/\\grave{e}/g
s/ë/\\ddot{e}/g
s/ô/\\hat{o}/g
s/ó/\\acute{o}/g
s/ò/\\grave{o}/g
s/ö/\\ddot{o}/g
s/û/\\hat{u}/g
s/ú/\\acute{u}/g
s/ù/\\grave{u}/g
s/ü/\\ddot{u}/g
}
```

Supongamos el siguiente fichero LaTeX de entrada:

```
> cat sp2en.tex
\documentclass[11pt,a4paper,oneside,onecolumn]{article}
\usepackage{isolatin1}
\title{Stream Editor. Convirtiendo a Ingles en LaTex}
\author{Casiano R. León \thanks{DEIOC Universidad de La Laguna}}
```

```
\begin{document}
\maketitle
Esto es un ejemplo de uso de sp2en.sed:
\begin{center}
áéíóú ÁÉÍÓÚ ñÑ ¿? ¡!\\
\begin{math}
â^{é^{2}} \neq â^{2é}
\end{math}
\end{center}
comienza el verbatim\\
\begin{listing}{1}
Lo que salga en verbatim depende de los packages
que hayan sido cargados: \'a \'e áéíóú ÁÉÍÓÚ ñÑ ¿? ¡!
\end {verbatim}
Termina el verbatim:\\
\begin{center}
áéíóú ÁÉÍÓÚ ñÑ ¿? ¡!
\end{center}
\end{document}
```

Al ejecutar:

```
> sed -f sp2en.sed sp2en.tex
```

Obtenemos la salida:

```
\documentclass[11pt,a4paper,oneside,onecolumn]{article}
\usepackage{isolatin1}
\title{Stream Editor. Convirtiendo a Ingles en LaTex}
\author{Casiano R. Le\'on \thanks{DEIOC Universidad de La Laguna}}
\begin{document}
\maketitle
Esto es un ejemplo de uso de sp2en.sed:
\begin{center}
\'a\'e\'{\i}\'o\'u \'A\'E\'I\'O\'U \~n\~N >? <!\\
\begin{math}
\hat{a}^{\acute{e}^{2}} \neq \hat{a}^{2\acute{e}}
\end{math}
\end{center}
comienza el verbatim\\
\begin{listing}{1}
Lo que salga en verbatim depende de los packages
que hayan sido cargados: \'a \'e áéíóú ÁÉÍÓÚ ñÑ ¿? ¡!
\end{ verbatim}
Termina el verbatim:\\
\begin{center}
\'a\'e\'{\i}\'o\'u \'A\'E\'I\'O\'U \~n\~N >? <!
\end{center}
\end{document}
```

30.6. Siguiente Línea: La orden n

Mediante la orden n podemos reemplazar el espacio de patrones actual por la siguiente línea. Si no se ha utilizado la opción `-n` en la ejecución de `sed`, el contenido del espacio de patrones se imprimirá antes de ser eliminada. El control pasa al comando siguiendo al comando n y no al primer

comando del *script*. Se trata, por tanto, de una orden que altera la conducta habitual de `sed`: Lo común es que, cuando se lee una nueva línea, se ejecute el primer comando del guión.

El siguiente ejemplo extrae los comentarios de un programa *C*. Se asume que, aunque los comentarios se pueden extender sobre varias líneas, no existe mas de un comentario por línea.

```
$ cat n.sed
# run it with: sed -n -f n.sed n.test
# write commented lines in a C program
#If current line matches /* ...
/\/\*/{
# Comienzo de comentario, aseguremonos que la línea no casa con un cierre
# de comentario.
:loop
/\*\//!{
p
n
b loop
}
p
}
```

Supongamos el siguiente programa de prueba:

```
$ cat n.test
#include <stdio.h>

fact(int n) { /* recursive function */
if(n == 0) return 1;
else(return n*fact(n-1));
}

void toto(id) { /* This function */ /* is
                   still empty */
}

main() {
  toto();
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
  toto();
  /* and here is
     a comment
     extended on 3 lines
  */
}
```

La salida al ejecutar el programa es la siguiente:

```
$ sed -n -f n.sed n.test
fact(int n) { /* recursive function */
void toto(id) { /* This function */ /* is
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
  /* and here is
     a comment
```

```
    extended on 3 lines
  */
```

Observe la desaparición de la línea " still empty */" debido a la existencia de dos comentarios, uno de ellos inacabado en la línea anterior.

30.7. Manipulando tablas numéricas

El siguiente ejemplo intercambia la primera y ultima columnas de un fichero como el que sigue:

```
$ cat columns.test
11111 22222 33333 44444 55555
  111    22    33  4444  5555
11111 22222 33333 44444 55555
   11    22     3   444  5555
```

La parte mas complicada es preservar el sangrado. El truco reside, en parte, en el patrón central \2, que memoriza los blancos después de la primera columna y un sólo blanco antes de la última.

```
$ cat columns.sed
s/^\( *[0-9][0-9]*\)\( .*[0-9] \)\( *[0-9][0-9]*\)$/\3\2\1/
$ sed -f columns.sed columns.test
55555 22222 33333 44444 11111
 5555    22    33  4444   111
55555 22222 33333 44444 11111
 5555    22     3   444    11
```

El siguiente ejemplo utiliza una opción del operador de sustitución que permite decidir que aparición del patrón deseamos sustituir. Asi,

`s/A/B/3`

sustituirá la 3 aparición de `A` por `B`, obviando las otras.

El ejemplo selecciona la columna dos del fichero:

```
$ cat col2.sed
#extracts the second column
s/^  *//
s/\<[0-9][0-9]*\>//1
s/\<[0-9][0-9]*\>//2
s/\<[0-9][0-9]*\>//2
s/\<[0-9][0-9]*\>//2
s/  *$//
$ sed -f col2.sed columns.test
 22222
    22
 22222
    22
```

Mas general que el anterior, el siguiente ejemplo elimina un número de columnas arbitrario `$1` por la izquierda y otro número `$2` por la derecha. Para lograrlo, es necesario utilizar un un guión para la `shell` que llama al correspondiente guión `sed`. Los parámetros son introducidos en el guión `sed` mediante el uso apropiado de las comillas dobles y simples:

```
> cat colbranch.sh
#!/bin/bash
sed -e '
s/^\( *[0-9]\+\)\{'"$1"'\}//
s/\( *[0-9]\+\)\{'"$2"'\}$//
'
```

Veamos un ejemplo de uso:

```
> cat columns.test
11111 22222 33333 44444 55555
  111    22    33  4444  5555
11111 22222 33333 44444 55555
   11    22     3   444  5555
> colbranch.sh 2 1 < columns.test
 33333 44444
    33  4444
 33333 44444
     3   444
```

30.8. Traducción entre Tablas

El comando *y* realiza una traducción entre dos tablas de caracteres. Observa el ejemplo:

```
$ cat toupper.sed
y/aáéíóúäëïöübcdefghijklmnopqrstuvxyzñ/AÁÉÍÓÚÄËÏÖÜBCDEFGHIJKLMNOPQRSTUVXYZÑ/
$ cat sp2en.test
¡Coño! ¿Es plím el que hizo plúm?
```

Obtenemos asi los contenidos del fichero en mayúsculas:

```
$ sed -f toupper.sed sp2en.test
¡COÑO! ¿ES PLÍM EL QUE HIZO PLÚM?
```

30.9. Del espacio de Patrones al de Mantenimiento

En `sed` se dispone, como en muchos otros editores de una zona de memoria a la que se puede enviar texto "cortado" o 'copiado" y desde la cual se puede recuperar el texto para insertarlo en la zona de trabajo (*pattern space*). en la jerga `sed` dicho espacio se conoce como *hold space*. El contenido del espacio de patrones (*pattern space*) puede moverse al espacio de mantenimiento (*hold space*) y recíprocamente:

Hold	h ó H	Copia o añade (append) los contenidos del *pattern space* al *hold space*.
Get	g ó G	Copia o añade los contenidos del *hold space* al *pattern space*.
Exchange	x	Intercambia los contenidos del *hold space* y el *pattern space*

Los comandos en minúsculas sobrescriben mientras que los que van en mayúsculas añaden. Asi *h* copia los contenidos del *pattern space* en el *hold space*, borrando los contenidos previos que estuvieran en el *hold space*. Las orden *G* añade (paste) los contenidos del *hold space* al espacio de patrones actual (por el final). Las dos cadenas se enlazan a través de un retorno de carro.
Ejemplo Este guión intenta mediante una heurística poner la definiciones de funciones C al final del fichero, suponiendo que una definición de función comienza en la columna 1 y que se caracterizan mediante uno o dos identificadores seguidos de un paréntesis:

```
$ cat G.sed
/\<if\>/s/.*/&/
t
/\<else\>/s/.*/&/
t
#... lo mismo para las otras palabras clave
```

```
/^[a-zA-Z][a-zA-Z]*([A-Z]*/H
t
/^[a-zA-Z][a-zA-Z]*  *[a-zA-Z][a-zA-Z]*(/H
${
G
}
```

Ejemplo de uso:

```
$ cat p.test
#include <stdio.h>

fact(int n) { /* recursive function */
if(n == 0) return 1;
else(return n*fact(n-1));
}

void toto(id) {
}

main() {
  toto();
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
}
```

Al ejecutar nuestro *script* obtenemos la salida:

```
$ sed -f G.sed p.test
#include <stdio.h>

fact(int n) { /* recursive function */
if(n == 0) return 1;
else(return n*fact(n-1));
}

void toto(id) {
}

main() {
  toto();
  printf("Hello world! factorial of 4 = %d\n",fact(4)); /* the
         comment takes two lines */
}

fact(int n) { /* recursive function */
void toto(id) {
main() {
```

Ejemplo El siguiente guión invierte los contenidos de un fichero:

```
> cat reverse.sed
$!{
:loop
h
n
```

```
G
$!b loop
}
${
p
}
```

He aqui el resultado de una ejecución:

```
> cat reverse.test
one
two
three
> sed -n -f reverse.sed reverse.test
three
two
one
```

Ejemplo

Supuesto un fichero de entrada con formato: ...

```
NAME: xxx
SURNAME: xxx
TITLE: SISTEMAS
OtherTitle:
BIRTHDAY: xxx
BIRTHMONTH: xxx
BIRTHYEAR: xxx
ADDRESSFAMILY: xxx
ADDRESSACTUAL: xxx
POSTALCODE: xxx
EMAIL: XXXXXXX@csi.ull.es
TELEPHONE: xxx
FAX: xxx
LABGROUP: xxx
COMMENTS:

...
```

Se pretenden extraer los apellidos y el nombre, concatenados con una coma. He aqui una posible solución:

```
#!/bin/sed -f
/^NAME:/ {
  s/^NAME://
  h
  n
  s/^SURNAME://
  G
  s/\n/,/
  y/áéíóúabcdefghijklmnñopqrstuvwxyz/ÁÉÍÓÚABCDEFGHIJKLMNÑOPQRSTUVWXYZ/
  p
}
```

30.10. La orden *N*

N añade la siguiente línea de entrada al espacio de patrones. Las dos líneas se separan mediante un retorno de carro. Después de su ejecución, el comando *N* pasa el control a los subsiguientes comandos en el *script*.
El siguiente ejemplo propuesto en [6] muestra una búsqueda y sustitución multilínea. Se trata de sustituir la cadena "Owner and Operator Guide" por "Installation Guide", cualquiera que sea su posición en una línea o entre ellas. Los autores de [6] y [7] proponen la siguiente solución:

```
$ cat multiline2.sed
#Assume the pattern is in no more than two lines
s/Owner and Operator Guide/Installation Guide/g
/Owner/{
  N
  s/ *\n/ /g
  s/Owner  *and  *Operator  *Guide/Installation Guide/g
}
```

Veamos primero los contenidos del fichero de prueba:

```
$ cat multiline.test
Dear Owner: Consult
Section 3.1 in the Owner and
Operator Guide for a description of the tape drives available for the Owner
of your system.

Consult Section 3.1 in the Owner and Operator
Guide for a description of the tape drives
available on your system.

Look in the Owner and Operator Guide, we mean the Owner
and Operator Guide shipped with your system.

Two manuals are provided including the Owner and
Operator Guide and the User Guide.

The Owner and Operator Guide is shipped with your system.

Look in the Owner
and Operator Guide shipped with your system.

The Owner
and
Operator
Guide is shipped with your system.
```

La ejecución del *script* da la siguiente salida:

```
$ sed -f multiline2.sed multiline.test
Dear Owner: Consult Section 3.1 in the Owner and
Operator Guide for a description of the tape drives available for the Owner
of your system.

Consult Section 3.1 in the Installation Guide for a description of the tape
drives
available on your system.
```

```
Look in the Installation Guide, we mean the Installation Guide shipped with
your system.

Two manuals are provided including the Installation Guide and the User Guide.

The Installation Guide is shipped with your system.

Look in the Installation Guide shipped with your system.

The Owner and
Operator
Guide is shipped with your system.
```

Uno de los problemas, que aparece en el primer párrafo del ejemplo de prueba, es que la segunda línea leída debe ser reciclada"para su uso en la siguiente búsqueda. El segundo fallo, que aparece en el último párrafo, es consecuencia de la limitación del *script* para trabajar con patrones partidos en más de dos líneas.
Consideremos esta otra solución:

```
$ cat multiline.sed
s/Owner and Operator Guide/Installation Guide/g
/Owner/{
:label
  N
  s/\n/ /g
  s/Owner  *and  *Operator  *Guide/Installation Guide/g
  /Owner *$/b label
  /Owner  *and *$/b label
  /Owner  *and  *Operator *$/b label
  }
```

Este otro *script* hace que `sed` permanezca en un bucle mientras la línea adjuntada en segundo lugar contenga un prefijo estricto de la cadena buscada.

```
$sed -f multiline.sed multiline.test
Dear Owner: Consult Section 3.1 in the Installation Guide for \
a description of the tape drives available for the Owner  of \
your system.

Consult Section 3.1 in the Installation Guide for a description of the tape drives
available on your system.

Look in the Installation Guide, we mean the Installation Guide \
shipped with your system.

Two manuals are provided including the Installation Guide and the User Guide.

The Installation Guide is shipped with your system.

Look in the Installation Guide shipped with your system.

The Installation Guide is shipped with your system.
```

Un problema que aparece con esta aproximación es la presencia de líneas muy largas. Las líneas permanecen en el espacio de trabajo mientras terminen en un prefijo de la cadena buscada. Para que

la salida quepa en la hoja he tenido que partir las líneas del fichero de salida, lo que he indicado con los símbolos \. Considere esta modificación:

```
#!/bin/sed -f
s/Owner and Operator Guide/Installation Guide/g
/Owner/{
:label
  N
  s/Owner\([ \n]*\)and\([ \n]*\)Operator\([ \n]*\)Guide/Installation\1\2Guide\3/g
  /Owner *$/b label
  /Owner  *and *$/b label
  /Owner  *and  *Operator *$/b label
  }
```

Es indudable la ventaja de disponer de esta capacidad de búsqueda multilínea no puede realizarse con otras utilidades como *ex* o *vi*.

30.11. Suprimir: El Comando *D*

El comando *D* suprime la primera parte (hasta el retorno de carro empotrado) en un espacio de patrones multilínea y bifurca al primer comando en el *script*. El retorno de carro empotrado puede describirse mediante la secuencia de escape `\n`. En el caso en que el espacio de patrones quede vacío como consecuencia de la supresión, se lee una nueva línea.

El siguiente ejemplo compacta una secuencia de líneas vacías en una sóla línea vacía.

```
1 > cat N.sed
2 /^$/{
3 N
4 /^\n$/D
5 }
```

Si la línea es vacía se lee la línea siguiente. Si esta también es vacía el espacio de patrones contiene `^\n$`. La orden `D` deja en el espacio de trabajo una línea vacía y bifurca al comienzo del *script* (sin que se lea una nueva línea). Por tanto nada ha sido impreso, no se ejecuta el comnado final `p` que actúa por defecto. Como el espacio de trabajo contiene `^$`, "casa" con el patrón especificado en línea 2 y se lee la siguiente línea. Si esta nueva línea es no vacía, no se ejecutará la orden `D` de la línea 4 y si que lo hará la orden por defecto final, imprimiéndose la línea vacía y la nueva línea no vacía.

Al ejecutar este "*script*" sobre un fichero conteniendo una secuencia de líneas en blanco:

```
> cat N.test
one empty

two empty lines


three empty lines



end of file
```

Se obtiene el resultado:

```
> sed -f N.sed N.test
one empty
```

```
two empty lines


three empty lines


end of file
```

Un buen ejercicio es intentar predecir la conducta de esta otra solución alternativa, en la que la supresión D es sustituida por la d:

```
/^$/{
N
/^\n$/d
}
```

¿Qué ocurrirá? ¿Es correcta esta segunda solución?

30.12. Búsqueda entre líneas

Este otro caso, tambien esta tomado de [6] y [7]. Se trata de extender la capacidad de búsqueda de *grep*, de modo que el patrón pueda ser encontrado incluso si se encuentra diseminado entre a lo mas dos líneas. El *script* presentado es una ligera variante del que aparece en [6] y escribe la(s) líneas que casan precedidas del número de la segunda línea.

```
$ cat phrase
#! /bin/sh
# phrase -- search for words across two lines.
# Prints the line number
# $1 = search string; remaining args = filenames
search=$1
shift
for file
do
sed -n '
  /'"$search"'/b final
  N
  h
  s/.*\n//
  /'"$search"'/b final
  g
  s/ *\n//
  /'"$search"'/{
    g
    b final
  }
  g
  D
:final
  =
  p
' $file
done
```

Así, con el ejemplo “multiline.test” usado anteriormente obtenemos la siguiente salida:

```
$ phrase "Owner and Operator Guide" multiline.test
3
Section 3.1 in the Owner and
Operator Guide for a description of the tape drives available for the Owner
7
Consult Section 3.1 in the Owner and Operator
Guide for a description of the tape drives
10
Look in the Owner and Operator Guide, we mean the Owner
14
Two manuals are provided including the Owner and
Operator Guide and the User Guide.
16
The Owner and Operator Guide is shipped with your system.
19
Look in the Owner
and Operator Guide shipped with your system.
```

Primero se busca el patrón `/'"$search"'/` en la línea actual. Observe el habilidoso uso de las comillas simples y dobles para permitir la sustitución de la variable. La primera comilla simple cierra la comilla simple al final de la línea 10. Las comillas dobles hacen que la *shell* sustituya `$search` por su valor. La nueva comilla simple permite continuar el texto sin sustituciones.

Si se encuentra el patrón de búsqueda, imprimimos el número de línea (comando `=`) y la línea. Si no, leemos la siguiente línea, formando un patrón multilínea. Salvamos las dos líneas en el *hold space*. Entonces intentamos buscar el patrón `/'"$search"'/` en la línea que acaba de incorporarse. Es por eso que eliminamos del espacio de patrones la primera línea con la orden `s/ *\n//`. Si se encuentra, imprimimos y se repite el ciclo. Si no, recuperamos las dos líneas del *hold space* sustituimos el retorno de carro por un blanco y realizamos una nueva busqueda. Si tiene éxito, se obtienen las dos líneas y se imprimen. En caso contrario, esto es, si el patrón no se ha encontrado en ninguna de las dos líneas, es necesario preservar la última para el siguiente ciclo. Por eso, se obtienen una vez mas las líneas del *hold space* y se suprime con *D* la primera de ellas. Dado que *D* devuelve el control al comienzo del *script*, la segunda línea no es eliminada. De todos modos, el *script* no es capaz de captar cuando un prefijo del patrón aparece al final de esta segunda línea, como muestra el ejemplo de prueba. En el primer párrafo el patrón se encuentra dos veces y sólo es encontrado una.

30.13. Seleccionando Items en un Registro Multilínea

El ejercicio resuelto aqui consiste en listar los alumnos que han seleccionado un determinado grupo de prácticas. Suponemos la organización del fichero de entrada descrita en la sección 30.9. El *script* recibe como primer argumento el nombre del fichero conteniendo la carpeta de correo (`pine`) asociada con la asignatura y como segundo argumento el grupo. Un primer *script* que no es descrito aqui, denominado `makefichas.sed` produce como salida el archivo con la estructura descrita en la sección 30.9. El segundo guión, denominado `grupo.sh` y que es el que nos ocupa, produce como salida los alumnos que pertenecen a ese grupo ed prácticas.

Estos son los contenidos del *script* `grupo`:

```
~/bin/makefichas.sed -n ~/mail/$1 | grupo.sh $2 | sort -u
```

Los contenidos del fichero `grupo.sh` son:

```
1  #!/bin/bash
2  search=$1
3  sed -n '
4  /^NAME:/ {
5    s/^NAME://
```

```
 6    h
 7    n
 8    s/^SURNAME://
 9    G
10    s/\n/,/
11    y/ÁÉÍÓÚáéíóúabcdefghijklmnñopqrstuvxyz/AEIOUAEIOUABCDEFGHIJKLMNÑOPQRSTUVXYZ/
12   h
13 }
14 /^LABGROUP:/ {
15   y/ÁÉÍÓÚáéíóúabcdefghijklmnñopqrstuvxyz/AEIOUAEIOUABCDEFGHIJKLMNÑOPQRSTUVXYZ/
16   s/'"$search"'/&/
17   t print
18   b
19 :print
20   g
21   p
22 }
23 '
```

De nuevo hacemos uso de las comillas simples y dobles en este ejemplo. Obsérvese como se proteje el guión `sed` entre las líneas 3 y 16. En la línea 16 el cierre de la comilla simple y el uso de la doble comilla permite la actuación de la interpretación de la `shell`, sustituyendo `$search` que coincide con el parámetro pasado en la llamada como `$2`. La siguiente comilla simple en esa línea permite la protección del resto del guión.

Capítulo 31

Expresiones Regulares en Perl

31.1. Introducción

Los rudimentos de las expresiones regulares pueden encontrarse en los trabajos pioneros de McCullogh y Pitts (1940) sobre redes neuronales. El lógico Stephen Kleene definió formalmente el algebra que denominó *conjuntos regulares* y desarrollo una notación para la descripción de dichos conjuntos, las *expresiones regulares*.

Durante las décadas de 1960 y 1970 hubo un desarrollo formal de las expresiones regulares. Una de las priemras publicaciones que utilizan las expresiones regulares en un marco informático es el artículo de 1968 de Ken Thompson *Regular Expression Search Algorithm* en el que describe un compilador de expresiones regulares que produce código objeto para un IBM 7094. Este compilador dió lugar al editor *qed*, en el cual se basó el editor de Unix *ed*. Aunque las expresiones regulares de este último no eran tan sofisticadas como las de *qed*, fueron las primeras en ser utilizadas en un contexto no académico. Se dice que el comando global `g` en su formato `g/re/p` que utilizaba para imprimir (opción `p`) las líneas que casan con la expresión regular `re` dió lugar a un programa separado al que se denomino `grep`.

Las expresiones regulares facilitadas por las primeras versiones de estas herramientas eran limitadas. Por ejemplo, se disponía del cierre de Kleene `*` pero no del cierre positivo `+` o del operador opcional `?`. Por eso, posteriormente, se han introducido los metacaracteres `\+` y `\?`. Existían numerosas limitaciones en dichas versiones, por ej. `$` sólo significa "final de línea" al final de la expresión regular. Eso dificulta expresiones como

```
grep 'cierre$\|^Las' viq.tex
```

Sin embargo, la mayor parte de las versiones actuales resuelven correctamente estos problemas:

```
nereida:~/viq> grep 'cierre$\|^Las' viq.tex
Las expresiones regulares facilitadas por las primeras versiones de estas herramientas
eran limitadas. Por ejemplo, se disponía del cierre de Kleene \verb|*| pero no del cierre
nereida:~/viq>
```

De hecho AT&T Bell labs añadió numerosas funcionalidades, como por ejemplo, el uso de `\{min, max\}`, tomada de *lex*. Por esa época, Alfred Aho escribió *egrep* que, no sólo proporciona un conjunto mas rico de operadores sino que mejoró la implementación. Mientras que el *grep* de Ken Thompson usaba un autómata finito no determinista (NFA), la versión de *egrep* de Aho usa un autómata finito determinista (DFA).

En 1986 Henry Spencer desarrolló la librería *regex* para el lenguaje `C`, que proporciona un conjunto consistente de funciones que permiten el manejo de expresiones regulares. Esta librería ha contribuido a "homogeneizar" la sintáxis y semántica de las diferentes herramientas que utilizan expresiones regulares (como *awk*, *lex*, *sed*, ...).

Véase También

- La sección *Expresiones Regulares en Otros Lenguajes* 31.3

- Regular Expressions Cookbook. Jan Goyvaerts, Steven Levithan
- PCRE (Perl Compatible Regular Expressions) en la Wikipedia
- PCRE (Perl Compatible Regular Expressions)
- Java Regular Expressions
- C# Regular Expressions
- .NET Framework Regular Expressions

31.1.1. Un ejemplo sencillo

Matching en Contexto Escalar

```
pl@nereida:~/Lperltesting$ cat -n c2f.pl
  1    #!/usr/bin/perl -w
  2    use strict;
  3
  4    print "Enter a temperature (i.e. 32F, 100C):\n";
  5    my $input = <STDIN>;
  6    chomp($input);
  7
  8    if ($input !~ m/^([-+]?[0-9]+(\.[0-9]*)?)\s*([CF])$/i) {
  9      warn "Expecting a temperature, so don't understand \"$input\".\n";
 10    }
 11    else {
 12      my $InputNum = $1;
 13      my $type = $3;
 14      my ($celsius, $farenheit);
 15      if ($type eq "C" or $type eq "c") {
 16        $celsius = $InputNum;
 17        $farenheit = ($celsius * 9/5)+32;
 18      }
 19      else {
 20        $farenheit = $InputNum;
 21        $celsius = ($farenheit -32)*5/9;
 22      }
 23      printf "%.2f C = %.2f F\n", $celsius, $farenheit;
 24    }
```

Véase también:

- `perldoc perlrequick`
- `perldoc perlretut`
- `perldoc perlre`
- `perldoc perlreref`

Ejecución con el depurador:

```
pl@nereida:~/Lperltesting$ perl -wd c2f.pl
Loading DB routines from perl5db.pl version 1.28
Editor support available.
```

```
Enter h or ‘h h’ for help, or ‘man perldebug’ for more help.
main::(c2f.pl:4):       print "Enter a temperature (i.e. 32F, 100C):\n";
DB<1> c 8
Enter a temperature (i.e. 32F, 100C):
32F
main::(c2f.pl:8):       if ($input !~ m/^([-+]?[0-9]+(\.[0-9]*)?)\s*([CF])$/i) {
DB<2> n
main::(c2f.pl:12):        my $InputNum = $1;
DB<2> x ($1, $2, $3)
0  32
1  undef
2  'F'
DB<3> use YAPE::Regex::Explain
DB<4> p YAPE::Regex::Explain->new('([-+]?[0-9]+(\.[0-9]*)?)\s*([CF])$')->explain
The regular expression:
(?-imsx:([-+]?[0-9]+(\.[0-9]*)?)\s*([CF])$)
matches as follows:

NODE                     EXPLANATION
----------------------------------------------------------------------
(?-imsx:                 group, but do not capture (case-sensitive)
                         (with ^ and $ matching normally) (with . not
                         matching \n) (matching whitespace and #
                         normally):
----------------------------------------------------------------------
  (                        group and capture to \1:
----------------------------------------------------------------------
    [-+]?                    any character of: '-', '+' (optional
                             (matching the most amount possible))
----------------------------------------------------------------------
    [0-9]+                   any character of: '0' to '9' (1 or more
                             times (matching the most amount
                             possible))
----------------------------------------------------------------------
    (                        group and capture to \2 (optional
                             (matching the most amount possible)):
----------------------------------------------------------------------
      \.                       '.'
----------------------------------------------------------------------
      [0-9]*                   any character of: '0' to '9' (0 or
                               more times (matching the most amount
                               possible))
----------------------------------------------------------------------
    )?                       end of \2 (NOTE: because you're using a
                             quantifier on this capture, only the
                             LAST repetition of the captured pattern
                             will be stored in \2)
----------------------------------------------------------------------
  )                        end of \1
----------------------------------------------------------------------
  \s*                      whitespace (\n, \r, \t, \f, and " ") (0 or
                           more times (matching the most amount
                           possible))
```

```
------------------------------------------------------------------------
  (                       group and capture to \3:
------------------------------------------------------------------------
    [CF]                    any character of: 'C', 'F'
------------------------------------------------------------------------
  )                       end of \3
------------------------------------------------------------------------
  $                       before an optional \n, and the end of the
                          string
------------------------------------------------------------------------
)                       end of grouping
------------------------------------------------------------------------
```

Dentro de una expresión regular es necesario referirse a los textos que casan con el primer, paréntesis, segundo, etc. como `\1`, `\2`, etc. La notación `$1` se refieré a lo que casó con el primer paréntesis en el último *matching*, no en el actual. Veamos un ejemplo:

```
pl@nereida:~/Lperltesting$ cat -n dollar1slash1.pl
   1    #!/usr/bin/perl -w
   2    use strict;
   3
   4    my $a = "hola juanito";
   5    my $b = "adios anita";
   6
   7    $a =~ /(ani)/;
   8    $b =~ s/(adios) *($1)/\U$1 $2/;
   9    print "$b\n";
```

Observe como el `$1` que aparece en la cadena de reemplazo (línea 8) se refiere a la cadena `adios` mientras que el `$1` en la primera parte contiene `ani`:

```
pl@nereida:~/Lperltesting$ ./dollar1slash1.pl
ADIOS ANIta
```

Ejercicio 31.1.1. *Indique cuál es la salida del programa anterior si se sustituye la línea 8 por*

```
$b =~ s/(adios) *(\1)/\U$1 $2/;
```

Número de Paréntesis

El número de paréntesis con memoria no está limitado:

```
pl@nereida:~/Lperltesting$ perl -wde 0
main::(-e:1):   0
            123456789ABCDEF
DB<1> $x = "123456789AAAAAA"
                  1  2  3  4  5  6  7  8  9 10 11  12
DB<2> $r = $x =~ /(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)\11/; print "$r\n$10\n$11\n"
1
A
A
```

Véase el siguiente párrafo de `perlre` (sección Capture buffers):

> *There is no limit to the number of captured substrings that you may use. However Perl also uses* `\10`*,* `\11`*, etc. as aliases for* `\010`*,* `\011`*, etc. (Recall that* `0` *means octal, so* `\011` *is the character at number* `9` *in your coded character set; which would be the 10th*

character, a horizontal tab under ASCII.) Perl resolves this ambiguity by interpreting `\10` *as a backreference only if at least* `10` *left parentheses have opened before it. Likewise* `\11` *is a backreference only if at least* `11` *left parentheses have opened before it. And so on.* `\1` *through* `\9` *are always interpreted as backreferences.*

Contexto de Lista

Si se utiliza en un contexto que requiere una lista, el "pattern match" retorna una lista consistente en las subexpresiones casadas mediante los paréntesis, esto es `$1`, `$2`, `$3`, ... Si no hubiera emparejamiento se retorna la lista vacía. Si lo hubiera pero no hubieran paréntesis se retorna la lista `($&)`.

```
pl@nereida:~/src/perl/perltesting$ cat -n escapes.pl
     1  #!/usr/bin/perl -w
     2  use strict;
     3
     4  my $foo = "one two three four five\nsix seven";
     5  my ($F1, $F2, $Etc) = ($foo =~ /^\s*(\S+)\s+(\S+)\s*(.*)/);
     6  print "List Context: F1 = $F1, F2 = $F2, Etc = $Etc\n";
     7
     8  # This is 'almost' the same than:
     9  ($F1, $F2, $Etc) = split(/\s+/, $foo, 3);
    10  print "Split: F1 = $F1, F2 = $F2, Etc = $Etc\n";
```

Observa el resultado de la ejecución:

```
pl@nereida:~/src/perl/perltesting$ ./escapes.pl
List Context: F1 = one, F2 = two, Etc = three four five
Split: F1 = one, F2 = two, Etc = three four five
six seven
```

El modificador `s`

La opción `s` usada en una regexp hace que el punto `'.'` case con el retorno de carro:

```
pl@nereida:~/src/perl/perltesting$ perl -wd ./escapes.pl
main::(./escapes.pl:4): my $foo = "one two three four five\nsix seven";
DB<1> c 9
List Context: F1 = one, F2 = two, Etc = three four five
main::(./escapes.pl:9): ($F1, $F2, $Etc) = split(' ',$foo, 3);
DB<2> ($F1, $F2, $Etc) = ($foo =~ /^\s*(\S+)\s+(\S+)\s*(.*)/s)
DB<3> p "List Context: F1 = $F1, F2 = $F2, Etc = $Etc\n"
List Context: F1 = one, F2 = two, Etc = three four five
six seven
```

La opción `/s` hace que `.` se empareje con un `\n`. Esto es, casa con cualquier carácter.

Veamos otro ejemplo, que imprime los nombres de los ficheros que contienen cadenas que casan con un patrón dado, incluso si este aparece disperso en varias líneas:

```
1   #!/usr/bin/perl -w
2   #use:
3   #smodifier.pl 'expr' files
4   #prints the names of the files that match with the give expr
5   undef $/; # input record separator
6   my $what = shift @ARGV;
7   while(my $file = shift @ARGV) {
8     open(FILE, "<$file");
```

```
 9     $line =  <FILE>;
10     if ($line =~ /$what/s) {
11       print "$file\n";
12     }
13   }
```

Ejemplo de uso:

```
> smodifier.pl 'three.*three' double.in split.pl doublee.pl
double.in
doublee.pl
```

Vea la sección 31.4.2 para ver los contenidos del fichero `double.in`. En dicho fichero, el patrón `three.*three` aparece repartido entre varias líneas.

El modificador `m`

El modificador `s` se suele usar conjuntamente con el modificador `m`. He aquí lo que dice la seccion *Using character classes* de la sección 'Using-character-classes' en `perlretut` al respecto:

> - `m` *modifier (*`//m`*): Treat string as a set of multiple lines.* `'.'` *matches any character except* `\n`*.* `^` *and* `$` *are able to match at the start or end of any line within the string.*
> - *both* `s` *and* `m` *modifiers (*`//sm`*): Treat string as a single long line, but detect multiple lines.* `'.'` *matches any character, even* `\n` *.* `^` *and* `$` *, however, are able to match at the start or end of any line within the string.*
>
> *Here are examples of //s and //m in action:*
>
> ```
> 1. $x = "There once was a girl\nWho programmed in Perl\n";
> 2.
> 3. $x =~ /^Who/; # doesn't match, "Who" not at start of string
> 4. $x =~ /^Who/s; # doesn't match, "Who" not at start of string
> 5. $x =~ /^Who/m; # matches, "Who" at start of second line
> 6. $x =~ /^Who/sm; # matches, "Who" at start of second line
> 7.
> 8. $x =~ /girl.Who/; # doesn't match, "." doesn't match "\n"
> 9. $x =~ /girl.Who/s; # matches, "." matches "\n"
> 10. $x =~ /girl.Who/m; # doesn't match, "." doesn't match "\n"
> 11. $x =~ /girl.Who/sm; # matches, "." matches "\n"
> ```
>
> *Most of the time, the default behavior is what is wanted, but* `//s` *and* `//m` *are occasionally very useful. If* `//m` *is being used, the start of the string can still be matched with* `\A` *and the end of the string can still be matched with the anchors* `\Z` *(matches both the end and the newline before, like* `$`*), and* `\z` *(matches only the end):*
>
> ```
> 1. $x =~ /^Who/m; # matches, "Who" at start of second line
> 2. $x =~ /\AWho/m; # doesn't match, "Who" is not at start of string
> 3.
> 4. $x =~ /girl$/m; # matches, "girl" at end of first line
> 5. $x =~ /girl\Z/m; # doesn't match, "girl" is not at end of string
> 6.
> 7. $x =~ /Perl\Z/m; # matches, "Perl" is at newline before end
> 8. $x =~ /Perl\z/m; # doesn't match, "Perl" is not at end of string
> ```

Normalmente el carácter `^` casa solamente con el comienzo de la cadena y el carácter `$` con el final. Los `\n` empotrados no casan con `^` ni `$`. El modificador `/m` modifica esta conducta. De este modo `^` y `$` casan con cualquier frontera de línea interna. Las anclas `\A` y `\Z` se utilizan entonces para casar con el comienzo y final de la cadena. Véase un ejemplo:

```
nereida:~/perl/src> perl -de 0
  DB<1> $a = "hola\npedro"
  DB<2> p "$a"
hola
pedro
  DB<3> $a =~ s/.*/x/m
  DB<4> p $a
x
pedro
  DB<5> $a =~ s/^pedro$/juan/
  DB<6> p "$a"
x
pedro
  DB<7> $a =~ s/^pedro$/juan/m
  DB<8>  p "$a"
x
juan
```

El conversor de temperaturas reescrito usando contexto de lista

Reescribamos el ejemplo anterior usando un contexto de lista:

```
casiano@millo:~/Lperltesting$ cat -n c2f_list.pl
  1    #!/usr/bin/perl -w
  2    use strict;
  3
  4    print "Enter a temperature (i.e. 32F, 100C):\n";
  5    my $input = <STDIN>;
  6    chomp($input);
  7
  8    my ($InputNum, $type);
  9
 10    ($InputNum, $type) = $input =~ m/^
 11                                       ([-+]?[0-9]+(?:\.[0-9]*)?) # Temperature
 12                                       \s*
 13                                       ([cCfF]) # Celsius or Farenheit
 14                                     $/x;
 15
 16    die "Expecting a temperature, so don't understand \"$input\".\n" unless defined($InputN
 17
 18    my ($celsius, $fahrenheit);
 19    if ($type eq "C" or $type eq "c") {
 20      $celsius = $InputNum;
 21      $fahrenheit = ($celsius * 9/5)+32;
 22    }
 23    else {
 24      $fahrenheit = $InputNum;
 25      $celsius = ($fahrenheit -32)*5/9;
 26    }
 27    printf "%.2f C = %.2f F\n", $celsius, $fahrenheit;
```

La opción `x`

La opción `/x` en una regexp permite utilizar comentarios y espacios dentro de la expresión regular. Los espacios dentro de la expresión regular dejan de ser significativos. Si quieres conseguir un espacio

que sea significativo, usa `\s` o bien escápalo. Véase la sección 'Modifiers' en `perlre` y la sección 'Building-a-regexp' en `perlretut`.

Paréntesis sin memoria

La notación `(?: ... )` se usa para introducir paréntesis de agrupamiento sin memoria. `(?: ...)` Permite agrupar las expresiones tal y como lo hacen los paréntesis ordinarios. La diferencia es que no "memorizan" esto es no guardan nada en `$1`, `$2`, etc. Se logra así una compilación mas eficiente. Veamos un ejemplo:

```
> cat groupingpar.pl
#!/usr/bin/perl

  my $a = shift;

  $a =~ m/(?:hola )*(juan)/;
  print "$1\n";
nereida:~/perl/src> groupingpar.pl 'hola juan'
juan
```

Interpolación en los patrones: La opción `o`

El patrón regular puede contener variables, que serán interpoladas (en tal caso, el patrón será re-compilado). Si quieres que dicho patrón se compile una sóla vez, usa la opción `/o`.

```
pl@nereida:~/Lperltesting$ cat -n mygrep.pl
     1  #!/usr/bin/perl -w
     2  my $what = shift @ARGV || die "Usage $0 regexp files ...\n";
     3  while (<>) {
     4    print "File $ARGV, rel. line $.: $_" if (/$what/o); # compile only once
     5  }
     6
```

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lperltesting$ ./mygrep.pl
Usage ./mygrep.pl regexp files ...
pl@nereida:~/Lperltesting$ ./mygrep.pl if labels.c
File labels.c, rel. line 7:        if (a < 10) goto LABEL;
```

El siguiente texto es de la sección 'Using-regular-expressions-in-Perl' en `perlretut`:

> *If* `$pattern` *won't be changing over the lifetime of the script, we can add the //o modifier, which directs Perl to only perform variable substitutions once*

Otra posibilidad es hacer una compilación previa usando el operador `qr` (véase la sección 'Regexp-Quote-Like-Operators' en `perlop`). La siguiente variante del programa anterior también compila el patrón una sóla vez:

```
pl@nereida:~/Lperltesting$ cat -n mygrep2.pl
     1  #!/usr/bin/perl -w
     2  my $what = shift @ARGV || die "Usage $0 regexp files ...\n";
     3  $what = qr{$what};
     4  while (<>) {
     5    print "File $ARGV, rel. line $.: $_" if (/$what/);
     6  }
```

Véase

- El nodo en perlmonks /o is dead, long live qr//! por diotalevi

Cuantificadores greedy

El siguiente extracto de la sección *Matching Repetitions* en la sección 'Matching-repetitions' en `perlretut` ilustra la semántica *greedy* de los operadores de repetición `*+{}?` etc.

> *For all of these quantifiers, Perl will try to match as much of the string as possible, while still allowing the regexp to succeed. Thus with* `/a?.../`*, Perl will first try to match the regexp with the a present; if that fails, Perl will try to match the regexp without the a present. For the quantifier* `*` *, we get the following:*
>
> ```
> 1. $x = "the cat in the hat";
> 2. $x =~ /^(.*)(cat)(.*)$/; # matches,
> 3. # $1 = 'the '
> 4. # $2 = 'cat'
> 5. # $3 = ' in the hat'
> ```
>
> *Which is what we might expect, the match finds the only cat in the string and locks onto it. Consider, however, this regexp:*
>
> ```
> 1. $x =~ /^(.*)(at)(.*)$/; # matches,
> 2. # $1 = 'the cat in the h'
> 3. # $2 = 'at'
> 4. # $3 = '' (0 characters match)
> ```
>
> *One might initially guess that Perl would find the* `at` *in* `cat` *and stop there, but that wouldn't give the longest possible string to the first quantifier* `.*`*. Instead, the first quantifier* `.*` *grabs as much of the string as possible while still having the regexp match. In this example, that means having the* `at` *sequence with the final* `at` *in the string.*
>
> *The other important principle illustrated here is that when there are two or more elements in a regexp, the leftmost quantifier, if there is one, gets to grab as much the string as possible, leaving the rest of the regexp to fight over scraps. Thus in our example, the first quantifier* `.*` *grabs most of the string, while the second quantifier* `.*` *gets the empty string. Quantifiers that grab as much of the string as possible are called maximal match or greedy quantifiers.*
>
> *When a regexp can match a string in several different ways, we can use the principles above to predict which way the regexp will match:*
>
> - **Principle 0:** *Taken as a whole, any regexp will be matched at the earliest possible position in the string.*
> - **Principle 1:** *In an alternation* `a|b|c...` *, the leftmost alternative that allows a match for the whole regexp will be the one used.*
> - **Principle 2:** *The maximal matching quantifiers* `?`*,* `*`*,* `+` *and* `{n,m}` *will in general match as much of the string as possible while still allowing the whole regexp to match.*
> - **Principle 3:** *If there are two or more elements in a regexp, the leftmost greedy quantifier, if any, will match as much of the string as possible while still allowing the whole regexp to match. The next leftmost greedy quantifier, if any, will try to match as much of the string remaining available to it as possible, while still allowing the whole regexp to match. And so on, until all the regexp elements are satisfied.*

Regexp y Bucles Infinitos

El siguiente párrafo está tomado de la sección 'Repeated-Patterns-Matching-a-Zero-length-Substring' en `perlre`:

> *Regular expressions provide a terse and powerful programming language. As with most other power tools, power comes together with the ability to wreak havoc.*
>
> *A common abuse of this power stems from the ability to make infinite loops using regular expressions, with something as innocuous as:*

```
1. 'foo' =~ m{ ( o? )* }x;
```

The `o?` *matches at the beginning of* `'foo'` *, and since the position in the string is not moved by the match,* `o?` *would match again and again because of the* `*` *quantifier.*

Another common way to create a similar cycle is with the looping modifier `//g` *:*

```
1. @matches = ( 'foo' =~ m{ o? }xg );
```

or

```
1. print "match: <$&>\n" while 'foo' =~ m{ o? }xg;
```

or the loop implied by `split()`*.*

... Perl allows such constructs, by forcefully breaking the infinite loop. The rules for this are different for lower-level loops given by the greedy quantifiers `*+{}` *, and for higher-level ones like the* `/g` *modifier or* `split()` *operator.*

The lower-level loops are interrupted (that is, the loop is broken) when Perl detects that a repeated expression matched a zero-length substring. Thus

```
1.  m{ (?: NON_ZERO_LENGTH | ZERO_LENGTH )* }x;
```

is made equivalent to

```
1.  m{ (?: NON_ZERO_LENGTH )*
2.  |
3.  (?: ZERO_LENGTH )?
4.  }x;
```

The higher level-loops preserve an additional state between iterations: whether the last match was zero-length. To break the loop, the following match after a zero-length match is prohibited to have a length of zero. This prohibition interacts with backtracking (see Backtracking), and so the second best match is chosen if the best match is of zero length.

For example:

```
1. $_ = 'bar';
2. s/\w??/<$&>/g;
```

results in `<><b><><a><><r><>` *. At each position of the string the best match given by non-greedy* `??` *is the zero-length match, and the second best match is what is matched by* `\w` *. Thus zero-length matches alternate with one-character-long matches.*

Similarly, for repeated `m/()/g` *the second-best match is the match at the position one notch further in the string.*

The additional state of being matched with zero-length is associated with the matched string, and is reset by each assignment to `pos()`*. Zero-length matches at the end of the previous match are ignored during* `split`*.*

Ejercicio 31.1.2. ▪ *Explique la conducta del siguiente matching:*

```
  DB<25> $c = 0

  DB<26>  print(($c++).": <$&>\n") while 'aaaababab' =~ /a*(ab)*/g;
0: <aaaa>
1: <>
2: <a>
3: <>
4: <a>
5: <>
6: <a>
7: <>
8: <>
```

Cuantificadores *lazy*

Las expresiones *lazy* o *no greedy* hacen que el NFA se detenga en la cadena mas corta que casa con la expresión. Se denotan como sus análogas *greedy* añadiéndole el postfijo `?`:

- `{n,m}?`
- `{n,}?`
- `{n}?`
- `*?`
- `+?`
- `??`

Repasemos lo que dice la sección Matching Repetitions en la sección 'Matching-repetitions' en `perlretut`:

> *Sometimes greed is not good. At times, we would like quantifiers to match a minimal piece of string, rather than a maximal piece. For this purpose, Larry Wall created the minimal match or non-greedy quantifiers* `??` ,`*?`, `+?`, *and* `{}?`. *These are the usual quantifiers with a ? appended to them. They have the following meanings:*
>
> - `a??` *means: match 'a' 0 or 1 times. Try 0 first, then 1.*
> - `a*?` *means: match 'a' 0 or more times, i.e., any number of times, but as few times as possible*
> - `a+?` *means: match 'a' 1 or more times, i.e., at least once, but as few times as possible*
> - `a{n,m}?` *means: match at least n times, not more than m times, as few times as possible*
> - `a{n,}?` *means: match at least n times, but as few times as possible*
> - `a{n}?` *means: match exactly n times. Because we match exactly n times, an? is equivalent to an and is just there for notational consistency.*
>
> *Let's look at the example above, but with minimal quantifiers:*

```
1. $x = "The programming republic of Perl";
2. $x =~ /^(.+?)(e|r)(.*)$/; # matches,
3. # $1 = 'Th'
4. # $2 = 'e'
5. # $3 = ' programming republic of Perl'
```

The minimal string that will allow both the start of the string `^` *and the alternation to match is* `Th` *, with the alternation* `e|r` *matching* `e`*. The second quantifier* `.*` *is free to gobble up the rest of the string.*

```
1. $x =~ /(m{1,2}?)(.*?)$/; # matches,
2. # $1 = 'm'
3. # $2 = 'ming republic of Perl'
```

The first string position that this regexp can match is at the first `m` *in programming . At this position, the minimal* `m{1,2}?` *matches just one* `m` *. Although the second quantifier* `.*?` *would prefer to match no characters, it is constrained by the end-of-string anchor* `$` *to match the rest of the string.*

```
1. $x =~ /(.*?)(m{1,2}?)(.*)$/; # matches,
2. # $1 = 'The progra'
3. # $2 = 'm'
4. # $3 = 'ming republic of Perl'
```

In this regexp, you might expect the first minimal quantifier `.*?` *to match the empty string, because it is not constrained by a* `^` *anchor to match the beginning of the word. Principle 0 applies here, however. Because it is possible for the whole regexp to match at the start of the string, it will match at the start of the string. Thus the first quantifier has to match everything up to the first m. The second minimal quantifier matches just one* `m` *and the third quantifier matches the rest of the string.*

```
1. $x =~ /(.??)(m{1,2})(.*)$/; # matches,
2. # $1 = 'a'
3. # $2 = 'mm'
4. # $3 = 'ing republic of Perl'
```

Just as in the previous regexp, the first quantifier `.??` *can match earliest at position* `a` *, so it does. The second quantifier is greedy, so it matches mm , and the third matches the rest of the string.*

We can modify principle 3 above to take into account non-greedy quantifiers:

- **Principle 3:** *If there are two or more elements in a regexp, the leftmost greedy (non-greedy) quantifier, if any, will match as much (little) of the string as possible while still allowing the whole regexp to match. The next leftmost greedy (non-greedy) quantifier, if any, will try to match as much (little) of the string remaining available to it as possible, while still allowing the whole regexp to match. And so on, until all the regexp elements are satisfied.*

Ejercicio 31.1.3. *Explique cuál será el resultado de el segundo comando de matching introducido en el depurador:*

```
casiano@millo:~/Lperltesting$ perl -wde 0
main::(-e:1):   0
  DB<1> x ('1'x34) =~ m{^(11+)\1+$}
0  11111111111111111
  DB<2> x ('1'x34) =~ m{^(11+?)\1+$}
????????????????????????????????????
```

Descripción detallada del proceso de matching Veamos en detalle lo que ocurre durante un matching. Repasemos lo que dice la sección Matching Repetitions en la sección 'Matching-repetitions' en `perlretut`:

> *Just like alternation, quantifiers are also susceptible to backtracking. Here is a step-by-step analysis of the example*
>
> ```
> 1. $x = "the cat in the hat";
> 2. $x =~ /^(.*)(at)(.*)$/; # matches,
> 3. # $1 = 'the cat in the h'
> 4. # $2 = 'at'
> 5. # $3 = '' (0 matches)
> ```
>
> 1. *Start with the first letter in the string 't'.*
> 2. *The first quantifier '.*' starts out by matching the whole string 'the cat in the hat'.*
> 3. *'a' in the regexp element 'at' doesn't match the end of the string. Backtrack one character.*
> 4. *'a' in the regexp element 'at' still doesn't match the last letter of the string 't', so backtrack one more character.*
> 5. *Now we can match the 'a' and the 't'.*
> 6. *Move on to the third element '.*'. Since we are at the end of the string and '.*' can match 0 times, assign it the empty string.*
> 7. *We are done!*

Rendimiento

La forma en la que se escribe una regexp puede dar lugar agrandes variaciones en el rendimiento. Repasemos lo que dice la sección Matching Repetitions en la sección 'Matching-repetitions' en `perlretut`:

> *Most of the time, all this moving forward and backtracking happens quickly and searching is fast. There are some pathological regexps, however, whose execution time exponentially grows with the size of the string. A typical structure that blows up in your face is of the form*
>
> ```
> /(a|b+)*/;
> ```
>
> *The problem is the nested indeterminate quantifiers. There are many different ways of partitioning a string of length n between the* `+` *and* `*`*: one repetition with* `b+` *of length n, two repetitions with the first* `b+` *length k and the second with length* $n - k$*, m repetitions whose bits add up to length n, etc.*
>
> *In fact there are an exponential number of ways to partition a string as a function of its length. A regexp may get lucky and match early in the process, but if there is no match, Perl will try every possibility before giving up. So be careful with nested* `*`*'s,* `{n,m}`*'s, and* `+` *'s.*
>
> *The book Mastering Regular Expressions by Jeffrey Friedl [8] gives a wonderful discussion of this and other efficiency issues.*

Eliminación de Comentarios de un Programa C

El siguiente ejemplo elimina los comentarios de un programa `C`.

```
casiano@millo:~/Lperltesting$ cat -n comments.pl
   1    #!/usr/bin/perl -w
   2    use strict;
```

```
  3
  4    my $progname = shift @ARGV or die "Usage:\n$0 prog.c\n";
  5    open(my $PROGRAM,"<$progname") || die "can't find $progname\n";
  6    my $program = '';
  7    {
  8      local $/ = undef;
  9      $program = <$PROGRAM>;
 10    }
 11    $program =~ s{
 12      /\*  # Match the opening delimiter
 13      .*?  # Match a minimal number of characters
 14      \*/  # Match the closing delimiter
 15    }[]gsx;
 16
 17    print $program;
```

Veamos un ejemplo de ejecución. Supongamos el fichero de entrada:

```
> cat hello.c
#include <stdio.h>
/* first
comment
*/
main() {
  printf("hello world!\n"); /* second comment */
}
```

Entonces la ejecución con ese fichero de entrada produce la salida:

```
> comments.pl hello.c
#include <stdio.h>

main() {
  printf("hello world!\n");
}
```

Veamos la diferencia de comportamiento entre `*` y `*?` en el ejemplo anterior:

```
pl@nereida:~/src/perl/perltesting$ perl5_10_1 -wde 0
main::(-e:1):   0
  DB<1>  use re 'debug'; 'main() /* 1c */ { /* 2c */ return; /* 3c */ }' =~ qr{(/\*.*\*/)}; pr
Compiling REx "(/\*.*\*/)"
Final program:
   1: OPEN1 (3)
   3:   EXACT </*> (5)
   5:   STAR (7)
   6:     REG_ANY (0)
   7:   EXACT <*/> (9)
   9: CLOSE1 (11)
  11: END (0)
anchored "/*" at 0 floating "*/" at 2..2147483647 (checking floating) minlen 4
Guessing start of match in sv for REx "(/\*.*\*/)" against "main() /* 1c */ { /* 2c */ return;
Found floating substr "*/" at offset 13...
Found anchored substr "/*" at offset 7...
Starting position does not contradict /^/m...
Guessed: match at offset 7
```

```
Matching REx "(/\*.*\*/)" against "/* 1c */ { /* 2c */ return; /* 3c */ }"
   7 <in() > </* 1c */ {>    |  1:OPEN1(3)
   7 <in() > </* 1c */ {>    |  3:EXACT </*>(5)
   9 <() /*> < 1c */ { />    |  5:STAR(7)
                                 REG_ANY can match 36 times out of 2147483647...
  41 <; /* 3c > <*/ }>       |  7:  EXACT <*/>(9)
  43 <; /* 3c */> < }>       |  9:  CLOSE1(11)
  43 <; /* 3c */> < }>       | 11:  END(0)
Match successful!

/* 1c */ { /* 2c */ return; /* 3c */
Freeing REx: "(/\*.*\*/)"

  DB<2>  use re 'debug'; 'main() /* 1c */ { /* 2c */ return; /* 3c */ }' =~ qr{(/\*.*?\*/)}; p
Compiling REx "(/\*.*?\*/)"
Final program:
   1: OPEN1 (3)
   3:   EXACT </*> (5)
   5:   MINMOD (6)
   6:   STAR (8)
   7:     REG_ANY (0)
   8:   EXACT <*/> (10)
  10: CLOSE1 (12)
  12: END (0)
anchored "/*" at 0 floating "*/" at 2..2147483647 (checking floating) minlen 4
Guessing start of match in sv for REx "(/\*.*?\*/)" against "main() /* 1c */ { /* 2c */ return
Found floating substr "*/" at offset 13...
Found anchored substr "/*" at offset 7...
Starting position does not contradict /^/m...
Guessed: match at offset 7
Matching REx "(/\*.*?\*/)" against "/* 1c */ { /* 2c */ return; /* 3c */ }"
   7 <in() > </* 1c */ {>    |  1:OPEN1(3)
   7 <in() > </* 1c */ {>    |  3:EXACT </*>(5)
   9 <() /*> < 1c */ { />    |  5:MINMOD(6)
   9 <() /*> < 1c */ { />    |  6:STAR(8)
                                 REG_ANY can match 4 times out of 4...
  13 <* 1c > <*/ { /* 2c>    |  8:  EXACT <*/>(10)
  15 <1c */> < { /* 2c *>    | 10:  CLOSE1(12)
  15 <1c */> < { /* 2c *>    | 12:  END(0)
Match successful!

/* 1c */
Freeing REx: "(/\*.*?\*/)"

  DB<3>
```

Véase también la documentación en la sección 'Matching-repetitions' en `perlretut` y la sección 'Quantifiers' en `perlre`.

Negaciones y operadores *lazy* A menudo las expresiones `X[^X]*X` y `X.*?X`, donde `X` es un carácter arbitrario se usan de forma casi equivalente.

- La primera significa:

  Una cadena que no contiene `X` *en su interior y que está delimitada por* `X`*s*

- La segunda significa:

  Una cadena que comienza en `X` *y termina en la* `X` *mas próxima a la* `X` *de comienzo*

Esta equivalencia se rompe si no se cumplen las hipótesis establecidas.

En el siguiente ejemplo se intentan detectar las cadenas entre comillas dobles que terminan en el signo de exclamación:

```
pl@nereida:~/Lperltesting$ cat -n negynogreedy.pl
     1  #!/usr/bin/perl -w
     2  use strict;
     3
     4  my $b = 'Ella dijo "Ana" y yo contesté: "Jamás!". Eso fué todo.';
     5  my $a;
     6  ($a = $b) =~ s/".*?!"/-$&-/;
     7  print "$a\n";
     8
     9  $b =~ s/"[^"]*!"/-$&-/;
    10  print "$b\n";
```

Al ejecutar el programa obtenemos:

```
> negynogreedy.pl
Ella dijo -"Ana" y yo contesté: "Jamás!"-. Eso fué todo.
Ella dijo "Ana" y yo contesté: -"Jamás!"-. Eso fué todo.
```

Copia y sustitución simultáneas El operador de *binding* `=~` nos permite "asociar" la variable con la operación de casamiento o sustitución. Si se trata de una sustitución y se quiere conservar la cadena, es necesario hacer una copia:

```
$d = $s;
$d =~ s/esto/por lo otro/;
```

en vez de eso, puedes abreviar un poco usando la siguiente "perla":

```
($d = $s) =~ s/esto/por lo otro/;
```

Obsérvese la asociación por la izquierda del operador de asignación.

Referencias a Paréntesis Previos

Las referencias relativas permiten escribir expresiones regulares mas reciclables. Véase la documentación en la sección 'Relative-backreferences' en `perlretut`:

> *Counting the opening parentheses to get the correct number for a backreference is error-prone as soon as there is more than one capturing group. A more convenient technique became available with Perl 5.10: relative backreferences. To refer to the immediately preceding capture group one now may write* `\g{-1}` *, the next but last is available via* `\g{-2}`*, and so on.*
>
> *Another good reason in addition to readability and maintainability for using relative backreferences is illustrated by the following example, where a simple pattern for matching peculiar strings is used:*
>
> ```
> 1. $a99a = '([a-z])(\d)\2\1'; # matches a11a, g22g, x33x, etc.
> ```
>
> *Now that we have this pattern stored as a handy string, we might feel tempted to use it as a part of some other pattern:*

```
1. $line = "code=e99e";
2. if ($line =~ /^(\w+)=$a99a$/){ # unexpected behavior!
3.   print "$1 is valid\n";
4. } else {
5.   print "bad line: '$line'\n";
6. }
```

But this doesn't match – at least not the way one might expect. Only after inserting the interpolated `$a99a` *and looking at the resulting full text of the regexp is it obvious that the backreferences have backfired – the subexpression* `(\w+)` *has snatched number 1 and demoted the groups in* `$a99a` *by one rank. This can be avoided by using relative backreferences:*

```
1. $a99a = '([a-z])(\d)\g{-1}\g{-2}'; # safe for being interpolated
```

El siguiente programa ilustra lo dicho:

```
casiano@millo:~/Lperltesting$ cat -n backreference.pl
   1    use strict;
   2    use re 'debug';
   3
   4    my $a99a = '([a-z])(\d)\2\1';
   5    my $line = "code=e99e";
   6    if ($line =~ /^(\w+)=$a99a$/){ # unexpected behavior!
   7      print "$1 is valid\n";
   8    } else {
   9      print "bad line: '$line'\n";
  10    }
```

Sigue la ejecución:

```
casiano@millo:~/Lperltesting$ perl5.10.1 -wd backreference.pl
main::(backreference.pl:4):     my $a99a = '([a-z])(\d)\2\1';
  DB<1> c 6
main::(backreference.pl:6):     if ($line =~ /^(\w+)=$a99a$/){ # unexpected behavior!
  DB<2> x ($line =~ /^(\w+)=$a99a$/)
  empty array
  DB<4> $a99a = '([a-z])(\d)\g{-1}\g{-2}'
  DB<5> x ($line =~ /^(\w+)=$a99a$/)
0  'code'
1  'e'
2  9
```

Usando Referencias con Nombre (Perl 5.10)

El siguiente texto esta tomado de la sección 'Named-backreferences' en `perlretut`:

Perl 5.10 also introduced named capture buffers and named backreferences. To attach a name to a capturing group, you write either `(?<name>...)` *or* `(?'name'...)`*. The backreference may then be written as* `\g{name}` *.*

It is permissible to attach the same name to more than one group, but then only the leftmost one of the eponymous set can be referenced. Outside of the pattern a named capture buffer is accessible through the `%+` *hash.*

Assuming that we have to match calendar dates which may be given in one of the three formats `yyyy-mm-dd`*,* `mm/dd/yyyy` *or* `dd.mm.yyyy`*, we can write three suitable patterns where we use* `'d'`*,* `'m'` *and* `'y'` *respectively as the names of the buffers capturing the pertaining components of a date. The matching operation combines the three patterns as alternatives:*

```
1.  $fmt1 = '(?<y>\d\d\d\d)-(?<m>\d\d)-(?<d>\d\d)';
2.  $fmt2 = '(?<m>\d\d)/(?<d>\d\d)/(?<y>\d\d\d\d)';
3.  $fmt3 = '(?<d>\d\d)\.(?<m>\d\d)\.(?<y>\d\d\d\d)';
4.  for my $d qw( 2006-10-21 15.01.2007 10/31/2005 ){
5.    if ( $d =~ m{$fmt1|$fmt2|$fmt3} ){
6.      print "day=$+{d} month=$+{m} year=$+{y}\n";
7.    }
8.  }
```

If any of the alternatives matches, the hash `%+` *is bound to contain the three key-value pairs.*

En efecto, al ejecutar el programa:

```
casiano@millo:~/Lperltesting$ cat -n namedbackreferences.pl
     1  use v5.10;
     2  use strict;
     3
     4  my $fmt1 = '(?<y>\d\d\d\d)-(?<m>\d\d)-(?<d>\d\d)';
     5  my $fmt2 = '(?<m>\d\d)/(?<d>\d\d)/(?<y>\d\d\d\d)';
     6  my $fmt3 = '(?<d>\d\d)\.(?<m>\d\d)\.(?<y>\d\d\d\d)';
     7
     8  for my $d qw( 2006-10-21 15.01.2007 10/31/2005 ){
     9    if ( $d =~ m{$fmt1|$fmt2|$fmt3} ){
    10      print "day=$+{d} month=$+{m} year=$+{y}\n";
    11    }
    12  }
```

Obtenemos la salida:

```
casiano@millo:~/Lperltesting$ perl5.10.1 -w namedbackreferences.pl
day=21 month=10 year=2006
day=15 month=01 year=2007
day=31 month=10 year=2005
```

Como se comentó:

... It is permissible to attach the same name to more than one group, but then only the leftmost one of the eponymous set can be referenced.

Veamos un ejemplo:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
DB<1>  # ... only the leftmost one of the eponymous set can be referenced
DB<2> $r = qr{(?<a>[a-c])(?<a>[a-f])}
DB<3> print $+{a} if 'ad' =~ $r
a
DB<4> print $+{a} if 'cf' =~ $r
c
DB<5> print $+{a} if 'ak' =~ $r
```

Reescribamos el ejemplo de conversión de temperaturas usando paréntesis con nombre:

```
pl@nereida:~/Lperltesting$ cat -n c2f_5_10v2.pl
 1  #!/usr/local/bin/perl5_10_1 -w
 2  use strict;
 3
 4  print "Enter a temperature (i.e. 32F, 100C):\n";
 5  my $input = <STDIN>;
 6  chomp($input);
 7
 8  $input =~ m/^
 9              (?<farenheit>[-+]?[0-9]+(?:\.[0-9]*)?)\s*[fF]
10              |
11              (?<celsius>[-+]?[0-9]+(?:\.[0-9]*)?)\s*[cC]
12           $/x;
13
14  my ($celsius, $farenheit);
15  if (exists $+{celsius}) {
16    $celsius = $+{celsius};
17    $farenheit = ($celsius * 9/5)+32;
18  }
19  elsif (exists $+{farenheit}) {
20    $farenheit = $+{farenheit};
21    $celsius = ($farenheit -32)*5/9;
22  }
23  else {
24    die "Expecting a temperature, so don't understand \"$input\".\n";
25  }
26
27  printf "%.2f C = %.2f F\n", $celsius, $farenheit;
```

La función `exists` retorna verdadero si existe la clave en el hash y falso en otro caso.

Grupos con Nombre y Factorización

El uso de nombres hace mas robustas y mas factorizables las expresiones regulares. Consideremos la siguiente regexp que usa notación posicional:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wde 0
main::(-e:1):   0
  DB<1> x "abbacddc" =~ /(.)(.)\2\1/
0  'a'
1  'b'
```

Supongamos que queremos reutilizar la regexp con repetición

```
  DB<2> x "abbacddc" =~ /((.)(.)\2\1){2}/
  empty array
```

¿Que ha ocurrido? La introducción del nuevo paréntesis nos obliga a renombrar las referencias a las posiciones:

```
  DB<3> x "abbacddc" =~ /((.)(.)\3\2){2}/
0  'cddc'
1  'c'
2  'd'
  DB<4> "abbacddc" =~ /((.)(.)\3\2){2}/; print "$&\n"
abbacddc
```

Esto no ocurre si utilizamos nombres. El operador `\k<a>` sirve para hacer referencia al valor que ha casado con el paréntesis con nombre `a`:

```
  DB<5> x "abbacddc" =~ /((?<a>.)(?<b>.)\k<b>\k<a>){2}/
0  'cddc'
1  'c'
2  'd'
```

El uso de grupos con nombre y `\k`[1] en lugar de referencias numéricas absolutas hace que la regexp sea mas reutilizable.

LLamadas a expresiones regulares via paréntesis con memoria

Es posible también llamar a la expresión regular asociada con un paréntesis.
Este parrafo tomado de la sección 'Extended-Patterns' en `perlre` explica el modo de uso:
`(?PARNO) (?-PARNO) (?+PARNO) (?R) (?0)`

> `PARNO` *is a sequence of digits (not starting with 0) whose value reflects the paren-number of the capture buffer to recurse to.*
>
> *....*
>
> *Capture buffers contained by the pattern will have the value as determined by the outermost recursion. ....*
>
> *If* `PARNO` **is preceded by a plus or minus sign then it is assumed to be relative, with negative numbers indicating preceding capture buffers and positive ones following**. *Thus* `(?-1)` *refers to the most recently declared buffer, and* `(?+1)` *indicates the next buffer to be declared.*
>
> **Note that the counting for relative recursion differs from that of relative backreferences, in that with recursion unclosed buffers are included.**

Veamos un ejemplo:

```
casiano@millo:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> x "AABB" =~ /(A)(?-1)(?+1)(B)/
0  'A'
1  'B'
  # Parenthesis:       1   2   2                   1
  DB<2> x 'ababa' =~ /^((?:([ab])(?1)\g{-1}|[ab]?))$/
0  'ababa'
1  'a'
  DB<3> x 'bbabababb' =~ /^((?:([ab])(?1)\g{-1}|[ab]?))$/
0  'bbabababb'
1  'b'
```

Véase también:

- Perl Training Australia: Regular expressions in Perl 5.10
- Perl 5.10 Advanced Regular Expressions by Yves Orton
- Gabor: Regular Expressions in Perl 5.10

[1] Una diferencia entre $\backslash k$ y $\backslash g$ es que el primero sólo admite un nombre como argumento mientras que $\backslash g$ admite enteros

Reutilizando Expresiones Regulares

La siguiente reescritura de nuestro ejemplo básico utiliza el módulo `Regexp::Common` para factorizar la expresión regular:

```
casiano@millo:~/src/perl/perltesting$ cat -n c2f_5_10v3.pl
 1  #!/soft/perl5lib/bin/perl5.10.1 -w
 2  use strict;
 3  use Regexp::Common;
 4
 5  print "Enter a temperature (i.e. 32F, 100C):\n";
 6  my $input = <STDIN>;
 7  chomp($input);
 8
 9  $input =~ m/^
10              (?<farenheit>$RE{num}{real})\s*[fF]
11              |
12              (?<celsius>$RE{num}{real})\s*[cC]
13           $/x;
14
15  my ($celsius, $farenheit);
16  if ('celsius' ~~ %+) {
17    $celsius = $+{celsius};
18    $farenheit = ($celsius * 9/5)+32;
19  }
20  elsif ('farenheit' ~~ %+) {
21    $farenheit = $+{farenheit};
22    $celsius = ($farenheit -32)*5/9;
23  }
24  else {
25    die "Expecting a temperature, so don't understand \"$input\".\n";
26  }
27
28  printf "%.2f C = %.2f F\n", $celsius, $farenheit;
```

Véase:

- La documentación del módulo `Regexp::Common` por Abigail
- Smart Matching: Perl Training Australia: Smart Match
- Rafael García Suárez: la sección 'Smart-matching-in-detail' en `perlsyn`
- Enrique Nell (Barcelona Perl Mongers): Novedades en Perl 5.10

El Módulo Regexp::Common

El módulo `Regexp::Common` provee un extenso número de expresiones regulares que son accesibles vía el hash `%RE`. sigue un ejemplo de uso:

```
casiano@millo:~/Lperltesting$ cat -n regexpcommonsynopsis.pl
     1  use strict;
     2  use Perl6::Say;
     3  use Regexp::Common;
     4
     5  while (<>) {
     6      say q{a number}              if /$RE{num}{real}/;
     7
```

```
 8      say q{a ['"`] quoted string} if /$RE{quoted}/;
 9
10      say q{a /.../ sequence}      if m{$RE{delimited}{'-delim'=>'/'}};
11
12      say q{balanced parentheses}  if /$RE{balanced}{'-parens'=>'()'}/;
13
14      die q{a #*@%-ing word}."\n"  if /$RE{profanity}/;
15
16  }
17
```

Sigue un ejemplo de ejecución:

```
casiano@millo:~/Lperltesting$ perl regexpcommonsynopsis.pl
43
a number
"2+2 es" 4
a number
a ['"`] quoted string
x/y/z
a /.../ sequence
(2*(4+5/(3-2)))
a number
balanced parentheses
fuck you!
a #*@%-ing word
```

El siguiente fragmento de la documentación de `Regexp::Common` explica el modo simplificado de uso:

> *To access a particular pattern,* `%RE` *is treated as a hierarchical hash of hashes (of hashes...), with each successive key being an identifier. For example, to access the pattern that matches real numbers, you specify:*
>
> ```
> $RE{num}{real}
> ```
>
> *and to access the pattern that matches integers:*
>
> ```
> $RE{num}{int}
> ```
>
> *Deeper layers of the hash are used to specify flags: arguments that modify the resulting pattern in some way.*
>
> - *The keys used to access these layers are prefixed with a minus sign and may have a value;*
> - *if a value is given, it's done by using a multidimensional key.*
>
> *For example, to access the pattern that matches base-2 real numbers with embedded commas separating groups of three digits (e.g.* `10,101,110.110101101`*):*
>
> ```
> $RE{num}{real}{-base => 2}{-sep => ','}{-group => 3}
> ```
>
> *Through the magic of Perl, these flag layers may be specified in any order (and even interspersed through the identifier keys!) so you could get the same pattern with:*

```
$RE{num}{real}{-sep => ','}{-group => 3}{-base => 2}
```

or:

```
$RE{num}{-base => 2}{real}{-group => 3}{-sep => ','}
```

or even:

```
$RE{-base => 2}{-group => 3}{-sep => ','}{num}{real}
```

etc.

Note, however, that the relative order of amongst the identifier keys is significant. That is:

```
$RE{list}{set}
```

would not be the same as:

```
$RE{set}{list}
```

Veamos un ejemplo con el depurador:

```
casiano@millo:~/Lperltesting$ perl -MRegexp::Common -wde 0
main::(-e:1):   0
  DB<1> x 'numero: 10,101,110.110101101 101.1e-1 234' =~ m{($RE{num}{real}{-base => 2}{-sep =>
0  '10,101,110.110101101'
1  '101.1e-1'
```

La expresión regular para un número real es relativamente compleja:

```
casiano@millo:~/src/perl/perltesting$ perl5.10.1 -wd c2f_5_10v3.pl
main::(c2f_5_10v3.pl:5):     print "Enter a temperature (i.e. 32F, 100C):\n";
  DB<1> p $RE{num}{real}
(?:(?i)(?:[+-]?)(?:(?=[0123456789]|[.])(?:[0123456789]*)(?:(?:[.])(?:[0123456789]{0,}))?)(?:(?
```

Si se usa la opción `-keep` el patrón proveído usa paréntesis con memoria:

```
casiano@millo:~/Lperltesting$ perl -MRegexp::Common -wde 0
main::(-e:1):   0
DB<2> x 'one, two, three, four, five' =~ /$RE{list}{-pat => '\w+'}/
0  1
DB<3> x 'one, two, three, four, five' =~ /$RE{list}{-pat => '\w+'}{-keep}/
0  'one, two, three, four, five'
1  ', '
```

Smart Matching

Perl 5.10 introduce el operador de smart matching. El siguiente texto es tomado casi verbatim del site de la compañía Perl Training Australia[2]:

[2] This Perl tip and associated text is copyright Perl Training Australia

Perl 5.10 introduces a new-operator, called smart-match, written ~~. As the name suggests, smart-match tries to compare its arguments in an intelligent fashion. Using smart-match effectively allows many complex operations to be reduces to very simple statements.

Unlike many of the other features introduced in Perl 5.10, there's no need to use the feature pragma to enable smart-match, as long as you're using 5.10 it's available.

The smart-match operator is always commutative. That means that `$x ~~ $y` *works the same way as* `$y ~~ $x`*. You'll never have to remember which order to place to your operands with smart-match. Smart-match in action.*

As a simple introduction, we can use smart-match to do a simple string comparison between simple scalars. For example:

```
use feature qw(say);

my $x = "foo";
my $y = "bar";
my $z = "foo";

say '$x and $y are identical strings' if $x ~~ $y;
say '$x and $z are identical strings' if $x ~~ $z;    # Printed
```

If one of our arguments is a number, then a numeric comparison is performed:

```
my $num   = 100;
my $input = <STDIN>;

say 'You entered 100' if $num ~~ $input;
```

This will print our message if our user enters 100, 100.00, +100, 1e2, or any other string that looks like the number 100.

We can also smart-match against a regexp:

```
my $input  = <STDIN>;

say 'You said the secret word!' if $input ~~ /xyzzy/;
```

Smart-matching with a regexp also works with saved regexps created with qr.

So we can use smart-match to act like eq, `==` *and* `=~`*, so what? Well, it does much more than that.*

We can use smart-match to search a list:

```
casiano@millo:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> @friends = qw(Frodo Meriadoc Pippin Samwise Gandalf)
  DB<2> print "You're a friend" if 'Pippin' ~~ @friends
You're a friend
  DB<3> print "You're a friend" if 'Mordok' ~~ @friends
```

It's important to note that searching an array with smart-match is extremely fast. It's faster than using grep, it's faster than using `first` *from* `Scalar::Util`*, and it's faster than walking through the loop with* `foreach`*, even if you do know all the clever optimisations.*

Esta es la forma típica de buscar un elemento en un array en versiones anteriores a la 5.10:

```
casiano@millo:~$ perl -wde 0
main::(-e:1):   0
  DB<1> use List::Util qw{first}
  DB<2> @friends = qw(Frodo Meriadoc Pippin Samwise Gandalf)
  DB<3> x first { $_ eq 'Pippin'} @friends
0  'Pippin'
  DB<4> x first { $_ eq 'Mordok'} @friends
0  undef
```

We can also use smart-match to compare arrays:

```
  DB<4> @foo = qw(x y z xyzzy ninja)
  DB<5> @bar = qw(x y z xyzzy ninja)
  DB<7> print "Identical arrays" if @foo ~~ @bar
Identical arrays
  DB<8> @bar = qw(x y z xyzzy nOnjA)
  DB<9> print "Identical arrays" if @foo ~~ @bar
  DB<10>
```

And even search inside an array using a string:

```
 DB<11> x @foo = qw(x y z xyzzy ninja)
0  'x'
1  'y'
2  'z'
3  'xyzzy'
4  'ninja'
  DB<12> print "Array contains a ninja " if @foo ~~ 'ninja'
```

or using a regexp:

```
  DB<13> print "Array contains magic pattern" if @foo ~~ /xyz/
Array contains magic pattern
  DB<14> print "Array contains magic pattern" if @foo ~~ /\d+/
```

Smart-match works with array references, too[3]*:*

```
  DB<16> $array_ref = [ 1..10 ]
  DB<17> print "Array contains 10" if 10 ~~ $array_ref
Array contains 10
  DB<18> print "Array contains 10" if $array_ref ~~ 10
  DB<19>
```

En el caso de un número y un array devuelve cierto si el escalar aparece en un array anidado:

```
casiano@millo:~/Lperltesting$ perl5.10.1 -E 'say "ok" if 42 ~~  [23, 17, [40..50], 70];'
ok
casiano@millo:~/Lperltesting$ perl5.10.1 -E 'say "ok" if 42 ~~  [23, 17, [50..60], 70];'
casiano@millo:~/Lperltesting$
```

Of course, we can use smart-match with more than just arrays and scalars, it works with searching for the key in a hash, too!

[3] En este caso la conmutatividad no funciona

```
  DB<19> %colour = ( sky   => 'blue', grass => 'green', apple => 'red',)
  DB<20> print "I know the colour" if 'grass' ~~ %colour
I know the colour
  DB<21> print "I know the colour" if 'cloud' ~~ %colour
  DB<22>
  DB<23> print "A key starts with 'gr'" if %colour ~~ /^gr/
A key starts with 'gr'
  DB<24> print "A key starts with 'clou'" if %colour ~~ /^clou/
  DB<25>
```

You can even use it to see if the two hashes have identical keys:

```
  DB<26> print 'Hashes have identical keys' if %taste ~~ %colour;
Hashes have identical keys
```

La conducta del operador de smart matching viene dada por la siguiente tabla tomada de la sección 'Smart-matching-in-detail' en `perlsyn`:

> *The behaviour of a smart match depends on what type of thing its arguments are. The behaviour is determined by the following table: the first row that applies determines the match behaviour (which is thus mostly determined by the type of the right operand). Note that the smart match implicitly dereferences any non-blessed hash or array ref, so the "Hash.*[a]*nd .*[A]*rray.*[en]*tries apply in those cases. (For blessed references, the .*[O]*bject.*[en]*tries apply.)*
>
> *Note that the "Matching Codeçolumn is not always an exact rendition. For example, the smart match operator short-circuits whenever possible, but grep does not.*

```
$a      $b        Type of Match Implied    Matching Code
======  =====     =====================    =============
Any     undef     undefined                !defined $a

Any     Object    invokes ~~ overloading on $object, or dies

Hash    CodeRef   sub truth for each key[1] !grep { !$b->($_) } keys %$a
Array   CodeRef   sub truth for each elt[1] !grep { !$b->($_) } @$a
Any     CodeRef   scalar sub truth          $b->($a)

Hash    Hash      hash keys identical (every key is found in both hashes)
Array   Hash      hash slice existence     grep { exists $b->{$_} } @$a
Regex   Hash      hash key grep            grep /$a/, keys %$b
undef   Hash      always false (undef can't be a key)
Any     Hash      hash entry existence     exists $b->{$a}

Hash    Array     hash slice existence     grep { exists $a->{$_} } @$b
Array   Array     arrays are comparable[2]
Regex   Array     array grep               grep /$a/, @$b
undef   Array     array contains undef     grep !defined, @$b
Any     Array     match against an array element[3]
                                           grep $a ~~ $_, @$b

Hash    Regex     hash key grep            grep /$b/, keys %$a
Array   Regex     array grep               grep /$b/, @$a
Any     Regex     pattern match            $a =~ /$b/
```

```
Object  Any       invokes ~~ overloading on $object, or falls back:
Any     Num       numeric equality         $a == $b
Num     numish[4] numeric equality         $a == $b
undef   Any       undefined                !defined($b)
Any     Any       string equality          $a eq $b
```

Ejercicios

Ejercicio 31.1.4. ▪ *Indique la salida del siguiente programa:*

```
 1  pl@nereida:~/Lperltesting$ cat twonumbers.pl
 2  $_ = "I have 2 numbers: 53147";
 3  @pats = qw{
 4    (.*)(\d*)
 5    (.*)(\d+)
 6    (.*?)(\d*)
 7    (.*?)(\d+)
 8    (.*)(\d+)$
 9    (.*?)(\d+)$
10    (.*)\b(\d+)$
11    (.*\D)(\d+)$
12  };
13
14  print "$_\n";
15  for $pat (@pats) {
16    printf "%-12s ", $pat;
17    <>;
18    if ( /$pat/ ) {
19      print "<$1> <$2>\n";
20    } else {
21      print "FAIL\n";
22    }
23  }
```

31.1.2. Depuración de Expresiones Regulares

Para obtener información sobre la forma en que es compilada una expresión regular y como se produce el proceso de matching podemos usar la opción `'debug'` del módulo `re`. La versión de Perl 5.10 da una información algo mas legible que la de las versiones anteriores:

```
pl@nereida:~/Lperltesting$ perl5_10_1 -wde 0

Loading DB routines from perl5db.pl version 1.32
Editor support available.

Enter h or ‘h h’ for help, or ‘man perldebug’ for more help.

main::(-e:1):   0
  DB<1> use re 'debug'; 'astr' =~ m{[sf].r}
Compiling REx "[sf].r"
Final program:
   1: ANYOF[fs][] (12)
  12: REG_ANY (13)
  13: EXACT <r> (15)
```

```
  15: END (0)
anchored "r" at 2 (checking anchored) stclass ANYOF[fs][] minlen 3
Guessing start of match in sv for REx "[sf].r" against "astr"
Found anchored substr "r" at offset 3...
Starting position does not contradict /^/m...
start_shift: 2 check_at: 3 s: 1 endpos: 2
Does not contradict STCLASS...
Guessed: match at offset 1
Matching REx "[sf].r" against "str"
   1 <a> <str>               |  1:ANYOF[fs][](12)
   2 <as> <tr>               | 12:REG_ANY(13)
   3 <ast> <r>               | 13:EXACT <r>(15)
   4 <astr> <>               | 15:END(0)
Match successful!
Freeing REx: "[sf].r"
```

Si se usa la opción `debug` de `re` con objetos expresión regular, se obtendrá información durante el proceso de matching:

```
 DB<3> use re 'debug'; $re = qr{[sf].r}
Compiling REx "[sf].r"
Final program:
   1: ANYOF[fs][] (12)
  12: REG_ANY (13)
  13: EXACT <r> (15)
  15: END (0)
anchored "r" at 2 (checking anchored) stclass ANYOF[fs][] minlen 3

  DB<4> 'astr' =~ $re
Guessing start of match in sv for REx "[sf].r" against "astr"
Found anchored substr "r" at offset 3...
Starting position does not contradict /^/m...
start_shift: 2 check_at: 3 s: 1 endpos: 2
Does not contradict STCLASS...
Guessed: match at offset 1
Matching REx "[sf].r" against "str"
   1 <a> <str>               |  1:ANYOF[fs][](12)
   2 <as> <tr>               | 12:REG_ANY(13)
   3 <ast> <r>               | 13:EXACT <r>(15)
   4 <astr> <>               | 15:END(0)
Match successful!
```

31.1.3. Tablas de Escapes, Metacarácteres, Cuantificadores, Clases

Sigue una sección de tablas con notaciones tomada de `perlre`:

Metacharacters

The following metacharacters have their standard egrep-ish meanings:

```
1. \ Quote the next metacharacter
2. ^ Match the beginning of the line
3. . Match any character (except newline)
4. $ Match the end of the line (or before newline at the end)
5. | Alternation
```

6. () Grouping
7. [] Character class

Standard greedy quantifiers

The following standard greedy quantifiers are recognized:

1. * Match 0 or more times
2. + Match 1 or more times
3. ? Match 1 or 0 times
4. {n} Match exactly n times
5. {n,} Match at least n times
6. {n,m} Match at least n but not more than m times

Non greedy quantifiers

The following non greedy quantifiers are recognized:

1. *? Match 0 or more times, not greedily
2. +? Match 1 or more times, not greedily
3. ?? Match 0 or 1 time, not greedily
4. {n}? Match exactly n times, not greedily
5. {n,}? Match at least n times, not greedily
6. {n,m}? Match at least n but not more than m times, not greedily

Possesive quantifiers

The following possesive quantifiers are recognized:

1. *+ Match 0 or more times and give nothing back
2. ++ Match 1 or more times and give nothing back
3. ?+ Match 0 or 1 time and give nothing back
4. {n}+ Match exactly n times and give nothing back (redundant)
5. {n,}+ Match at least n times and give nothing back
6. {n,m}+ Match at least n but not more than m times and give nothing back

Escape sequences

1. \t tab (HT, TAB)
2. \n newline (LF, NL)
3. \r return (CR)
4. \f form feed (FF)
5. \a alarm (bell) (BEL)
6. \e escape (think troff) (ESC)
7. \033 octal char (example: ESC)
8. \x1B hex char (example: ESC)
9. \x{263a} long hex char (example: Unicode SMILEY)
10. \cK control char (example: VT)
11. \N{name} named Unicode character
12. \l lowercase next char (think vi)
13. \u uppercase next char (think vi)
14. \L lowercase till \E (think vi)
15. \U uppercase till \E (think vi)
16. \E end case modification (think vi)
17. \Q quote (disable) pattern metacharacters till \E

Ejercicio 31.1.5. *Explique la salida:*

```
casiano@tonga:~$ perl -wde 0
main::(-e:1):   0
  DB<1> $x = '([a-z]+)'
  DB<2> x 'hola' =~ /$x/
0  'hola'
  DB<3> x 'hola' =~ /\Q$x/
  empty array
  DB<4> x '([a-z]+)' =~ /\Q$x/
0  1
```

Character Classes and other Special Escapes

```
 1. \w Match a "word" character (alphanumeric plus "_")
 2. \W Match a non-"word" character
 3. \s Match a whitespace character
 4. \S Match a non-whitespace character
 5. \d Match a digit character
 6. \D Match a non-digit character
 7. \pP Match P, named property. Use \p{Prop} for longer names.
 8. \PP Match non-P
 9. \X Match eXtended Unicode "combining character sequence",
10.    equivalent to (?>\PM\pM*)
11. \C Match a single C char (octet) even under Unicode.
12.    NOTE: breaks up characters into their UTF-8 bytes,
13.    so you may end up with malformed pieces of UTF-8.
14.    Unsupported in lookbehind.
15. \1 Backreference to a specific group.
16. '1' may actually be any positive integer.
17. \g1 Backreference to a specific or previous group,
18. \g{-1} number may be negative indicating a previous buffer and may
19.        optionally be wrapped in curly brackets for safer parsing.
20. \g{name} Named backreference
21. \k<name> Named backreference
22. \K Keep the stuff left of the \K, don't include it in $&
23. \v Vertical whitespace
24. \V Not vertical whitespace
25. \h Horizontal whitespace
26. \H Not horizontal whitespace
27. \R Linebreak
```

Zero width assertions

Perl defines the following zero-width assertions:

```
 1. \b Match a word boundary
 2. \B Match except at a word boundary
 3. \A Match only at beginning of string
 4. \Z Match only at end of string, or before newline at the end
 5. \z Match only at end of string
 6. \G Match only at pos() (e.g. at the end-of-match position
 7. of prior m//g)
```

The POSIX character class syntax

The POSIX character class syntax:

```
1. [:class:]
```

is also available. Note that the `[` and `]` brackets are literal; they must always be used within a character class expression.

```
1. # this is correct:
2. $string =~ /[[:alpha:]]/;
3.
4. # this is not, and will generate a warning:
5. $string =~ /[:alpha:]/;
```

Available classes

The available classes and their backslash equivalents (if available) are as follows:

```
1. alpha
2. alnum
3. ascii
4. blank
5. cntrl
6. digit \d
7. graph
8. lower
9. print
10. punct
11. space \s
12. upper
13. word \w
14. xdigit
```

For example use `[:upper:]` to match all the uppercase characters. Note that the `[]` are part of the `[::]` construct, not part of the whole character class. For example:

```
1. [01[:alpha:]%]
```

matches zero, one, any alphabetic character, and the percent sign.

Equivalences to Unicode

The following equivalences to Unicode `\p{}` constructs and equivalent backslash character classes (if available), will hold:

```
1. [[:...:]] \p{...} backslash
2.
3. alpha IsAlpha
4. alnum IsAlnum
5. ascii IsASCII
6. blank
7. cntrl IsCntrl
8. digit IsDigit \d
9. graph IsGraph
10. lower IsLower
11. print IsPrint
12. punct IsPunct
13. space IsSpace
14. IsSpacePerl \s
15. upper IsUpper
16. word IsWord \w
17. xdigit IsXDigit
```

Negated character classes

You can negate the `[::]` character classes by prefixing the class name with a `'^'`. This is a Perl extension. For example:

```
1. POSIX traditional Unicode
2.
3. [[:^digit:]] \D \P{IsDigit}
4. [[:^space:]] \S \P{IsSpace}
5. [[:^word:]] \W \P{IsWord}
```

31.1.4. Variables especiales después de un emparejamiento

Despues de un emparejamiento con éxito, las siguientes variables especiales quedan definidas:

`$&`	El texto que casó
`$‘`	El texto que está a la izquierda de lo que casó
`$’`	El texto que está a la derecha de lo que casó
`$1, $2, $3`, etc.	Los textos capturados por los paréntesis
`$+`	Una copia del `$1, $2, ...` con número mas alto
`@-`	Desplazamientos de las subcadenas que casan en `$1 ...`
`@+`	Desplazamientos de los finales de las subcadenas en `$1 ...`
`$#-`	El índice del último paréntesis que casó
`$#+`	El índice del último paréntesis en la última expresión regular

Las Variables de match, pre-match y post-mach

Ejemplo:

```
1 #!/usr/bin/perl -w
2 if ("Hello there, neighbor" =~ /\s(\w+),/) {
3   print "That was: ($‘)($&)($’).\n",
4 }
```

```
> matchvariables.pl
That was: (Hello)( there,)( neighbor).
```

El uso de estas variables tenía un efecto negativo en el rendimiento de la regexp. Véase `perlfaq6` la sección `Why does using $&, $‘, or $’ slow my program down?`.

> *Once Perl sees that you need one of these variables anywhere in the program, it provides them on each and every pattern match. That means that on every pattern match the entire string will be copied, part of it to* `$‘`*, part to* `$&`*, and part to* `$’`*. Thus the penalty is most severe with long strings and patterns that match often. Avoid* `$&`*,* `$’`*, and* `$‘` *if you can, but if you can’t, once you’ve used them at all, use them at will because you’ve already paid the price. Remember that some algorithms really appreciate them. As of the 5.005 release, the* `$&` *variable is no longer .*^{ex}*pensive”the way the other two are.*
>
> *Since Perl 5.6.1 the special variables* `@-` *and* `@+` *can functionally replace* `$‘`*,* `$&` *and* `$’`*. These arrays contain pointers to the beginning and end of each match (see* `perlvar` *for the full story), so they give you essentially the same information, but without the risk of excessive string copying.*
>
> *Perl 5.10 added three specials,* `${^MATCH}`*,* `${^PREMATCH}`*, and* `${^POSTMATCH}` *to do the same job but without the global performance penalty. Perl 5.10 only sets these variables if you compile or execute the regular expression with the* `/p` *modifier.*

```
pl@nereida:~/Lperltesting$ cat ampersandoldway.pl
#!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
use strict;
use Benchmark qw(cmpthese timethese);

'hola juan' =~ /ju/;
my ($a, $b, $c) = ($`, $&, $');


cmpthese( -1, {
    oldway => sub { 'hola juan' =~ /ju/  },
});
pl@nereida:~/Lperltesting$ cat ampersandnewway.pl
#!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
use strict;
use Benchmark qw(cmpthese timethese);

'hola juan' =~ /ju/p;
my ($a, $b, $c) = (${^PREMATCH}, ${^MATCH}, ${^POSTMATCH});


cmpthese( -1, {
    newway => sub { 'hola juan' =~ /ju/  },
});

pl@nereida:~/Lperltesting$ time ./ampersandoldway.pl
            Rate oldway
oldway 2991861/s     --

real    0m3.761s
user    0m3.740s
sys     0m0.020s
pl@nereida:~/Lperltesting$ time ./ampersandnewway.pl
            Rate newway
newway 8191999/s     --

real    0m6.721s
user    0m6.704s
sys     0m0.016s
```

Véase

- `perlvar` (busque por `$MATCH`)

Texto Asociado con el Último Paréntesis

La variable `$+` contiene el texto que casó con el último paréntesis en el patrón. Esto es útil en situaciones en las cuáles una de un conjunto de alternativas casa, pero no sabemos cuál:

```
  DB<9> "Revision: 4.5" =~ /Version: (.*)|Revision: (.*)/ && ($rev = $+);
  DB<10> x $rev
0  4.5
  DB<11> "Version: 4.5" =~ /Version: (.*)|Revision: (.*)/ && ($rev = $+);
  DB<12> x $rev
0  4.5
```

Los Offsets de los Inicios de los Casamientos: `@-`

El vector `@-` contiene los *offsets* o desplazamientos de los casamientos en la última expresión regular. La entrada `$-[0]` es el desplazamiento del último casamiento con éxito y `$-[n]` es el desplazamiento de la subcadena que casa con el `n`-ésimo paréntesis (o `undef` si el párentesis no casó). Por ejemplo:

```
#           012345678
DB<1> $z = "hola13.47"
DB<2> if ($z =~ m{a(\d+)(\.(\d+))?}) { print "@-\n"; }
3 4 6 7
```

El resultado se interpreta como sigue:

- 3 = desplazamiento de comienzo de `$& = a13.47`
- 4 = desplazamiento de comienzo de `$1 = 13`
- 6 = desplazamiento de comienzo de `$2 = .`
- 7 = desplazamiento de comienzo de `$3 = 47`

Esto es lo que dice `perlvar` sobre `@-`:

> *This array holds the offsets of the beginnings of the last successful submatches in the currently active dynamic scope.* `$-[0]` *is the offset into the string of the beginning of the entire match. The nth element of this array holds the offset of the nth submatch, so* `$-[1]` *is the offset where* `$1` *begins,* `$-[2]` *the offset where* `$2` *begins, and so on.*
>
> *After a match against some variable* `$var`*:*
>
> ```
> $` is the same as substr($var, 0, $-[0])
> $& is the same as substr($var, $-[0], $+[0] - $-[0])
> $' is the same as substr($var, $+[0])
> $1 is the same as substr($var, $-[1], $+[1] - $-[1])
> $2 is the same as substr($var, $-[2], $+[2] - $-[2])
> $3 is the same as substr($var, $-[3], $+[3] - $-[3])
> ```

Desplazamientos de los Finales de los Emparejamientos: `@+`

El array `@+` contiene los desplazamientos de los finales de los emparejamientos. La entrada `$+[0]` contiene el desplazamiento del final de la cadena del emparejamiento completo. Siguiendo con el ejemplo anterior:

```
#            0123456789
DB<17> $z = "hola13.47x"
DB<18> if ($z =~ m{a(\d+)(\.)(\d+)?}) { print "@+\n"; }
9 6 7 9
```

El resultado se interpreta como sigue:

- 9 = desplazamiento final de `$& = a13.47x`
- 6 = desplazamiento final de `$1 = 13`
- 7 = desplazamiento final de `$2 = .`
- 9 = desplazamiento final de `$3 = 47`

Número de paréntesis en la última regexp con éxito

Se puede usar `$#+` para determinar cuantos parentesis había en el último emparejamiento que tuvo éxito.

```
  DB<29> $z = "h"
  DB<30> print "$#+\n" if ($z =~ m{(a)(b)}) || ($z =~ m{(h)(.)?(.)?})
3
  DB<31> $z = "ab"
  DB<32> print "$#+\n" if ($z =~ m{(a)(b)}) || ($z =~ m{(h)(.)?(.)?})
2
```

Indice del Ultimo Paréntesis

La variable `$#-` contiene el índice del último paréntesis que casó. Observe la siguiente ejecución con el depurador:

```
  DB<1> $x = '13.47'; $y = '125'
  DB<2> if ($y =~ m{(\d+)(\.(\d+))?}) { print "last par = $#-, content = $+\n"; }
last par = 1, content = 125
  DB<3> if ($x =~ m{(\d+)(\.(\d+))?}) { print "last par = $#-, content = $+\n"; }
last par = 3, content = 47
```

`@-` y `@+` no tienen que tener el mismo tamaño

En general no puede asumirse que `@-` y `@+` sean del mismo tamaño.

```
  DB<1> "a" =~ /(a)|(b)/; @a = @-; @b = @+
  DB<2> x @a
0  0
1  0
  DB<3> x @b
0  1
1  1
2  undef
```

Véase También

Para saber más sobre las variables especiales disponibles consulte

- `perldoc perlretut`
- `perldoc perlvar`.

31.1.5. Ambito Automático

Como sabemos, ciertas variables (como `$1`, `$&` ...) reciben automáticamente un valor con cada operación de "matching".

Considere el siguiente código:

```
if (m/(...)/) {
  &do_something();
  print "the matched variable was $1.\n";
}
```

Puesto que `$1` es automáticamente declarada `local` a la entrada de cada bloque, no importa lo que se haya hecho en la función `&do_something()`, el valor de `$1` en la sentencia `print` es el correspondiente al "matching" realizado en el `if`.

31.1.6. Opciones

Modificador	Significado
e	evaluar: evaluar el lado derecho de una sustitución como una expresión
g	global: Encontrar todas las ocurrencias
i	ignorar: no distinguir entre mayúsculas y minúsculas
m	multilínea (`^` y `$` casan con `\n` internos)
o	optimizar: compilar una sola vez
s	`^` y `$` ignoran `\n` pero el punto `.` "casa" con `\n`
x	extendida: permitir comentarios

El Modificador `/g` La conducta de este modificador depende del contexto. En un contexto de listas devuelve una lista con todas las subcadenas casadas por todos los paréntesis en la expresión regular. Si no hubieran paréntesis devuelve una lista con todas las cadenas casadas (como si hubiera paréntesis alrededor del patrón global).

```
1 #!/usr/bin/perl -w
2 ($one, $five, $fifteen) = (‘uptime‘ =~ /(\d+\.\d+)/g);
3 print "$one, $five, $fifteen\n";
```

Observe la salida:

```
> uptime
  1:35pm  up 19:22,  0 users,  load average: 0.01, 0.03, 0.00
> glist.pl
0.01, 0.03, 0.00
```

En un contexto escalar `m//g` itera sobre la cadena, devolviendo cierto cada vez que casa, y falso cuando deja de casar. En otras palabras, recuerda donde se quedo la última vez y se recomienza la búsqueda desde ese punto. Se puede averiguar la posicion del emparejamiento utilizando la función `pos`. Si por alguna razón modificas la cadena en cuestión, la posición de emparejamiento se reestablece al comienzo de la cadena.

```
 1 #!/usr/bin/perl -w
 2 # count sentences in a document
 3 #defined as ending in [.!?] perhaps with
 4 # quotes or parens on either side.
 5 $/ = ""; # paragraph mode
 6 while ($paragraph = <>) {
 7   print $paragraph;
 8   while ($paragraph =~ /[a-z]['")]*[.!?]+['")]*\s/g) {
 9     $sentences++;
10   }
11 }
12 print "$sentences\n";
```

Observe el uso de la variable especial `$/`. Esta variable contiene el separador de registros en el fichero de entrada. Si se iguala a la cadena vacía usará las líneas en blanco como separadores. Se le puede dar el valor de una cadena multicarácter para usarla como delimitador. Nótese que establecerla a `\n\n` es diferente de asignarla a `""`. Si se deja `undef`, la siguiente lectura leerá todo el fichero.

Sigue un ejemplo de ejecución. El programa se llama `gscalar.pl`. Introducimos el texto desde `STDIN`. El programa escribe el número de párrafos:

```
> gscalar.pl
este primer parrafo. Sera seguido de un
segundo parrafo.
```

```
"Cita de Seneca".

3
```

La opción e: Evaluación del remplazo La opción `/e` permite la evaluación como expresión perl de la cadena de reemplazo (En vez de considerarla como una cadena delimitada por doble comilla).

```
1 #!/usr/bin/perl -w
2 $_ = "abc123xyz\n";
3 s/\d+/$&*2/e;
4 print;
5 s/\d+/sprintf("%5d",$&)/e;
6 print;
7 s/\w/$& x 2/eg;
8 print;
```

El resultado de la ejecución es:

```
> replacement.pl
abc246xyz
abc  246xyz
aabbcc  224466xxyyzz
```

Véase un ejemplo con anidamiento de `/e`:

```
1 #!/usr/bin/perl
2 $a ="one";
3 $b = "two";
4 $_ = '$a $b';
5 print "_ = $_\n\n";
6 s/(\$\w+)/$1/ge;
7 print "After 's/(\$\w+)/$1/ge' _ = $_\n\n";
8 s/(\$\w+)/$1/gee;
9 print "After 's/(\$\w+)/$1/gee' _ = $_\n\n";
```

El resultado de la ejecución es:

```
> enested.pl
_ = $a $b

After 's/($w+)/$b/ge' _ = $a $b

After 's/($w+)/$b/gee' _ = one two
```

He aqui una solución que hace uso de `e` al siguiente ejercicio (véase 'Regex to add space after punctuation sign' en PerlMonks) Se quiere poner un espacio en blanco después de la aparición de cada coma:

```
s/,/, /g;
```

pero se quiere que la sustitución no tenga lugar si la coma esta incrustada entre dos dígitos. Además se pide que si hay ya un espacio después de la coma, no se duplique

```
s/(\d[,.]\d)|(,(?!\s))/$1 || "$2 "/ge;
```

Se hace uso de un lookahead negativo `(?!\s)`. Véase la sección 31.2.3 para entender como funciona un lookahead negativo.

31.2. Algunas Extensiones

31.2.1. Comentarios

`(?#text)` Un comentario. Se ignora `text`. Si se usa la opción `x` basta con poner `#`.

31.2.2. Modificadores locales

Los modificadores de la conducta de una expresión regular pueden ser empotrados en una subexpresión usando el formato `(?pimsx-imsx)`.

Véase el correspondiente texto *Extended Patterns* de la sección 'Extended-Patterns' en `perlre`:

> *One or more embedded pattern-match modifiers, to be turned on (or turned off, if preceded by '-') for the remainder of the pattern or the remainder of the enclosing pattern group (if any). This is particularly useful for dynamic patterns, such as those read in from a configuration file, taken from an argument, or specified in a table somewhere. Consider the case where some patterns want to be case sensitive and some do not: The case insensitive ones merely need to include (?i) at the front of the pattern. For example:*
>
> ```
> 1. $pattern = "foobar";
> 2. if ( /$pattern/i ) { }
> 3.
> 4. # more flexible:
> 5.
> 6. $pattern = "(?i)foobar";
> 7. if ( /$pattern/ ) { }
> ```
>
> *These modifiers are restored at the end of the enclosing group. For example,*
>
> ```
> 1. ( (?i) blah ) \s+ \1
> ```
>
> *will match* `blah` *in any case, some spaces, and an exact (including the case!) repetition of the previous word, assuming the* `/x` *modifier, and no* `/i` *modifier outside this group.*

El siguiente ejemplo extiende el ejemplo visto en la sección 31.1.1 eliminando los comentarios `/* ... */` y `// ...` de un programa C. En dicho ejemplo se usaba el modificador `s` para hacer que el punto casara con cualquier carácter:

```
casiano@tonga:~/Lperltesting$ cat -n extendedcomments.pl
   1    #!/usr/bin/perl -w
   2    use strict;
   3
   4    my $progname = shift @ARGV or die "Usage:\n$0 prog.c\n";
   5    open(my $PROGRAM,"<$progname") || die "can't find $progname\n";
   6    my $program = '';
   7    {
   8      local $/ = undef;
   9      $program = <$PROGRAM>;
  10    }
  11    $program =~ s{(?xs)
  12      /\*  # Match the opening delimiter
  13      .*?  # Match a minimal number of characters
  14      \*/  # Match the closing delimiter
  15      |
  16        (?-s)//.* # C++ // comments. No s modifier
  17    }[]g;
  18
  19    print $program;
```

Sigue un ejemplo de ejecución. Usaremos como entrada el programa C:

```
casiano@tonga:~/Lperltesting$ cat -n ehello.c
     1  #include <stdio.h>
     2  /* first
     3  comment
     4  */
     5  main() { // A C++ comment
     6    printf("hello world!\n"); /* second comment */
     7  } // final comment
```

Al ejecutar el programa eliminamos los comentarios:

```
casiano@tonga:~/Lperltesting$ extendedcomments.pl ehello.c | cat -n
     1  #include <stdio.h>
     2
     3  main() {
     4    printf("hello world!\n");
     5  }
```

31.2.3. Mirando hacia adetrás y hacia adelante

El siguiente fragmento esta 'casi' literalmente tomado de la sección 'Looking-ahead-and-looking-behind' en `perlretut`:

Las zero-width assertions como caso particular de mirar atrás-adelante

In Perl regular expressions, most regexp elements 'eat up' a certain amount of string when they match. For instance, the regexp element `[abc}]` *eats up one character of the string when it matches, in the sense that Perl moves to the next character position in the string after the match. There are some elements, however, that don't eat up characters (advance the character position) if they match.*

The examples we have seen so far are the anchors. The anchor `^` *matches the beginning of the line, but doesn't eat any characters.*

Similarly, the word boundary anchor `\b` *matches wherever a character matching* `\w` *is next to a character that doesn't, but it doesn't eat up any characters itself.*

Anchors are examples of zero-width assertions. *Zero-width, because they consume no characters, and assertions, because they test some property of the string.*

In the context of our walk in the woods analogy to regexp matching, most regexp elements move us along a trail, but anchors have us stop a moment and check our surroundings. If the local environment checks out, we can proceed forward. But if the local environment doesn't satisfy us, we must backtrack.

Checking the environment entails either looking ahead on the trail, looking behind, or both.

- `^` *looks behind, to see that there are no characters before.*
- `$` *looks ahead, to see that there are no characters after.*
- `\b` *looks both ahead and behind, to see if the characters on either side differ in their "word-ness".*

The lookahead and lookbehind assertions are generalizations of the anchor concept. Lookahead and lookbehind are zero-width assertions that let us specify which characters we want to test for.

Lookahead assertion

The lookahead assertion is denoted by `(?=regexp)` *and the lookbehind assertion is denoted by* `(?<=fixed-regexp)`.

En español, operador de "trailing" o "mirar-adelante" positivo. Por ejemplo, `/\w+(?=\t)/` *solo casa una palabra si va seguida de un tabulador, pero el tabulador no formará parte de* `$&`. *Ejemplo:*

```
> cat -n lookahead.pl
   1  #!/usr/bin/perl
   2
   3   $a = "bugs the rabbit";
   4   $b = "bugs the frog";
   5   if ($a =~ m{bugs(?= the cat| the rabbit)}i) { print "$a matches. \$& = $&\n"; }
   6   else { print "$a does not match\n"; }
   7   if ($b =~ m{bugs(?= the cat| the rabbit)}i) { print "$b matches. \$& = $&\n"; }
   8   else { print "$b does not match\n"; }
```

Al ejecutar el programa obtenemos:

```
> lookahead.pl
bugs the rabbit matches. $& = bugs
bugs the frog does not match
>
```

Some examples using the debugger[4]*:*

```
  DB<1>      #01234567890123456789
  DB<2> $x = "I catch the housecat 'Tom-cat' with catnip"
  DB<3> print "($&) (".pos($x).")\n" if $x  =~ /cat(?=\s)/g
(cat) (20)                    # matches 'cat' in 'housecat'

  DB<5> $x = "I catch the housecat 'Tom-cat' with catnip" # To reset pos
  DB<6> x @catwords = ($x =~ /(?<=\s)cat\w+/g)
0  'catch'
1  'catnip'

  DB<7>      #012345678901234567890123456789
  DB<8> $x = "I catch the housecat 'Tom-cat' with catnip"
  DB<9> print "($&) (".pos($x).")\n" if $x =~ /\bcat\b/g
(cat) (29) # matches 'cat' in 'Tom-cat'

  DB<10> $x = "I catch the housecat 'Tom-cat' with catnip"
  DB<11> x  $x =~ /(?<=\s)cat(?=\s)/
  empty array
  DB<12> # doesn't match; no isolated 'cat' in middle of $x
```

A hard RegEx problem

Véase el nodo A hard RegEx problem en PerlMonks. Un monje solicita:

> *Hi Monks,*
> *I wanna to match this issues:*
>
> 1. *The string length is between 3 and 10*
> 2. *The string ONLY contains [0-9] or [a-z] or [A-Z], but*
> 3. *The string must contain a number AND a letter at least*
>
> *Pls help me check. Thanks*

Solución:

[4]catnip: La nepeta cataria, también llamada menta de los gatos, de la familia del tomillo y la lavanda. Su perfume desencadena un comportamiento en el animal, similar al del celo

```
casiano@millo:~$ perl -wde 0
main::(-e:1):   0
  DB<1> x 'aaa2a1' =~  /\A(?=.*[a-z])(?=.*\d)\w{3,10}\z/i
0  1
  DB<2> x 'aaaaaa' =~  /\A(?=.*[a-z])(?=.*\d)\w{3,10}\z/i
  empty array
  DB<3> x '1111111' =~  /\A(?=.*[a-z])(?=.*\d)\w{3,10}\z/i
  empty array
  DB<4> x '1111111bbbbb' =~  /\A(?=.*[a-z])(?=.*\d)\w{3,10}\z/i
  empty array
  DB<5> x '111bbbbb' =~  /\A(?=.*[a-z])(?=.*\d)\w{3,10}\z/i
0  1
```

Los paréntesis looakehaed and lookbehind no capturan

Note that the parentheses in `(?=regexp)` *and* `(?<=regexp)` *are non-capturing, since these are zero-width assertions.*

Limitaciones del lookbehind

Lookahead `(?=regexp)` *can match arbitrary regexps, but lookbehind* `(?<=fixed-regexp)` *only works for regexps of fixed width, i.e., a fixed number of characters long.*

Thus `(?<=(ab|bc))` *is fine, but* `(?<=(ab)*)` *is not.*

Negación de los operadores de lookahead y lookbehind

The negated versions of the lookahead and lookbehind assertions are denoted by `(?!regexp)` *and* `(?<!fixed-regexp)` *respectively. They evaluate true if the regexps do not match:*

```
$x = "foobar";
$x =~ /foo(?!bar)/;  # doesn't match, 'bar' follows 'foo'
$x =~ /foo(?!baz)/;  # matches, 'baz' doesn't follow 'foo'
$x =~ /(?<!\s)foo/;  # matches, there is no \s before 'foo'
```

Ejemplo: split con lookahead y lookbehind

Here is an example where a string containing blank-separated words, numbers and single dashes is to be split into its components.

Using `/\s+/` *alone won't work, because spaces are not required between dashes, or a word or a dash. Additional places for a split are established by looking ahead and behind:*

```
casiano@tonga:~$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> $str = "one two - --6-8"
  DB<2> x @toks = split / \s+ | (?<=\S) (?=-) | (?<=-)  (?=\S)/x, $str
0  'one'
1  'two'
2  '-'
3  '-'
4  '-'
5  6
6  '-'
7  8
```

Look Around en perlre

El siguiente párrafo ha sido extraído la sección 'Look-Around-Assertions' en `pelre`. Usémoslo como texto de repaso:

Look-around assertions are zero width patterns which match a specific pattern without including it in `$&`*. Positive assertions match when their subpattern matches, negative assertions match when their subpattern fails. Look-behind matches text up to the current match position, look-ahead matches text following the current match position.*

- `(?=pattern)`

  A zero-width positive look-ahead assertion. For example, `/\w+(?=\t)/` *matches a word followed by a tab, without including the tab in* `$&`*.*

- `(?!pattern)`

  A zero-width negative look-ahead assertion. For example `/foo(?!bar)/` *matches any occurrence of* `foo` *that isn't followed by* `bar`*.*

  Note however that look-ahead and look-behind are NOT the same thing. You cannot use this for look-behind.

  If you are looking for a `bar` *that isn't preceded by a* `foo`*,* `/(?!foo)bar/` *will not do what you want.*

  That's because the `(?!foo)` *is just saying that the next thing cannot be* `foo` *–and it's not, it's a* `bar`*, so* `foobar` *will match.*

  You would have to do something like `/(?!foo)...bar/` *for that.*

  We say "like"because there's the case of your `bar` *not having three characters before it.*

  You could cover that this way: `/(?:(?!foo)...|^.{0,2})bar/`*. Sometimes it's still easier just to say:*

```
if (/bar/ && $‘ !~ /foo$/)
```

  For look-behind see below.

- `(?<=pattern)`

  A zero-width positive look-behind assertion.

  For example, `/(?<=\t)\w+/` *matches a word that follows a tab, without including the tab in* `$&`*. Works only for fixed-width look-behind.*

- `\K`

  There is a special form of this construct, called `\K` *, which causes the regex engine to 'keep' everything it had matched prior to the* `\K` *and not include it in* `$&`*. This effectively provides variable length look-behind. The use of* `\K` *inside of another look-around assertion is allowed, but the behaviour is currently not well defined.*

  For various reasons `\K` *may be significantly more efficient than the equivalent* `(?<=...)` *construct, and it is especially useful in situations where you want to efficiently remove something following something else in a string. For instance*

```
s/(foo)bar/$1/g;
```

  can be rewritten as the much more efficient

```
s/foo\Kbar//g;
```

  Sigue una sesión con el depurador que ilustra la semántica del operador:

```
casiano@millo:~$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> print "& = <$&> 1 = <$1>\n" if "alphabet" =~ /([^aeiou][a-z][aeiou])[a-z]/
& = <phab> 1 = <pha>
  DB<2> print "& = <$&> 1 = <$1>\n" if "alphabet" =~ /\K([^aeiou][a-z][aeiou])[a-z]/
& = <phab> 1 = <pha>
  DB<3> print "& = <$&> 1 = <$1>\n" if "alphabet" =~ /([^aeiou]\K[a-z][aeiou])[a-z]/
& = <hab> 1 = <pha>
```

```
  DB<4> print "& = <$&> 1 = <$1>\n" if "alphabet" =~ /([^aeiou][a-z]\K[aeiou])[a-z]/
& = <ab> 1 = <pha>
  DB<5> print "& = <$&> 1 = <$1>\n" if "alphabet" =~ /([^aeiou][a-z][aeiou])\K[a-z]/
& = <b> 1 = <pha>
  DB<6> print "& = <$&> 1 = <$1>\n" if "alphabet" =~ /([^aeiou][a-z][aeiou])[a-z]\K/
& = <> 1 = <pha>
  DB<7> @a = "alphabet" =~ /([aeiou]\K[^aeiou])/g; print "$&\n"
t
  DB<8> x @a
0  'al'
1  'ab'
2  'et'
```

Otro ejemplo: eliminamos los blancos del final en una cadena:

```
  DB<23> $x = '  cadena entre blancos  '
  DB<24> ($y = $x) =~ s/.*\b\K.*//g
  DB<25> p "<$y>"
<  cadena entre blancos>
```

- `(?<!pattern)`

  A zero-width negative look-behind assertion.

  For example `/(?<!bar)foo/` *matches any occurrence of* `foo` *that does not follow* `bar`. *Works only for fixed-width look-behind.*

Veamos un ejemplo de uso. Se quiere sustituir las extensiones `.something` por `.txt` en cadenas que contienen una ruta a un fichero:

```
casiano@millo:~$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> ($b = $a = 'abc/xyz.something') =~ s{\.[^.]*$}{.txt}
  DB<2> p $b
abc/xyz.txt
  DB<3> ($b = $a = 'abc/xyz.something') =~ s/\.\K[^.]*$/txt/;
  DB<4> p $b
abc/xyz.txt
  DB<5> p $a
abc/xyz.something
```

Véase también:

- `Regexp::Keep` por Jeff Pinyan
- El nodo *positive look behind regexp mystery* en PerlMonks

Operador de predicción negativo: Última ocurrencia

Escriba una expresión regular que encuentre la última aparición de la cadena `foo` en una cadena dada.

```
  DB<6> x ($a = 'foo foo bar bar foo bar bar') =~ /foo(?!.*foo)/g; print pos($a)."\n"
19
  DB<7> x ($a = 'foo foo bar bar foo bar bar') =~ s/foo(?!.*foo)/\U$&/
0  1
  DB<8> x $a
0  'foo foo bar bar FOO bar bar'
```

Diferencias entre mirar adelante negativo y mirar adelante con clase negada

Aparentemente el operador "mirar-adelante" negativo es parecido a usar el operador "mirar-adelante" positivo con la negación de una clase.

`/regexp(?![abc])/`	`/regexp(?=[^abc])/`

Sin embargo existen al menos dos diferencias:

- Una negación de una clase debe casar algo para tener éxito. Un 'mirar-adelante" negativo tiene éxito si, en particular no logra casar con algo. Por ejemplo:

  `\d+(?!\.)` casa con `$a = '452'`, mientras que `\d+(?=[^.])` lo hace, pero porque `452` es `45` seguido de un carácter que no es el punto:

```
> cat lookaheadneg.pl
#!/usr/bin/perl

 $a = "452";
 if ($a =~ m{\d+(?=[^.])}i) { print "$a casa clase negada. \$& = $&\n"; }
 else { print "$a no casa\n"; }
 if ($a =~ m{\d+(?!\.)}i) { print "$a casa predicción negativa. \$& = $&\n"; }
 else { print "$b no casa\n"; }
nereida:~/perl/src> lookaheadneg.pl
452 casa clase negada. $& = 45
452 casa predicción negativa. $& = 452
```

- Una clase negada casa un único carácter. Un 'mirar-adelante" negativo puede tener longitud arbitraria.

AND y AND NOT

Otros dos ejemplos:

- `^(?![A-Z]*$)[a-zA-Z]*$`

  casa con líneas formadas por secuencias de letras tales que no todas son mayúsculas. (Obsérvese el uso de las anclas).

- `^(?=.*?esto)(?=.*?eso)`

  casan con cualquier línea en la que aparezcan `esto` y `eso`. Ejemplo:

```
> cat estoyeso.pl
#!/usr/bin/perl

 my $a = shift;

 if ($a =~ m{^(?=.*?esto)(?=.*?eso)}i) { print "$a matches.\n"; }
 else { print "$a does not match\n"; }

>estoyeso.pl 'hola eso y esto'
hola eso y esto matches.
> estoyeso.pl 'hola esto y eso'
hola esto y eso matches.
> estoyeso.pl 'hola aquello y eso'
```

```
hola aquello y eso does not match
> estoyeso.pl 'hola esto y aquello'
hola esto y aquello does not match
```

El ejemplo muestra que la interpretación es que cada operador mirar-adelante se interpreta siempre a partir de la posición actual de búsqueda. La expresión regular anterior es básicamente equivalente a `(/esto/ && /eso/)`.

- `(?!000)(\d\d\d)`

  casa con cualquier cadena de tres dígitos que no sea la cadena `000`.

Lookahead negativo versus lookbehind

Nótese que el "mirar-adelante" negativo no puede usarse fácilmente para imitar un "mirar-atrás", esto es, que no se puede imitar la conducta de `(?<!foo)bar` mediante algo como `(/?!foo)bar`. Tenga en cuenta que:

- Lo que dice `(?!foo)` es que los tres caracteres que siguen no puede ser `foo`.
- Así, `foo` no pertenece a `/(?!foo)bar/`, pero `foobar` pertenece a `(?!foo)bar/` porque `bar` es una cadena cuyos tres siguientes caracteres son `bar` y no son `foo`.
- Si quisieramos conseguir algo parecido a `(?<!foo)bar` usando un lookahead negativo tendríamos que escribir algo asi como `/(?!foo)...bar/` que casa con una cadena de tres caracteres que no sea `foo` seguida de `bar` (pero que tampoco es exactamente equivalente):

```
pl@nereida:~/Lperltesting$ cat -n foobar.pl
     1  use v5.10;
     2  use strict;
     3
     4  my $a = shift;
     5
     6  for my $r (q{(?<!foo)bar}, q{(?!foo)bar}, q{(?!foo)...bar}) {
     7    if ($a =~ /$r/) {
     8      say "$a casa con $r"
     9    }
    10    else {
    11      say "$a no casa con $r"
    12    }
    13  }
```

- Al ejecutar con diferentes entradas el programa anterior vemos que la solución `q{(?!foo)...bar}` se apróxima mas a `(q{(?<!foo)bar}`:

```
pl@nereida:~/Lperltesting$ perl5.10.1 foobar.pl foobar
foobar no casa con (?<!foo)bar
foobar casa con (?!foo)bar
foobar no casa con (?!foo)...bar

pl@nereida:~/Lperltesting$ perl5.10.1 foobar.pl bar
bar casa con (?<!foo)bar
bar casa con (?!foo)bar
bar no casa con (?!foo)...bar
```

Ejercicio 31.2.1. *Explique porqué* `bar` *casa con* `(?<!foo)bar` *pero no con* `(?!foo)...bar`. *¿Sabría encontrar una expresión regular mas apropiada usando lookahead negativo?*

- En realidad, posiblemente sea mas legible una solución como:

```
if (/bar/ and $` !~ /foo$/)
```

o aún mejor (véase 31.1.4):

```
if (/bar/p && ${^PREMATCH} =~ /foo$/)
```

El siguiente programa puede ser utilizado para ilustrar la equivalencia:

```
pl@nereida:~/Lperltesting$ cat -n foobarprematch.pl
 1  use v5.10;
 2  use strict;
 3
 4  $_ = shift;
 5
 6  if (/bar/p && ${^PREMATCH} =~ /foo$/) {
 7    say "$_ no cumple ".q{/bar/p && ${^PREMATCH} =~ /foo$/};
 8  }
 9  else {
10    say "$_ cumple ".q{/bar/p && ${^PREMATCH} =~ /foo$/};
11  }
12  if (/(?<!foo)bar/) {
13    say "$_ casa con (?<!foo)bar"
14  }
15  else {
16    say "$_ no casa con (?<!foo)bar"
17  }
```

Siguen dos ejecuciones:

```
pl@nereida:~/Lperltesting$ perl5.10.1 foobarprematch.pl bar
bar cumple /bar/p && ${^PREMATCH} =~ /foo$/
bar casa con (?<!foo)bar
pl@nereida:~/Lperltesting$ perl5.10.1 foobarprematch.pl foobar
foobar no cumple /bar/p && ${^PREMATCH} =~ /foo$/
foobar no casa con (?<!foo)bar
```

Ejercicios

Ejercicio 31.2.2.

- *Escriba una sustitución que reemplaze todas las apariciones de* `foo` *por* `foo`, *usando* `\K` *o lookbehind*
- *Escriba una sustitución que reemplaze todas las apariciones de* `lookahead` *por* `look-ahead` *usando lookaheads y lookbehinds*
- *Escriba una expresión regular que capture todo lo que hay entre las cadenas* `foo` *y* `bar` *siempre que no se incluya la palabra* `baz`
- *¿Cuál es la salida?*

```
DB<1> x 'abc' =~ /(?=(.)(.)(.))a(b)/
```

- *Se quiere poner un espacio en blanco después de la aparición de cada coma:*

```
s/,/, /g;
```

pero se quiere que la sustitución no tenga lugar si la coma esta incrustada entre dos dígitos.

- *Se quiere poner un espacio en blanco después de la aparición de cada coma:*

  ```
  s/,/, /g;
  ```

  pero se quiere que la sustitución no tenga lugar si la coma esta incrustada entre dos dígitos. Además se pide que si hay ya un espacio después de la coma, no se duplique

- *¿Cuál es la salida?*

```
pl@nereida:~/Lperltesting$ cat -n ABC123.pl
  1  use warnings;
  2  use strict;
  3
  4  my $c = 0;
  5  my @p = ('^(ABC)(?!123)', '^(\D*)(?!123)',);
  6
  7  for my $r (@p) {
  8    for my $s (qw{ABC123 ABC445}) {
  9      $c++;
 10      print "$c: '$s' =~ /$r/ : ";
 11      <>;
 12      if ($s =~ /$r/) {
 13        print " YES ($1)\n";
 14      }
 15      else {
 16        print " NO\n";
 17      }
 18    }
 19  }
```

31.2.4. Definición de Nombres de Patrones

Perl 5.10 introduce la posibilidad de definir subpatrones en una sección del patrón.

Lo que dice `perlretut` sobre la definición de nombres de patrones

Citando la sección *Defining named patterns* en el documento la sección 'Defining-named-patterns' en `perlretut` para perl5.10:

> *Some regular expressions use identical subpatterns in several places. Starting with Perl 5.10, it is possible to define named subpatterns in a section of the pattern so that they can be called up by name anywhere in the pattern. This syntactic pattern for this definition group is* `"(?(DEFINE)(?<name>pattern)...)"` *An insertion of a named pattern is written as* `(?&name)`*.*

Veamos un ejemplo que define el lenguaje de los números en punto flotante:

```
pl@nereida:~/Lperltesting$ cat -n definingnamedpatterns.pl
 1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
 2  use v5.10;
 3
 4  my $regexp = qr{
 5     ^ (?<num>
 6               (?&osg)[\t\ ]* (?: (?&int)(?&dec)? | (?&dec) )
```

```
 7       )
 8       (?: [eE]
 9       (?<exp> (?&osg)(?&int)) )?
10     $
11        (?(DEFINE)
12         (?<osg>[-+]?)         # optional sign
13         (?<int>\d++)          # integer
14         (?<dec>\.(?&int))     # decimal fraction
15        )
16  }x;
17
18  my $input = <>;
19  chomp($input);
20  my @r;
21  if (@r = $input =~ $regexp) {
22    my $exp = $+{exp} || '';
23    say "$input matches: (num => '$+{num}', exp => '$exp')";
24  }
25  else {
26    say "does not match";
27  }
```

`perlretut` comenta sobre este ejemplo:

> *The example above illustrates this feature. The three subpatterns that are used more than once are the optional sign, the digit sequence for an integer and the decimal fraction. The* `DEFINE` *group at the end of the pattern contains their definition. Notice that the decimal fraction pattern is the first place where we can reuse the integer pattern.*

Lo que dice `perlre` sobre la definición de patrones

Curiosamente, `(DEFINE)` se considera un caso particular de las expresiones regulares condicionales de la forma `(?(condition)yes-pattern)` (véase la sección 31.2.10). Esto es lo que dice la sección 'Extended-Patterns' en `perlre` al respecto:

> *A special form is the* `(DEFINE)` *predicate, which never executes directly its yes-pattern, and does not allow a no-pattern. This allows to define subpatterns which will be executed only by using the recursion mechanism. This way, you can define a set of regular expression rules that can be bundled into any pattern you choose.*
>
> *It is recommended that for this usage you put the* `DEFINE` *block at the end of the pattern, and that you name any subpatterns defined within it.*
>
> *Also, it's worth noting that patterns defined this way probably will not be as efficient, as the optimiser is not very clever about handling them.*
>
> *An example of how this might be used is as follows:*
>
> ```
> 1. /(?<NAME>(?&NAME_PAT))(?<ADDR>(?&ADDRESS_PAT))
> 2.         (?(DEFINE)
> 3.           (?<NAME_PAT>....)
> 4.           (?<ADRESS_PAT>....)
> 5. )/x
> ```
>
> *Note that* **capture buffers matched inside of recursion are not accessible after the recursion returns**, *so the extra layer of capturing buffers is necessary. Thus* `$+{NAME_PAT}` *would not be defined even though* `$+{NAME}` *would be.*

Lo que dice `perlvar` sobre patrones con nombre Esto es lo que dice `perlvar` respecto a las variables implicadas `%+` y `%-`. Con respecto a el hash `%+`:

- `%LAST_PAREN_MATCH`, `%+`

  Similar to `@+` *, the* `%+` *hash allows access to the named capture buffers, should they exist, in the last successful match in the currently active dynamic scope.*

  For example, `$+{foo}` *is equivalent to* `$1` *after the following match:*

  ```
  1. 'foo' =~ /(?<foo>foo)/;
  ```

  The keys of the `%+` *hash list only the names of buffers that have captured (and that are thus associated to defined values).*

  The underlying behaviour of `%+` *is provided by the* `Tie::Hash::NamedCapture` *module.*

  Note: `%-` *and* `%+` *are tied views into a common internal hash associated with the last successful regular expression. Therefore mixing iterative access to them via* `each` *may have unpredictable results. Likewise, if the last successful match changes, then the results may be surprising.*

- `%-`

  Similar to `%+` *, this variable allows access to the named capture buffers in the last successful match in the currently active dynamic scope.* **To each capture buffer name found in the regular expression, it associates a reference to an array containing the list of values captured by all buffers with that name (should there be several of them), in the order where they appear**.

  Here's an example:

  ```
  1. if ('1234' =~ /(?<A>1)(?<B>2)(?<A>3)(?<B>4)/) {
  2.   foreach my $bufname (sort keys %-) {
  3.     my $ary = $-{$bufname};
  4.     foreach my $idx (0..$#$ary) {
  5.       print "\$-{$bufname}[$idx] : ",
  6.             (defined($ary->[$idx]) ? "'$ary->[$idx]'" : "undef"),
  7.             "\n";
  8.     }
  9.   }
 10. }
  ```

  would print out:

  ```
  1. $-{A}[0] : '1'
  2. $-{A}[1] : '3'
  3. $-{B}[0] : '2'
  4. $-{B}[1] : '4'
  ```

  The keys of the `%-` *hash correspond to all buffer names found in the regular expression.*

31.2.5. Patrones Recursivos

Perl 5.10 introduce la posibilidad de definir subpatrones en una sección del patrón. Citando la versión del documento `perlretut` para perl5.10:

> *This feature (introduced in Perl 5.10) significantly extends the power of Perl's pattern matching. By referring to some other capture group anywhere in the pattern with the construct* `(?group-ref)`*, the pattern within the referenced group is used as an independent subpattern in place of the group reference itself. Because the group reference may be contained within the group it refers to, it is now possible to apply pattern matching to tasks that hitherto required a recursive parser.*

...

In `(?...)` *both absolute and relative backreferences may be used. The entire pattern can be reinserted with* `(?R)` *or* `(?0)`*. If you prefer to name your buffers, you can use* `(?&name)` *to recurse into that buffer.*

Palíndromos

Véase un ejemplo que reconoce los palabra-palíndromos (esto es, la lectura directa y la inversa de la cadena pueden diferir en los signos de puntuación):

```
casiano@millo:~/Lperltesting$ cat -n palindromos.pl
  1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
  2  use v5.10;
  3
  4  my $regexp = qr/^(\W*
  5                       (?:
  6                             (\w) (?1) \g{-1}  # palindromo estricto
  7                          |
  8                             \w?               # no recursiva
  9                       )
 10                    \W*)$/ix;
 11
 12  my $input = <>;
 13  chomp($input);
 14  if ($input =~ $regexp) {
 15    say "$input is a palindrome";
 16  }
 17  else {
 18    say "does not match";
 19  }
```

Ejercicio 31.2.3. *¿Cuál es el efecto del modificador* `i` *en la regexp* `qr/^(\W* (?: (\w) (?1) \g{-1} | \w? ) \W`

Siguen algunos ejemplos de ejecución[5]

```
pl@nereida:~/Lperltesting$ ./palindromos.pl
A man, a plan, a canal: Panama!
A man, a plan, a canal: Panama! is a palindrome
pl@nereida:~/Lperltesting$ ./palindromos.pl
A man, a plan, a cam, a yak, a yam, a canal { Panama!
A man, a plan, a cam, a yak, a yam, a canal { Panama! is a palindrome
pl@nereida:~/Lperltesting$ ./palindromos.pl
A man, a plan, a cat, a ham, a yak, a yam, a hat, a canal { Panama!
A man, a plan, a cat, a ham, a yak, a yam, a hat, a canal { Panama! is a palindrome
pl@nereida:~/Lperltesting$ ./palindromos.pl
saippuakauppias
saippuakauppias is a palindrome
pl@nereida:~/Lperltesting$ ./palindromos.pl
dfghjgfd
does not match
```

[5]

- saippuakauppias: Vendedor de jabón (suomi)
- yam: batata (inglés)
- cam: leva

```
pl@nereida:~/Lperltesting$ ./palindromos.pl
...;;;;
...;;;; is a palindrome
```

Lo que dice `perlre` sobre recursividad

> `(?PARNO) (?-PARNO) (?+PARNO) (?R) (?0)`
>
> *Similar to* `(??{ code })` *(véase la sección 31.2.9) except it does not involve compiling any code, instead it treats the contents of a capture buffer as an independent pattern that must match at the current position. Capture buffers contained by the pattern will have the value as determined by the outermost recursion.*
>
> `PARNO` *is a sequence of digits (not starting with 0) whose value reflects the paren-number of the capture buffer to recurse to.*
>
> `(?R)` *recurses to the beginning of the whole pattern.* `(?0)` *is an alternate syntax for* `(?R)`.
>
> *If* `PARNO` *is preceded by a plus or minus sign then it is assumed to be relative, with negative numbers indicating preceding capture buffers and positive ones following. Thus (?-1) refers to the most recently declared buffer, and (?+1) indicates the next buffer to be declared.*
>
> *Note that the counting for relative recursion differs from that of relative backreferences, in that with recursion unclosed buffers are included.*

Hay una diferencia fundamental entre `\g{-1}` y `(?-1)`. El primero significa *lo que casó con el último paréntesis*. El segundo significa que se debe *llamar a la expresión regular que define el último paréntesis*. Véase un ejemplo:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1>  x ($a = "12 aAbB 34") =~ s/([aA])(?-1)(?+1)([bB])/-\1\2-/g
0  1
  DB<2> p $a
12 -aB- 34
```

En `perlre` también se comenta sobre este punto:

> *If there is no corresponding capture buffer defined, then it is a fatal error. Recursing deeper than 50 times without consuming any input string will also result in a fatal error. The maximum depth is compiled into perl, so changing it requires a custom build.*

Paréntesis Equilibrados

El siguiente programa (inspirado en uno que aparece en `perlre`) reconoce una llamada a una función `foo()` que puede contener una secuencia de expresiones con paréntesis equilibrados como argumento:

```
 1  pl@nereida:~/Lperltesting$ cat perlrebalancedpar.pl
 2  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1  -w
 3  use v5.10;
 4  use strict;
 5
 6  my $regexp = qr{ (                    # paren group 1 (full function)
 7                  foo
 8                     (                  # paren group 2 (parens)
 9                       \(
10                          (             # paren group 3 (contents of parens)
11                             (?:
```

```
12                                  [^()]+  # Non-parens
13                                |
14                                  (?2) # Recurse to start of paren group 2
15                              )*
16                           )              # 3
17                        \)
18                       )                 # 2
19                )                        # 1
20      }x;
21
22  my $input = <>;
23  chomp($input);
24  my @res = ($input =~ /$regexp/);
25  if (@res) {
26    say "<$&> is balanced\nParen: (@res)";
27  }
28  else {
29    say "does not match";
30  }
```

Al ejecutar obtenemos:

```
pl@nereida:~/Lperltesting$ ./perlrebalancedpar.pl
foo(bar(baz)+baz(bop))
<foo(bar(baz)+baz(bop))> is balanced
Paren: (foo(bar(baz)+baz(bop)) (bar(baz)+baz(bop)) bar(baz)+baz(bop))
```

Como se comenta en `perlre` es conveniente usar índices relativos si se quiere tener una expresión regular reciclable:

> *The following shows how using negative indexing can make it easier to embed recursive patterns inside of a* `qr//` *construct for later use:*
>
> ```
> 1. my $parens = qr/(\((?:[^()]++|(?-1))*+\))/;
> 2. if (/foo $parens \s+ + \s+ bar $parens/x) {
> 3.   # do something here...
> 4. }
> ```

Véase la sección 31.2.6 para comprender el uso de los operadores posesivos como `++`.

Capturando los bloques de un programa

El siguiente programa presenta una heurística para determinar los bloques de un programa:

```
1    pl@nereida:~/Lperltesting$ cat blocks.pl
2    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
3    use v5.10;
4    use strict;
5    #use re 'debug';
6
7    my $rb = qr{(?x)
8        (
9          \{               # llave abrir
10            (?:
11               [^{}]++   # no llaves
```

```
12              |
13                   [^{}]*+  # no llaves
14                   (?1)     # recursivo
15                   [^{}]*+  # no llaves
16              )*+
17            \}              # llave cerrar
18        )
19      };
20
21    local $/ = undef;
22    my $input = <>;
23    my@blocks = $input =~ m{$rb}g;
24    my $i = 0;
25    say($i++.":\n$_\n===") for @blocks;
```

Veamos una ejecución. Le daremos como entrada el siguiente programa: Al ejecutar el programa con esta entrada obtenemos:

```
pl@nereida:~/Lperltesting$ cat -n blocks.
     1  main() { /* 1 */
     2    { /* 2 */ }
     3    { /* 3 */ }
     4  }
     5
     6  f(){  /* 4 */
     7    {   /* 5 */
     8      { /* 6 */ }
     9    }
    10    {   /* 7 */
    11      { /* 8 */ }
    12    }
    13  }
    14
    15  g(){ /* 9 */
    16  }
    17
    18  h() {
    19  {{{}}}
    20  }
    21  /* end h */
```

```
pl@nereida:~/Lperltesting$ perl5.10.1 blocks.pl b
0:
{ /* 1 */
  { /* 2 */ }
  { /* 3 */ }
}
===
1:
{  /* 4 */
  {   /* 5 */
    { /* 6 */ }
  }
  {   /* 7 */
    { /* 8 */ }
  }
}
===
2:
{ /* 9 */
}
===
3:
{
{{{}}}
}
===
```

Reconocimiento de Lenguajes Recursivos: Un subconjunto de LaTeX

La posibilidad de combinar en las expresiones regulares Perl 5.10 la recursividad con los constructos `(?<name>...)` y `?&name)` así como las secciones `(?(DEFINE) ...)` permiten la escritura de expresiones regulares que reconocen lenguajes recursivos. El siguiente ejemplo muestra un reconocedor de un subconjunto del lenguaje LaTeX (véase la entrada LaTeX en la wikipedia):

```
1  pl@nereida:~/Lperltesting$ cat latex5_10.pl
2  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
```

```
 3  use strict;
 4  use v5.10;
 5
 6  my $regexp = qr{
 7      \A(?&File)\z
 8
 9      (?(DEFINE)
10          (?<File>     (?&Element)*+\s*
11          )
12
13          (?<Element>  \s* (?&Command)
14                     |  \s* (?&Literal)
15          )
16
17          (?<Command>  \\ \s* (?<L>(?&Literal)) \s* (?<Op>(?&Options)?) \s* (?<A>(?&Args))
18             (?{
19                say "command: <$+{L}> options: <$+{Op}> args: <$+{A}>"
20             })
21          )
22
23          (?<Options>  \[ \s* (?:(?&Option) (?:\s*,\s* (?&Option) )*)? \s* \]
24          )
25
26          (?<Args>     (?: \{ \s* (?&Element)* \s* \} )*
27          )
28
29          (?<Option>   \s* [^][\$&%#_{}~^\s,]+
30          )
31
32          (?<Literal>  \s* ([^][\$&%#_{}~^\s]+)
33          )
34      )
35  }xms;
36
37  my $input = do{ local $/; <>};
38  if ($input =~ $regexp) {
39    say "$@: matches:\n$&";
40  }
41  else {
42    say "does not match";
43  }
```

Añadimos una acción semántica al final de la aceptación de un `<Command>`.

```
(?<Command>  \\ \s* (?<L>(?&Literal)) \s* (?<Op>(?&Options)?) \s* (?<A>(?&Args)?)
   (?{
      say "command: <$+{L}> options: <$+{Op}> args: <$+{A}>"
   })
)
```

Esta acción es ejecutada pero no afecta al proceso de análisis. (véase la sección 31.2.8 para mas información sobre las acciones semánticas en medio de una regexp). La acción se limita a mostrar que ha casado con cada una de las tres componentes: el comando, las opciones y los argumentos.

Los paréntesis adicionales, como en `(?<L>(?&Literal))` son necesarios para guardar lo que casó.

Cuando se ejecuta produce la siguiente salida[6]:

```
pl@nereida:~/Lperltesting$ cat prueba.tex
\documentclass[a4paper,11pt]{article}
\usepackage{latexsym}
\author{D. Conway}
\title{Parsing \LaTeX{}}
\begin{document}
\maketitle
\tableofcontents
\section{Description}
...is easy \footnote{But not\\ \emph{necessarily} simple}.
In fact it's easy peasy to do.
\end{document}

pl@nereida:~/Lperltesting$ ./latex5_10.pl prueba.tex
command: <documentclass> options: <[a4paper,11pt]> args: <{article}>
command: <usepackage> options: <> args: <{latexsym}>
command: <author> options: <> args: <{D. Conway}>
command: <LaTeX> options: <> args: <{}>
command: <title> options: <> args: <{Parsing \LaTeX{}}>
command: <begin> options: <> args: <{document}>
command: <maketitle> options: <> args: <>
command: <tableofcontents> options: <> args: <>
command: <section> options: <> args: <{Description}>
command: <emph> options: <> args: <{necessarily}>
command: <footnote> options: <> args: <{But not\\ \emph{necessarily} simple}>
command: <end> options: <> args: <{document}>
: matches:
\documentclass[a4paper,11pt]{article}
\usepackage{latexsym}
\author{D. Conway}
\title{Parsing \LaTeX{}}
\begin{document}
\maketitle
\tableofcontents
\section{Description}
...is easy \footnote{But not\\ \emph{necessarily} simple}.
In fact it's easy peasy to do.
```

[6]

- peasy:A disagreeable taste of very fresh green peas
- easy peasy:

  1. (uk) very easy (short for easy-peasy-lemon-squeezy)
  2. the first half of a rhyming phrase with several alternate second halves, all of which connote an activity or a result that is, respectively, simple to perform or achieve.

     Tie your shoes? Why that's easy peasy lemon squeezy!
     Beat your meat? Why that's easy peasy Japanesey!
     As a red-stater, condemn books and films without having read or seen them? Why that's easy peasy puddin'n'pie!

  3. It comes from a 1970's british TV commercial for Lemon Squeezy detergent. They were with a little girl who points out dirty greasy dishes to an adult (mom or relative) and then this adult produces Lemon Squeezy and they clean the dishes quickly. At the end of the commercial the girl says *Easy Peasy Lemon Squeezy.* Today it is a silly way to state something was or will be very easy.

```
\end{document}
```

La siguiente entrada `prueba3.tex` no pertenece al lenguaje definido por el patrón regular, debido a la presencia de la cadena `$In$` en la última línea:

```
pl@nereida:~/Lperltesting$ cat prueba3.tex
\documentclass[a4paper,11pt]{article}
\usepackage{latexsym}
\author{D. Conway}
\title{Parsing \LaTeX{}}
\begin{document}
\maketitle
\tableofcontents
\section{Description}
\comm{a}{b}
...is easy \footnote{But not\\ \emph{necessarily} simple}.
$In$ fact it's easy peasy to do.
\end{document}

pl@nereida:~/Lperltesting$ ./latex5_10.pl prueba3.tex
command: <documentclass> options: <[a4paper,11pt]> args: <{article}>
command: <usepackage> options: <> args: <{latexsym}>
command: <author> options: <> args: <{D. Conway}>
command: <LaTeX> options: <> args: <{}>
command: <title> options: <> args: <{Parsing \LaTeX{}}>
command: <begin> options: <> args: <{document}>
command: <maketitle> options: <> args: <>
command: <tableofcontents> options: <> args: <>
command: <section> options: <> args: <{Description}>
command: <comm> options: <> args: <{a}{b}>
command: <emph> options: <> args: <{necessarily}>
command: <footnote> options: <> args: <{But not\\ \emph{necessarily} simple}>
does not match
```

Ejercicio 31.2.4. *Obsérvese el uso del cuantificador posesivo en:*

```
 10          (?<File>     (?&Element)*+\s*
 11          )
```

¿Que ocurrre si se quita el posesivo y se vuelve a ejecutar `$ ./latex5_10.pl prueba3.tex`*?*

Reconocimiento de Expresiones Aritméticas

Véase el nodo Complex regex for maths formulas en perlmonks para la formulación del problema. Un monje pregunta:

> *Hiya monks,*
>
> *Im having trouble getting my head around a regular expression to match sequences. I need to catch all exceptions where a mathematical expression is illegal...*
>
> *There must be either a letter or a digit either side of an operator parenthesis must open and close next to letters or digits, not next to operators, and do not have to exist variables must not be more than one letter Nothing other than* `a-z,A-Z,0-9,+,-,*,/,(,)` *can be used*
>
> *Can anyone offer a hand on how best to tackle this problem?*
>
> *many thanks*

La solución parte de que una *expresión* es o bien un *término* o bien un *término* seguido de una operador y un *término*, esto es:

- `termino`
- `termino op termino op termino ...`

que puede ser unificado como `termino (op termino)*`.

Un *término* es un número o un identificador o una *expresión* entre paréntesis, esto es:

- `numero`
- `identificador`
- `( expresión )`

La siguiente expresión regular recursiva sigue esta idea:

```
pl@nereida:~/Lperltesting$ cat -n simpleexpressionsna.pl
  1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
  2    use v5.10;
  3    use strict;
  4    use warnings;
  5
  6    local our ($skip, $term, $expr);
  7    $skip = qr/\s*/;
  8    $expr = qr{ (?<EXPR>
  9                         (?<TERM>              # An expression is a TERM  ...
 10                                $skip (?<ID>[a-zA-Z]+)
 11                              | $skip (?<INT>[1-9]\d*)
 12                              | $skip \(
 13                                $skip  (?&EXPR)
 14                                $skip \)
 15                         ) (?: $skip           # possibly followed by a sequence of ...
 16                               (?<OP>[-+*/])
 17                               (?&TERM)        # ... operand TERM pairs
 18                           )*
 19                 )
 20               }x;
 21    my $re = qr/^ $expr $skip \z/x;
 22    sub is_valid { shift =~ /$re/o }
 23
 24    my @test = ( '(a + 3)', '(3 * 4)+(b + x)', '(5 - a)*z',
 25                 '((5 - a))*((((z)))+2)', '3 + 2', '!3 + 2', '3 + 2!',
 26                 '3 a', '3 3', '3 * * 3',
 27                 '2 - 3 * 4',  '2 - 3 + 4',
 28               );
 29    foreach (@test) {
 30      say("$_:");
 31      say(is_valid($_) ? "\n<$_> is valid" : "\n<$_> is not valid")
 32    }
```

Podemos usar acciones semánticas empotradas para ver la forma en la que trabaja la expresión regular (véase la sección 31.2.8):

```
pl@nereida:~/Lperltesting$ cat -n simpleexpressions.pl
  1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
  2    use v5.10;
  3    use strict;
  4    use warnings;
  5
  6    use re 'eval'; # to allow Eval-group at runtime
  7
  8    local our ($skip, $term, $expr);
  9    $skip = qr/\s*/;
 10    $expr = qr{ (?<EXPR>
 11                        (?<TERM>              # An expression is a TERM  ...
 12                               $skip (?<ID>[a-zA-Z]+)  (?{ print "[ID $+{ID}] "  })
 13                              | $skip (?<INT>[1-9]\d*)  (?{ print "[INT $+{INT}] " })
 14                              | $skip \(                (?{ print "[(] " })
 15                                $skip  (?&EXPR)
 16                                $skip \)                (?{ print "[)] " })
 17                        ) (?: $skip           # possibly followed by a sequence of ...
 18                              (?<OP>[-+*/])             (?{ print "[OP $+{OP}] " })
 19                              (?&TERM)        # ... operand TERM pairs
 20                          )*
 21                )
 22               }x;
 23    my $re = qr/^ $expr $skip \z/x;
 24    sub is_valid { shift =~ /$re/o }
 25
 26    my @test = ( '(a + 3)', '(3 * 4)+(b + x)', '(5 - a)*z',
 27                 '((5 - a))*((((z)))+2)', '3 + 2', '!3 + 2', '3 + 2!',
 28                 '3 a', '3 3', '3 * * 3',
 29                 '2 - 3 * 4',  '2 - 3 + 4',
 30               );
 31    foreach (@test) {
 32      say("$_:");
 33      say(is_valid($_) ? "\n<$_> is valid" : "\n<$_> is not valid")
 34    }
```

Ejecución:

```
pl@nereida:~/Lperltesting$ ./simpleexpressions.pl
(a + 3):
[(] [ID a] [OP +] [INT 3] [)]
<(a + 3)> is valid
(3 * 4)+(b + x):
[(] [INT 3] [OP *] [INT 4] [)] [OP +] [(] [ID b] [OP +] [ID x] [)]
<(3 * 4)+(b + x)> is valid
(5 - a)*z:
[(] [INT 5] [OP -] [ID a] [)] [OP *] [ID z]
<(5 - a)*z> is valid
((5 - a))*((((z)))+2):
[(] [(] [INT 5] [OP -] [ID a] [)] [)] [OP *] [(] [(] [(] [(] [ID z] [)] [)] [)] [OP +] [INT 2]
<((5 - a))*((((z)))+2)> is valid
3 + 2:
[INT 3] [OP +] [INT 2]
<3 + 2> is valid
```

```
!3 + 2:

<!3 + 2> is not valid
3 + 2!:
[INT 3] [OP +] [INT 2]
<3 + 2!> is not valid
3 a:
[INT 3]
<3 a> is not valid
3 3:
[INT 3]
<3 3> is not valid
3 * * 3:
[INT 3] [OP *]
<3 * * 3> is not valid
2 - 3 * 4:
[INT 2] [OP -] [INT 3] [OP *] [INT 4]
<2 - 3 * 4> is valid
2 - 3 + 4:
[INT 2] [OP -] [INT 3] [OP +] [INT 4]
<2 - 3 + 4> is valid
```

31.2.6. Cuantificadores Posesivos

Por defecto, cuando un subpatrón con un cuantificador impide que el patrón global tenga éxito, se produce un backtrack. Hay ocasiones en las que esta conducta da lugar a ineficiencia.

Perl 5.10 provee los cuantificadores posesivos: Un cuantificador posesivo actúa como un cuantificador greedy pero no se produce backtracking.

*+	Casar 0 o mas veces y no retroceder
++	Casar 1 o mas veces y no retroceder
?+	Casar 0 o 1 veces y no retroceder
{n}+	Casar exactamente n veces y no retroceder (redundante)
{n,}+	Casar al menos n veces y no retroceder
{n,m}+	Casar al menos n veces y no mas de m veces y no retroceder

Por ejemplo, la cadena `'aaaa'` no casa con `/(a++a)/` porque no hay retroceso después de leer las 4 aes:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wde 0
main::(-e:1):   0
  DB<1> x 'aaaa' =~ /(a+a)/
0  'aaaa'
  DB<2> x 'aaaa' =~ /(a++a)/
  empty array
```

Cadenas Delimitadas por Comillas Dobles

Los operadores posesivos sirven para poder escribir expresiones regulares mas eficientes en aquellos casos en los que sabemos que el retroceso no conducirá a nuevas soluciones, como es el caso del reconocimiento de las cadenas delimitadas por comillas dobles:

```
pl@nereida:~/Lperltesting$ cat -n ./quotedstrings.pl
     1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
     2  use v5.10;
     3
     4  my $regexp = qr/
     5    "             # double quote
```

```
 6    (?:           # no memory
 7        [^"\\]++  # no " or escape: Don't backtrack
 8      | \\.       # escaped character
 9    )*+
10    "             # end double quote
11  /x;
12
13  my $input = <>;
14  chomp($input);
15  if ($input =~ $regexp) {
16    say "$& is a string";
17  }
18  else {
19    say "does not match";
20  }
```

Paréntesis Posesivos

Los paréntesis posesivos `(?> ...)` dan lugar a un reconocedor que rechaza las demandas de retroceso. De hecho, los operadores posesivos pueden ser reescritos en términos de los paréntesis posesivos: La notación `X++` es equivalente a `(?>X+)`.

Paréntesis Balanceados

El siguiente ejemplo reconoce el lenguaje de los paréntesis balanceados:

```
pl@nereida:~/Lperltesting$ cat -n ./balancedparenthesis.pl
 1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
 2  use v5.10;
 3
 4  my $regexp =
 5      qr/^(
 6             [^()]*+ # no hay paréntesis, no backtrack
 7            \(
 8                (?>        # subgrupo posesivo
 9                   [^()]++ # no hay paréntesis, + posesivo, no backtrack
10                  |(?1)    # o es un paréntesis equilibrado
11                )*
12             \)
13             [^()]*+ # no hay paréntesis
14           )$/x;
15
16  my $input = <>;
17  chomp($input);
18  if ($input =~ $regexp) {
19    say "$& is a balanced parenthesis";
20  }
21  else {
22    say "does not match";
23  }
```

Cuando se ejecuta produce una salida como:

```
pl@nereida:~/Lperltesting$ ./balancedparenthesis.pl
(2*(3+4)-5)*2
(2*(3+4)-5)*2 is a balanced parenthesis
```

```
pl@nereida:~/Lperltesting$ ./balancedparenthesis.pl
(2*(3+4)-5))*2
does not match
pl@nereida:~/Lperltesting$ ./balancedparenthesis.pl
2*(3+4
does not match
pl@nereida:~/Lperltesting$ ./balancedparenthesis.pl
4*(2*(3+4)-5)*2
4*(2*(3+4)-5)*2 is a balanced parenthesis
```

Encontrando los bloques de un programa

El uso de los operadores posesivos nos permite reescribir la solución al problema de encontrar los bloques maximales de un código dada en la sección 31.2.5 de la siguiente manera:

```
 1  pl@nereida:~/Lperltesting$ cat blocksopti.pl
 2  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
 3  use v5.10;
 4  use strict;
 5  #use re 'debug';
 6
 7  my $rb = qr{(?x)
 8      (
 9        \{               # llave abrir
10           (?:
11               [^{}]++   # no llaves
12           |
13               (?1)      # recursivo
14               [^{}]*+   # no llaves
15           )*+
16         \}              # llave cerrar
17      )
18    };
19
20  local $/ = undef;
21  my $input = <>;
22  my@blocks = $input =~ m{$rb}g;
23  my $i = 0;
24  say($i++.":\n$_\n===") for @blocks;
```

Véase también

- Possessive Quantifiers en http://www.regular-expressions.info/
- Nodo *Possessive Quantifiers in Perl 5.10 regexps* en PerlMonks
- perldoc `perlre`

31.2.7. Perl 5.10: Numeración de los Grupos en Alternativas

A veces conviene tener una forma de acceso uniforme a la lista proporcionada por los paréntesis con memoria. Por ejemplo, la siguiente expresión regular reconoce el lenguaje de las horas en notaciones civil y militar:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wde 0
main::(-e:1):   0
```

```
  DB<1> '23:12' =~ /(\d\d|\d):(\d\d)|(\d\d)(\d\d)/; print "1->$1 2->$2\n"
1->23 2->12

  DB<2> '2312' =~ /(\d\d|\d):(\d\d)|(\d\d)(\d\d)/; print "3->$3 4->$4\n"
3->23 4->12
```

Parece inconveniente tener los resultados en variables distintas. El constructo `(?| ...)` hace que los paréntesis se enumeren relativos a las alternativas:

```
  DB<3> '2312' =~ /(?|(\d\d|\d):(\d\d)|(\d\d)(\d\d))/; print "1->$1 2->$2\n"
1->23 2->12

  DB<4> '23:12' =~ /(?|(\d\d|\d):(\d\d)|(\d\d)(\d\d))/; print "1->$1 2->$2\n"
1->23 2->12
```

Ahora en ambos casos `$1` y `$2` contienen las horas y minutos.

31.2.8. Ejecución de Código dentro de una Expresión Regular

Es posible introducir código Perl dentro de una expresión regular. Para ello se usa la notación `(?{code})`.

El siguiente texto esta tomado de la sección 'A-bit-of-magic:-executing-Perl-code-in-a-regular-expression' en `perlretut`:

> *Normally, regexps are a part of Perl expressions. Code evaluation expressions turn that around by allowing arbitrary Perl code to be a part of a regexp. A code evaluation expression is denoted (?code), with code a string of Perl statements.*
>
> *Be warned that this feature is considered experimental, and may be changed without notice.*
>
> *Code expressions are zero-width assertions, and the value they return depends on their environment.*
>
> *There are two possibilities: either the code expression is used as a conditional in a conditional expression* `(?(condition)...)`*, or it is not.*
>
> - *If the code expression is a conditional, the code is evaluated and the result (i.e., the result of the last statement) is used to determine truth or falsehood.*
> - *If the code expression is not used as a conditional, the assertion always evaluates true and the result is put into the special variable* `$^R` *. The variable* `$^R` *can then be used in code expressions later in the regexp*

Resultado de la última ejecución

Las expresiones de código son *zero-width assertions*: no consumen entrada. El resultado de la ejecución se salva en la variable especial `$^R`.

Veamos un ejemplo:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wde 0
main::(-e:1):   0
  DB<1> $x = "abcdef"
  DB<2> $x =~ /abc(?{ "Hi mom\n" })def(?{ print $^R })$/
Hi mom
  DB<3> $x =~ /abc(?{ print "Hi mom\n"; 4 })def(?{ print "$^R\n" })/
Hi mom
4
  DB<4> $x =~ /abc(?{ print "Hi mom\n"; 4 })ddd(?{ print "$^R\n" })/ # does not match
  DB<5>
```

En el último ejemplo (línea `DB<4>`) ninguno de los `print` se ejecuta dado que no hay matching.

El Código empotrado no es interpolado

Tomado de la sección 'Extended-Patterns' en `perlre`:

> *This zero-width assertion evaluates any embedded Perl code. It always succeeds, and its code is not interpolated. Currently, the rules to determine where the code ends are somewhat convoluted.*

Contenido del último paréntesis y la variable por defecto en acciones empotradas

Tomado de la sección 'Extended-Patterns' en `perlre`:

> *. . . can be used with the special variable* `$^N` *to capture the results of submatches in variables without having to keep track of the number of nested parentheses. For example:*

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> $x = "The brown fox jumps over the lazy dog"
  DB<2> x $x =~ /the (\S+)(?{ $color = $^N }) (\S+)(?{ $animal = $^N })/i
0  'brown'
1  'fox'
  DB<3> p "color=$color animal=$animal\n"
color=brown animal=fox
  DB<4> $x =~ /the (\S+)(?{ print (substr($_,0,pos($_))."\n" }) (\S+)/i
The brown
```

> *Inside the* `(?{...})` *block,* `$_` *refers to the string the regular expression is matching against. You can also use* `pos()` *to know what is the current position of matching within this string.*

Los cuantificadores y el código empotrado

Si se usa un cuantificador sobre un código empotrado, actúa como un bucle:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wde 0
main::(-e:1):   0
  DB<1> $x = "aaaa"
  DB<2>  $x =~ /(a(?{ $c++ }))*/
  DB<3> p $c
4
  DB<4> $y = "abcd"
  DB<5> $y =~ /(?:(.)(?{ print "-$1-\n" }))*/
-a-
-b-
-c-
-d-
```

Ámbito

Tomado (y modificado el ejemplo) de la sección 'Extended-Patterns' en `perlre`:

> *. . . The code is properly scoped in the following sense: If the assertion is backtracked (compare la sección 'Backtracking' en* `perlre`*), all changes introduced after localization are undone, so that*

```
pl@nereida:~/Lperltesting$ cat embededcodescope.pl
  use strict;

  our ($cnt, $res);
```

```
 sub echo {
local our $pre = substr($_,0,pos($_));
local our $post = (pos($_) < length)? (substr($_,1+pos($_))) : '';

print("$pre(count = $cnt)$post\n");
 }

 $_ = 'a' x 8;
 m<
(?{ $cnt = 0 }) # Initialize $cnt.
(
  a
  (?{
    local $cnt = $cnt + 1; # Update $cnt, backtracking-safe.
    echo();
  })
)*
aaaa
(?{ $res = $cnt }) # On success copy to non-localized
# location.
 >x;

 print "FINAL RESULT: cnt = $cnt res =$res\n";
```

will set `$res = 4` *. Note that after the match,* `$cnt` *returns to the globally introduced value, because the scopes that restrict local operators are unwound.*

```
pl@nereida:~/Lperltesting$ perl5.8.8 -w embededcodescope.pl
a(count = 1)aaaaaa
aa(count = 2)aaaaa
aaa(count = 3)aaaa
aaaa(count = 4)aaa
aaaaa(count = 5)aa
aaaaaa(count = 6)a
aaaaaaa(count = 7)
aaaaaaaa(count = 8)
FINAL RESULT: cnt = 0 res =4
```

Caveats

- *Due to an unfortunate implementation issue, the Perl code contained in these blocks is treated as a compile time closure that can have seemingly bizarre consequences when used with lexically scoped variables inside of subroutines or loops. There are various workarounds for this, including simply using global variables instead. If you are using this construct and strange results occur then check for the use of lexically scoped variables.*

- *For reasons of security, this construct is forbidden if the regular expression involves run-time interpolation of variables, unless the perilous* `use re 'eval'` *pragma has been used (see* `re`*), or the variables contain results of* `qr//` *operator (see* `"qr/STRING/imosx"` *in* `perlop`*).*
  This restriction is due to the wide-spread and remarkably convenient custom of using run-time determined strings as patterns. For example:

```
1. $re = <>;
2. chomp $re;
3. $string =~ /$re/;
```

Before Perl knew how to execute interpolated code within a pattern, this operation was completely safe from a security point of view, although it could raise an exception from an illegal pattern. If you turn on the `use re 'eval'` *, though, it is no longer secure, so you should only do so if you are also using taint checking. Better yet, use the carefully constrained evaluation within a* `Safe` *compartment. See perlsec for details about both these mechanisms. (Véase la sección 'Taint-mode' en* `perlsec`*)*

- *Because Perl's regex engine is currently* not re-entrant*, interpolated code may not invoke the regex engine either directly with* `m//` *or* `s///`*, or indirectly with functions such as* `split`*.*

Depurando con código empotrado Colisiones en los Nombres de las Subexpresiones Regulares

Las acciones empotradas pueden utilizarse como mecanismo de depuración y de descubrimiento del comportamiento de nuestras expresiones regulares.

En el siguiente programa se produce una colisión entre los nombres `<i>` y `<j>` de los patrones que ocurren en el patrón `<expr>` y en el patrón principal:

```
pl@nereida:~/Lperltesting$ cat -n clashofnamedofssets.pl
  1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
  2    use v5.10;
  3
  4    my $input;
  5
  6    local $" = ", ";
  7
  8    my $parser = qr{
  9        ^ (?<i> (?&expr)) (?<j> (?&expr)) \z
 10          (?{
 11               say "main \$+ hash:";
 12               say " ($_ => $+{$_}) " for sort keys %+;
 13           })
 14
 15        (?(DEFINE)
 16            (?<expr>
 17                (?<i> . )
 18                (?<j> . )
 19                  (?{
 20                      say "expr \$+ hash:";
 21                      say " ($_ => $+{$_}) " for sort keys %+;
 22                  })
 23            )
 24        )
 25    }x;
 26
 27    $input = <>;
 28    chomp($input);
 29    if ($input =~ $parser) {
 30      say "matches: ($&)";
 31    }
```

La colisión hace que la salida sea esta:

```
pl@nereida:~/Lperltesting$ ./clashofnamedofssets.pl
abab
expr $+ hash:
 (i => a)
 (j => b)
expr $+ hash:
 (i => ab)
 (j => b)
main $+ hash:
 (i => ab)
 (j => ab)
matches: (abab)
```

Si se evitan las colisiones, se evita la pérdida de información:

```
pl@nereida:~/Lperltesting$ cat -n namedoffsets.pl
  1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
  2    use v5.10;
  3
  4    my $input;
  5
  6    local $" = ", ";
  7
  8    my $parser = qr{
  9        ^ (?<i> (?&expr)) (?<j> (?&expr)) \z
 10          (?{
 11               say "main \$+ hash:";
 12               say " ($_ => $+{$_}) " for sort keys %+;
 13           })
 14
 15        (?(DEFINE)
 16            (?<expr>
 17                (?<i_e> . )
 18                (?<j_e> . )
 19                  (?{
 20                      say "expr \$+ hash:";
 21                      say " ($_ => $+{$_}) " for sort keys %+;
 22                  })
 23            )
 24        )
 25    }x;
 26
 27    $input = <>;
 28    chomp($input);
 29    if ($input =~ $parser) {
 30      say "matches: ($&)";
 31    }
```

que al ejecutarse produce:

```
pl@nereida:~/Lperltesting$ ./namedoffsets.pl
abab
expr $+ hash:
```

```
 (i_e => a)
 (j_e => b)
expr $+ hash:
 (i => ab)
 (i_e => a)
 (j_e => b)
main $+ hash:
 (i => ab)
 (j => ab)
matches: (abab)
```

31.2.9. Expresiones Regulares en tiempo de matching

Los paréntesis especiales:

```
(??{ Código Perl })
```

hacen que el `Código Perl` sea evaluado durante el tiempo de matching. El resultado de la evaluación se trata como una expresión regular. El match continuará intentando casar con la expresión regular retornada.

Paréntesis con memoria dentro de una *pattern code expression*

Los paréntesis en la expresión regular retornada no cuentan en el patrón exterior. Véase el siguiente ejemplo:

```
pl@nereida:~/Lperltesting$ cat -n postponedregexp.pl
   1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
   2    use v5.10;
   3    use strict;
   4
   5    my $r = qr{(?x)                # ignore spaces
   6                ([ab])             # save 'a' or 'b' in \$1
   7                (??{ "($^N)"x3 })  # 3 more of the same as in \$1
   8              };
   9    say "<$&> lastpar = $#-" if 'bbbb' =~ $r;
  10    say "<$&> lastpar = $#-" if 'aaaa' =~ $r;
  11    say "<abab> didn't match" unless 'abab' =~ $r;
  12    say "<aaab> didn't match" unless 'aaab' =~ $r;
```

Como se ve, hemos accedido desde el código interior al último paréntesis usando `$^N`. Sigue una ejecución:

```
pl@nereida:~/Lperltesting$ ./postponedregexp.pl
<bbbb> lastpar = 1
<aaaa> lastpar = 1
<abab> didn't match
<aaab> didn't match
```

Ejemplo: Secuencias de dígitos de longitud especificada por el primer dígito

Consideremos el problema de escribir una expresión regular que reconoce secuencias no vacías de dígitos tales que la longitud de la secuencia restante viene determinada por el primer dígito. Esta es una solución:

```
pl@nereida:~/Lperltesting$ cat -n intints.pl
   1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
```

```
 2  use v5.10;
 3  use strict;
 4
 5  my $r = qr{(?x)                # ignore spaces
 6             (\d)                # a digit
 7             ( (??{
 8                 "\\d{$^N}"      # as many as the former
 9               })                # digit says
10             )
11            };
12  say "<$&> <$1> <$2>" if '3428' =~ $r;
13  say "<$&> <$1> <$2>" if '228' =~ $r;
14  say "<$&> <$1> <$2>" if '14' =~ $r;
15  say "24 does not match" unless '24' =~ $r;
16  say "4324 does not match" unless '4324' =~ $r;
```

Cuando se ejecuta se obtiene:

```
pl@nereida:~/Lperltesting$ ./intints.pl
<3428> <3> <428>
<228> <2> <28>
<14> <1> <4>
24 does not match
4324 does not match
```

Ejemplo: Secuencias de dígitos no repetidos

Otro ejemplo: queremos escribir una expresión regular que reconozca secuencias de `$n` dígitos en las que no todos los dígitos se repiten. Donde quizá `$n` es capturado de un paréntesis anterior en la expresión regular. Para simplificar la ilustración de la técnica supongamos que `$n = 7`:

```
pl@nereida:~$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> x join '', map { "(?!".$_."{7})" } 0..9
0  '(?!0{7})(?!1{7})(?!2{7})(?!3{7})(?!4{7})(?!5{7})(?!6{7})(?!7{7})(?!8{7})(?!9{7})'
  DB<2> x '7777777' =~ /(??{join '', map { "(?!".$_."{7})" } 0..9})(\d{7})/
  empty array
  DB<3> x '7777778' =~ /(??{join '', map { "(?!".$_."{7})" } 0..9})(\d{7})/
0  7777778
  DB<4> x '4444444' =~ /(??{join '', map { "(?!".$_."{7})" } 0..9})(\d{7})/
  empty array
  DB<5> x '4422444' =~ /(??{join '', map { "(?!".$_."{7})" } 0..9})(\d{7})/
0  4422444
```

Palíndromos con independencia del acento

Se trata en este ejercicio de generalizar la expresión regular introducida en la sección 31.2.5 para reconocer los palabra-palíndromos.

Se trata de encontrar una regexp que acepte que la lectura derecha e inversa de una frase en Español pueda diferir en la acentuación (como es el caso del clásico palíndromo *dábale arroz a la zorra el abad*). Una solución trivial es preprocesar la cadena eliminando los acentos. Supondremos sin embargo que se quiere trabajar sobre la cadena original. He aquí una solucion:

```
 1  pl@nereida:~/Lperltesting$ cat actionspanishpalin.pl
 2  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w -CIOEioA
 3  use v5.10;
```

```
 4  use strict;
 5  use utf8;
 6  use re 'eval';
 7  use Switch;
 8
 9  sub f {
10    my $char = shift;
11
12    switch($char) {
13      case [ qw{a á} ] { return '[aá]' }
14      case [ qw{e é} ] { return '[eé]' }
15      case [ qw{i í} ] { return '[ií]' }
16      case [ qw{o ó} ] { return '[oó]' }
17      case [ qw{u ú} ] { return '[uú]' }
18      else             { return $char  };
19    }
20  }
21
22  my $regexp = qr/^(\W* (?:
23                              (\w) (?-2)(??{ f($^N) })
24                          | \w?
25                       ) \W*
26                   )
27                  $
28                 /ix;
29
30  my $input = <>; # Try: 'dábale arroz a la zorra el abad';
31  chomp($input);
32  if ($input =~ $regexp) {
33    say "$input is a palindrome";
34  }
35  else {
36    say "$input does not match";
37  }
```

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lperltesting$ ./actionspanishpalin.pl
dábale arroz a la zorra el abad
dábale arroz a la zorra el abad is a palindrome
pl@nereida:~/Lperltesting$ ./actionspanishpalin.pl
éoíúaáuioé
éoíúaáuioé is a palindrome
pl@nereida:~/Lperltesting$ ./actionspanishpalin.pl
dáed
dáed does not match
```

Postponiendo para conseguir recursividad

Véase el nodo Complex regex for maths formulas para la formulación del problema:

> *Hiya monks,*
> *Im having trouble getting my head around a regular expression to match sequences. I need to catch all exceptions where a mathematical expression is illegal...*

There must be either a letter or a digit either side of an operator parenthesis must open and close next to letters or digits, not next to operators, and do not have to exist variables must not be more than one letter Nothing other than `a-z,A-Z,0-9,+,-,*,/,(,)` *can be used*

Can anyone offer a hand on how best to tackle this problem?
many thanks

La respuesta dada por ikegami usa `(?{{ ... })` para conseguir una conducta recursiva en versiones de perl anteriores a la 5.10:

```
pl@nereida:~/Lperltesting$ cat -n complexformula.pl
  1    #!/usr/bin/perl
  2    use strict;
  3    use warnings;
  4
  5    sub is_valid_expr {
  6       use re 'eval'; # to allow Eval-group at runtime
  7
  8       local our ($skip, $term, $expr);
  9       $skip = qr! \s* !x;
 10       $term = qr! $skip [a-zA-Z]+              # A term is an identifier
 11                 | $skip [1-9][0-9]*            # or a number
 12                 | $skip \( (??{ $expr }) $skip # or an expression
 13                        \)
 14                 !x;
 15       $expr = qr! $term                         # A expr is a term
 16                   (?: $skip [-+*/] $term )*     # or a term + a term ...
 17                 !x;
 18
 19       return $_[0] =~ / ^ $expr $skip \z /x;
 20    }
 21
 22    print(is_valid_expr($_) ? "$_ is valid\n" : "$_ is not valid\n") foreach (
 23     '(a + 3)',
 24     '(3 * 4)+(b + x)',
 25     '(5 - a)*z',
 26     '3 + 2',
 27
 28     '!3 + 2',
 29     '3 + 2!',
 30
 31     '3 a',
 32     '3 3',
 33     '3 * * 3',
 34
 35     '2 - 3 * 4',
 36     '2 - 3 + 4',
 37    );
```

Sigue el resultado de la ejecución:

```
pl@nereida:~/Lperltesting$ perl complexformula.pl
(a + 3) is valid
(3 * 4)+(b + x) is valid
```

```
(5 - a)*z is valid
3 + 2 is valid
!3 + 2 is not valid
3 + 2! is not valid
3 a is not valid
3 3 is not valid
3 * * 3 is not valid
2 - 3 * 4 is valid
2 - 3 + 4 is valid
```

Caveats

Estos son algunos puntos a tener en cuenta cuando se usan patrones postpuestos. Véase la entrada `(??{ code })` en la sección 'Extended-Patterns' en `perlre`:

> *WARNING: This extended regular expression feature is considered experimental, and may be changed without notice. Code executed that has side effects may not perform identically from version to version due to the effect of future optimisations in the regex engine.*
>
> *This is a* postponed regular subexpression. *The code is evaluated at run time, at the moment this subexpression may match. The result of evaluation is considered as a regular expression and matched as if it were inserted instead of this construct.*
>
> *The code is not interpolated.*
>
> *As before, the rules to determine where the code ends are currently somewhat convoluted.*
>
> *Because perl's regex engine is not currently re-entrant, delayed code may not invoke the regex engine either directly with* `m//` *or* `s///`*), or indirectly with functions such as* `split`*.*
>
> *Recursing deeper than 50 times without consuming any input string will result in a fatal error. The maximum depth is compiled into perl, so changing it requires a custom build.*

31.2.10. Expresiones Condicionales

Citando a `perlre`:

> *A conditional expression is a form of if-then-else statement that allows one to choose which patterns are to be matched, based on some condition.*
>
> *There are two types of conditional expression:* `(?(condition)yes-regexp)` *and* `(?(condition)yes-rege`
>
> `(?(condition)yes-regexp)` *is like an* `if () {}` *statement in Perl. If the condition is true, the yes-regexp will be matched. If the condition is false, the yes-regexp will be skipped and Perl will move onto the next regexp element.*
>
> *The second form is like an* `if () {} else {}` *statement in Perl. If the condition is true, the yes-regexp will be matched, otherwise the no-regexp will be matched.*
>
> *The condition can have several forms.*
>
> - *The first form is simply an integer in parentheses (integer). It is true if the corresponding backreference* `\integer` *matched earlier in the regexp. The same thing can be done with a name associated with a capture buffer, written as* `(<name>)` *or* `('name')`*.*
> - *The second form is a bare zero width assertion* `(?...)`*, either a lookahead, a lookbehind, or a code assertion.*
> - *The third set of forms provides tests that return true if the expression is executed within a recursion* `(R)` *or is being called from some capturing group, referenced either by number* `(R1`*, or by name* `(R&name)`*.*

Condiciones: número de paréntesis

Una expresión condicional puede adoptar diversas formas. La mas simple es un entero en paréntesis. Es cierta si la correspondiente referencia `\integer` casó (también se puede usar un nombre si se trata de un paréntesis con nombre).

En la expresión regular `/^(.)(..)?(?(2)a|b)/` si el segundo paréntesis casa, la cadena debe ir seguida de una `a`, si no casa deberá ir seguida de una `b`:

```
  DB<1> x 'hola' =~ /^(.)(..)?(?(2)a|b)/
0  'h'
1  'ol'
  DB<2> x 'ha' =~ /^(.)(..)?(?(2)a|b)/
  empty array
  DB<3> x 'hb' =~ /^(.)(..)?(?(2)a|b)/
0  'h'
1  undef
```

Ejemplo: cadenas de la forma *una*-otra-otra-*una*

La siguiente búsqueda casa con patrones de la forma `$x$x` o `$x$y$y$x`:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wde 0
main::(-e:1):   0
  DB<1> x 'aa' =~ m{^(\w+)(\w+)?(?(2)\2\1|\1)$}
0  'a'
1  undef
  DB<2> x 'abba' =~ m{^(\w+)(\w+)?(?(2)\2\1|\1)$}
0  'a'
1  'b'
  DB<3> x 'abbc' =~ m{^(\w+)(\w+)?(?(2)\2\1|\1)$}
  empty array
  DB<4> x 'juanpedropedrojuan' =~ m{^(\w+)(\w+)?(?(2)\2\1|\1)$}
0  'juan'
1  'pedro'
```

Condiciones: Código

Una expresión condicional también puede ser un código:

```
  DB<1> $a = 0; print "$&" if 'hola' =~ m{(?(?{$a})hola|adios)} # No hay matching

  DB<2> $a = 1; print "$&" if 'hola' =~ m{(?(?{$a})hola|adios)}
hola
```

Ejemplo: Cadenas con posible paréntesis inicial (no anidados)

La siguiente expresión regular utiliza un condicional para forzar a que si una cadena comienza por un paréntesis abrir termina con un paréntesis cerrar. Si la cadena no comienza por paréntesis abrir no debe existir un paréntesis final de cierre:

```
pl@nereida:~/Lperltesting$ cat -n conditionalregexp.pl
   1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
   2    use v5.10;
   3    use strict;
   4
   5    my $r = qr{(?x)                # ignore spaces
   6                  ^
   7                  ( \( )?           # may be it comes an open par
```

```
 8                [^()]+             # no parenthesis
 9                (?(1)              # did we sart with par?
10                  \)               # if yes then close par
11                )
12                $
13              };
14    say "<$&>" if '(abcd)' =~ $r;
15    say "<$&>" if 'abc' =~ $r;
16    say "<(abc> does not match" unless '(abc' =~ $r;
17    say "<abc)> does not match" unless 'abc)' =~ $r;
```

Al ejecutar este programa se obtiene:

```
pl@nereida:~/Lperltesting$ ./conditionalregexp.pl
<(abcd)>
<abc>
<(abc> does not match
<abc)> does not match
```

Expresiones Condicionales con `(R)`

El siguiente ejemplo muestra el uso de la condición `(R)`, la cual comprueba si la expresión ha sido evaluada dentro de una recursión:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> x 'bbaaaabb' =~ /(b(?(R)a+|(?0))b)/
0  'bbaaaabb'
  DB<2> x 'bb' =~ /(b(?(R)a+|(?0))b)/
  empty array
  DB<3> x 'bab' =~ /(b(?(R)a+|(?0))b)/
  empty array
  DB<4> x 'bbabb' =~ /(b(?(R)a+|(?0))b)/
0  'bbabb'
```

La sub-expresión regular `(?(R)a+|(?0))` dice: si esta siendo evaluada recursivamente admite `a+` si no, evalúa la regexp completa recursivamente.

Ejemplo: Palíndromos con Equivalencia de Acentos Españoles

Se trata en este ejercicio de generalizar la expresión regular introducida en la sección 31.2.5 para reconocer los palabra-palíndromos[7]. Se trata de encontrar una regexp que acepte que la lectura derecha e inversa de una frase en Español pueda diferir en la acentuación (como es el caso del clásico palíndromo *dábale arroz a la zorra el abad*). Una solución trivial es preprocesar la cadena eliminando los acentos. Supondremos sin embargo que se quiere trabajar sobre la cadena original. He aquí una solucion parcial (por consideraciones de legibilidad sólo se consideran las vocales `a` y `o`:

```
 1  pl@nereida:~/Lperltesting$ cat spanishpalin.pl
 2  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w -CIOEioA
 3  use v5.10;
 4  use strict;
 5  use utf8;
 6
 7  my $regexp = qr/^(?<pal>\W* (?:
 8                              (?<L>(?<a>[áa])|(?<e>[ée])|\w) # letter
```

[7] No sé si existe el término. Significa que la lectura directa y la inversa pueden diferir en los signos de puntuación

```
 9                               (?&pal)                         # nested palindrome
10                               (?(<a>)[áa]                     # if is an "a" group
11                                     |(?:((?<e>)[ée]           # if is an "e" group
12                                                |\g{L}         # exact match
13                                           )                  # end if [ée]
14                                     )                        # end group
15                               )                               # end if [áa]
16                             | \w?                             # non rec. case
17                         ) \W*                                 # punctuation symbols
18                     )
19                   $
20                  /ix;
21
22  my $input = <>; # Try: 'dábale arroz a la zorra el abad';
23  chomp($input);
24  if ($input =~ $regexp) {
25    say "$input is a palindrome";
26  }
27  else {
28    say "$input does not match";
29  }
```

Ejecución:

```
pl@nereida:~/Lperltesting$ ./spanishpalin.pl
dábale arroz a la zorra el abad
dábale arroz a la zorra el abad is a palindrome
pl@nereida:~/Lperltesting$ ./spanishpalin.pl
óuuo
óuuo does not match
pl@nereida:~/Lperltesting$ ./spanishpalin.pl
éaáe
éaáe is a palindrome
```

Hemos usado la opción `-CIOEioA` para asegurarnos que los ficheros de entrada/saldia y error y la línea de comandos estan en modo UTF-8. (Véase la sección **??**)

Esto es lo que dice la documentación de `perlrun` al respecto:

> *The* `-C` *flag controls some of the Perl Unicode features.*
>
> *As of 5.8.1, the* `-C` *can be followed either by a number or a list of option letters. The letters, their numeric values, and effects are as follows; listing the letters is equal to summing the numbers.*

```
 1   I 1 STDIN is assumed to be in UTF-8
 2   O 2 STDOUT will be in UTF-8
 3   E 4 STDERR will be in UTF-8
 4   S 7 I + O + E
 5   i 8 UTF-8 is the default PerlIO layer for input streams
 6   o 16 UTF-8 is the default PerlIO layer for output streams
 7   D 24 i + o
 8   A 32 the @ARGV elements are expected to be strings encoded
 9   in UTF-8
10   L 64 normally the "IOEioA" are unconditional,
11   the L makes them conditional on the locale environment
```

```
12   variables (the LC_ALL, LC_TYPE, and LANG, in the order
13   of decreasing precedence) -- if the variables indicate
14   UTF-8, then the selected "IOEioA" are in effect
15   a 256 Set ${^UTF8CACHE} to -1, to run the UTF-8 caching code in
16   debugging mode.
```

For example, `-COE` *and* `-C6` *will both turn on UTF-8-ness on both* `STDOUT` *and* `STDERR`*. Repeating letters is just redundant, not cumulative nor toggling.*

The io options mean that any subsequent `open()` *(or similar I/O operations) will have the* `:utf8` *PerlIO layer implicitly applied to them, in other words,* `UTF-8` *is expected from any input stream, and* `UTF-8` *is produced to any output stream. This is just the default, with explicit layers in* `open()` *and with* `binmode()` *one can manipulate streams as usual.*

`-C` *on its own (not followed by any number or option list), or the empty string* `""` *for the* `PERL_UNICODE` *environment variable, has the same effect as* `-CSDL` *. In other words, the standard I/O handles and the default*`open()` *layer are UTF-8-fied but only if the locale environment variables indicate a UTF-8 locale. This behaviour follows the implicit (and problematic) UTF-8 behaviour of Perl 5.8.0.*

You can use `-C0` *(or* `0` *for* `PERL_UNICODE` *) to explicitly disable all the above Unicode features.*

El pragma `use utf8` hace que se utilice una semántica de carácteres (por ejemplo, la regexp `/./` casará con un carácter unicode), el pragma `use bytes` cambia de semántica de caracteres a semántica de bytes (la regexp `.` casará con un byte).

31.2.11. Verbos que controlan el retroceso

El verbo de control `(*FAIL)`

Tomado de la sección 'Backtracking-control-verbs' en `perlretut`:

The control verb `(*FAIL)` *may be abbreviated as* `(*F)`*. If this is inserted in a regexp it will cause to fail, just like at some mismatch between the pattern and the string. Processing of the regexp continues like after any "normal"failure, so that the next position in the string or another alternative will be tried. As failing to match doesn't preserve capture buffers or produce results, it may be necessary to use this in combination with embedded code.*

```
pl@nereida:~/Lperltesting$ cat -n vowelcount.pl
     1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1  -w
     2  use strict;
     3
     4  my $input = shift() || <STDIN>;
     5  my %count = ();
     6  $input =~ /([aeiou])(?{ $count{$1}++; })(*FAIL)/i;
     7  printf("'%s' => %3d\n", $_, $count{$_})  for (sort keys %count);
```

Al ejecutarse con entrada `supercalifragilistico` produce la salida:

```
pl@nereida:~/Lperltesting$ ./vowelcount.pl
supercalifragilistico
'a' =>   2
'e' =>   1
'i' =>   4
'o' =>   1
'u' =>   1
```

Ejercicio 31.2.5. *¿Que queda en* `$1` *depués de ejecutado el matching* `$input =~ /([aeiou])(?{ $count{$1}++;`

Véase también:

- El nodo en PerlMonks *The Oldest Plays the Piano*
- Véase el ejercicio *Las tres hijas* en la sección 31.4.4

El verbo de control `(*ACCEPT)`

Tomado de `perlretut`:

> *This pattern matches nothing and causes the end of successful matching at the point at which the* `(*ACCEPT)` *pattern was encountered, regardless of whether there is actually more to match in the string. When inside of a nested pattern, such as recursion, or in a subpattern dynamically generated via* `(??{})`*, only the innermost pattern is ended immediately.*
>
> *If the* `(*ACCEPT)` *is inside of capturing buffers then the buffers are marked as ended at the point at which the* `(*ACCEPT)` *was encountered. For instance:*

```
  DB<1> x 'AB' =~ /(A (A|B(*ACCEPT)|C) D)(E)/x
0  'AB'
1  'B'
2  undef
  DB<2> x 'ACDE'  =~ /(A (A|B(*ACCEPT)|C) D)(E)/x
0  'ACD'
1  'C'
2  'E'
```

El verbo `SKIP`

> *This zero-width pattern prunes the backtracking tree at the current point when backtracked into on failure. Consider the pattern* `A (*SKIP) B`*, where* `A` *and* `B` *are complex patterns. Until the* `(*SKIP)` *verb is reached,* `A` *may backtrack as necessary to match. Once it is reached, matching continues in* `B`*, which may also backtrack as necessary; however, should* `B` *not match, then no further backtracking will take place, and the pattern will fail outright at the current starting position.*
>
> *It also signifies that whatever text that was matched leading up to the* `(*SKIP)` *pattern being executed cannot be part of any match of this pattern. This effectively means that the regex engine* `skips` *forward to this position on failure and tries to match again, (assuming that there is sufficient room to match).*
>
> *The name of the* `(*SKIP:NAME)` *pattern has special significance. If a* `(*MARK:NAME)` *was encountered while matching, then it is that position which is used as the "skip point". If no* `(*MARK)` *of that name was encountered, then the* `(*SKIP)` *operator has no effect. When used without a name the "skip pointïs where the match point was when executing the* `(*SKIP)` *pattern.*

Ejemplo:

```
pl@nereida:~/Lperltesting$ cat -n SKIP.pl
     1  #!/soft/perl5lib/bin/perl5.10.1 -w
     2  use strict;
     3  use v5.10;
     4
     5  say "NO SKIP: /a+b?(*FAIL)/";
     6  our $count = 0;
     7  'aaab' =~ /a+b?(?{print "$&\n"; $count++})(*FAIL)/;
     8  say "Count=$count\n";
     9
```

```
10  say "WITH SKIP: a+b?(*SKIP)(*FAIL)/";
11  $count = 0;
12  'aaab' =~ /a+b?(*SKIP)(?{print "$&\n"; $count++})(*FAIL)/;
13  say "WITH SKIP: Count=$count\n";
14
15  say "WITH SKIP /a+(*SKIP)b?(*FAIL)/:";
16  $count = 0;
17  'aaab' =~ /a+(*SKIP)b?(?{print "$&\n"; $count++})(*FAIL)/;
18  say "Count=$count\n";
19
20  say "WITH SKIP /(*SKIP)a+b?(*FAIL): ";
21  $count = 0;
22  'aaab' =~ /(*SKIP)a+b?(?{print "$&\n"; $count++})(*FAIL)/;
23  say "Count=$count\n";
```

Ejecución:

```
pl@nereida:~/Lperltesting$ perl5.10.1 SKIP.pl
NO SKIP: /a+b?(*FAIL)/
aaab
aaa
aa
a
aab
aa
a
ab
a
Count=9

WITH SKIP: a+b?(*SKIP)(*FAIL)/
aaab
WITH SKIP: Count=1

WITH SKIP /a+(*SKIP)b?(*FAIL)/:
aaab
aaa
Count=2

WITH SKIP /(*SKIP)a+b?(*FAIL):
aaab
aaa
aa
a
aab
aa
a
ab
a
Count=9
```

Marcas

Tomado de la sección 'Backtracking-control-verbs' en `perlretut`:

(*MARK:NAME) (*:NAME)

This zero-width pattern can be used to mark the point reached in a string when a certain part of the pattern has been successfully matched. This mark may be given a name. A later (*SKIP) *pattern will then skip forward to that point if backtracked into on failure. Any number of* (*MARK) *patterns are allowed, and the* NAME *portion is optional and may be duplicated.*

In addition to interacting with the (*SKIP) *pattern,* (*MARK:NAME) *can be used to* label *a pattern branch, so that after matching, the program can determine which branches of the pattern were involved in the match.*

When a match is successful, the $REGMARK *variable will be set to the name of the most recently executed* (*MARK:NAME) *that was involved in the match.*

This can be used to determine which branch of a pattern was matched without using a separate capture buffer for each branch, which in turn can result in a performance improvement.

When a match has failed, and unless another verb has been involved in failing the match and has provided its own name to use, the $REGERROR *variable will be set to the name of the most recently executed* (*MARK:NAME).

```
pl@nereida:~/Lperltesting$ cat -n mark.pl
 1  use v5.10;
 2  use strict;
 3
 4  our $REGMARK;
 5
 6  $_ = shift;
 7  say $REGMARK if /(?:x(*MARK:mx)|y(*MARK:my)|z(*MARK:mz))/;
 8  say $REGMARK if /(?:x(*:xx)|y(*:yy)|z(*:zz))/;
```

Cuando se ejecuta produce:

```
pl@nereida:~/Lperltesting$ perl5.10.1 mark.pl y
my
yy
pl@nereida:~/Lperltesting$ perl5.10.1 mark.pl z
mz
zz
```

Poniendo un espacio después de cada signo de puntuación

Se quiere poner un espacio en blanco después de la aparición de cada coma:

```
s/,/, /g;
```

pero se quiere que la sustitución no tenga lugar si la coma esta incrustada entre dos dígitos. Además se pide que si hay ya un espacio después de la coma, no se duplique. Sigue una solución que usa marcas:

```
pl@nereida:~/Lperltesting$ perl5.10.1 -wdE 0
main::(-e:1):   0
  DB<1> $a = 'ab,cd, ef,12,34,efg,56,78,df, ef,'
  DB<2> x ($b = $a) =~ s/\d,\d(*:d)|,(?!\s)/($REGMARK eq 'd')? $& : ', '/ge
0  8
  DB<3> p "<$b>"
<ab, cd, ef, 12,34, efg, 56,78, df, ef, >
```

31.3. Expresiones Regulares en Otros Lenguajes

Vim

- Learn vi/vim in 50 lines and 15 minutes
- VIM Regular Expressions
- Editing features for advanced users
- Vim documentation: pattern
- Vim Regular Expressions Chart

Java

El siguiente ejemplo muestra un programa estilo `grep`: solicita una expresión regular para aplicarla luego a una serie de entradas leídas desde la entrada estandar.

```
casiano@nereida:~/projects/PA/regexp$ cat -n Application.java
 1  /**
 2   * javac Application.java
 3   * java Application
 4   */
 5
 6  import java.io.*;
 7  import java.util.regex.Pattern;
 8  import java.util.regex.Matcher;
 9
10  public class Application {
11
12      public static void main(String[] args){
13          String regexp = "";
14          BufferedReader br = new BufferedReader(new InputStreamReader(System.in));
15          try {
16              System.out.print("Enter your regex: ");
17              regexp = br.readLine();
18          } catch (IOException e) { System.exit(1); };
19          while (true) {
20
21              String input = "";
22              try {
23                System.out.print("Enter input string to search: ");
24                input = br.readLine();
25              } catch (IOException e) { System.exit(1); };
26
27              Pattern pattern = Pattern.compile(regexp);
28              Matcher matcher = pattern.matcher(input);
29
30              boolean found = false;
31              while (matcher.find()) {
32                  System.out.println("I found the text "
33                                     + matcher.group()
34                                     + " starting at index "
35                                     + matcher.start()
36                                     + " and ending at index "
```

```
37                                      +matcher.end()
38                  );
39                  found = true;
40              }
41              if(!found){
42                  System.out.println("No match found.");
43              }
44          }
45      }
46  }
```

Ejecución:

```
casiano@nereida:~/Ljavatesting$ java Application
Enter your regex: (\d+).(\d+)
Enter input string to search: a4b5d6c7efg
I found the text 4b5 starting at index 1 and ending at index 4
I found the text 6c7 starting at index 5 and ending at index 8
Enter input string to search: abc
No match found.
Enter input string to search:
```

Véase también Java Regular Expressions

bash

Esta es una versión en `bash` del conversor de temperaturas visto en las secciones anteriores:

```
pl@nereida:~/src/bash$ cat -n f2c
     1  #!/bin/bash
     2  echo "Enter a temperature (i.e. 32F, 100C):";
     3  read input;
     4
     5  if [ -z "$(echo $input | grep -i '^[-+]\?[0-9]\+\(\.[0-9]*\)\?\ *[CF]$')" ]
     6  then
     7    echo "Expecting a temperature, so don't understand \"$input\"." 1>&2;
     8  else
     9    input=$(echo $input | tr -d ' ');
    10    InputNum=${input:0:${#input}-1};
    11    Type=${input: -1}
    12
    13    if [ $Type = "c" -o $Type = "C" ]
    14    then
    15     celsius=$InputNum;
    16     fahrenheit=$(echo "scale=2; ($celsius * 9/5)+32" | bc -l);
    17    else
    18     fahrenheit=$InputNum;
    19     celsius=$(echo "scale=2; ($fahrenheit -32)*5/9" | bc -l);
    20    fi
    21
    22    echo "$celsius C = $fahrenheit F";
    23  fi
```

C

```
pl@nereida:~/src/regexpr$ cat -n pcregrep.c
 1  #include <stdio.h>
 2  #include <stdlib.h>
 3  #include <string.h>
 4  #include <assert.h>
 5  #include <pcre.h>
 6
 7  char enter_reverse_mode[] = "\33[7m";
 8  char exit_reverse_mode[] = "\33[0m";
 9
10  int main(int argc, char **argv)
11  {
12    const char *pattern;
13    const char *errstr;
14    int erroffset;
15    pcre *expr;
16    char line[512];
17    assert(argc == 2); /* XXX fixme */
18    pattern = argv[1];
19    if (!(expr = pcre_compile(pattern, 0, &errstr, &erroffset, 0))) {
20      fprintf(stderr, "%s: %s\n", pattern, errstr);
21      return EXIT_FAILURE;
22    }
23    while (fgets(line, sizeof line, stdin)) {
24      size_t len = strcspn(line, "\n");
25      int matches[2];
26      int offset = 0;
27      int flags = 0;
28      line[len] = '\0';
29      while (0 < pcre_exec(expr, 0, line, len, offset, flags, matches, 2)) {
30        printf("%.*s%s%.*s%s",
31          matches[0] - offset, line + offset,
32          enter_reverse_mode,
33          matches[1] - matches[0], line + matches[0],
34          exit_reverse_mode);
35        offset = matches[1];
36        flags |= PCRE_NOTBOL;
37      }
38      printf("%s\n", line + offset);
39    }
40    return EXIT_SUCCESS;
41  }
```

Compilación:

```
pl@nereida:~/src/regexpr$ gcc -lpcre pcregrep.c -o pcregrep
```

Cuando se ejecuta espera un patrón en la línea de comandos y pasa a leer desde la entrada estandar. Las cadenas que casan se muestran resaltadas:

```
pl@nereida:~/src/regexpr$ ./pcregrep '\d+'
435 otro 23
```

435 otro 23

```
hola
hola
```

Python

```
pl@nereida:~/src/python$ cat -n c2f.py
 1  #!/usr/local/bin/python
 2  import re
 3
 4  temp = raw_input( ' Introduzca una temperatura (i.e. 32F, 100C): ' )
 5  pattern = re.compile( "^([-+]?[0-9]+(\.[0-9]*)?)\s*([CF])$", re.IGNORECASE )
 6  mo = pattern.match( temp )
 7
 8  if mo:
 9    inputNum = float(mo.group( 1 ))
10    type = mo.group( 3 )
11    celsius = 0.0
12    fahrenheit = 0.0
13    if ( type == "C" or type == "c" ) :
14      celsius = inputNum
15      fahrenheit = ( celsius * 9/5 ) + 32
16    else :
17      fahrenheit = inputNum
18      celsius = ( fahrenheit - 32 ) * 5/9
19    print " ", '%.2f'%(celsius), " C = ", '%.2f'%(fahrenheit), " F\n"
20  else :
21    print " Se experaba una temperatura, no se entiende", temp, "\n"
```

Ruby

```
pl@nereida:~/src/ruby$ cat -n f2c_b
 1  #!/usr/bin/ruby
 2
 3  # Primero leemos una temperatura
 4  class Temperature_calculator
 5    def initialize temp
 6    comp = Regexp.new('^([-+]?\d+(\.\d*)?)\s*([CFcf])$')
 7    if temp =~ comp
 8    begin
 9      cifra = Float($1)
10      @C,@F = ( $3 == "F" or $3 == "f")? [(cifra -32) * 5/9, cifra] : [cifra , cifra * 9/5 +
11    end
12    else
13      raise("Entrada incorrecta")
14    end
15  end
16
17    def show
18      puts "Temperatura en Celsius: #{@C}, temperatura en Fahrenheit: #{@F}"
19    end
20  end
21
22  temperatura = Temperature_calculator.new(readline.chop)
23  temperatura.show
```

Javascript

```
<SCRIPT LANGUAGE="JavaScript"><!--
function demoMatchClick() {
  var re = new RegExp(document.demoMatch.regex.value);
  if (document.demoMatch.subject.value.match(re)) {
    alert("Successful match");
  } else {
    alert("No match");
  }
}

function demoShowMatchClick() {
  var re = new RegExp(document.demoMatch.regex.value);
  var m = re.exec(document.demoMatch.subject.value);
  if (m == null) {
    alert("No match");
  } else {
    var s = "Match at position " + m.index + ":\n";
    for (i = 0; i < m.length; i++) {
      s = s + m[i] + "\n";
    }
    alert(s);
  }
}

function demoReplaceClick() {
  var re = new RegExp(document.demoMatch.regex.value, "g");
  document.demoMatch.result.value =
    document.demoMatch.subject.value.replace(re,
      document.demoMatch.replacement.value);
}
// -->
</SCRIPT>

<FORM ID="demoMatch" NAME="demoMatch" METHOD=POST ACTION="javascript:void(0)">
<P>Regexp: <INPUT TYPE=TEXT NAME="regex" VALUE="\bt[a-z]+\b" SIZE=50></P>
<P>Subject string: <INPUT TYPE=TEXT NAME="subject"
   VALUE="This is a test of the JavaScript RegExp object" SIZE=50></P>
<P><INPUT TYPE=SUBMIT VALUE="Test Match" ONCLICK="demoMatchClick()">
<INPUT TYPE=SUBMIT VALUE="Show Match" ONCLICK="demoShowMatchClick()"></P>

<P>Replacement text: <INPUT TYPE=TEXT NAME="replacement" VALUE="replaced" SIZE=50></P>
<P>Result: <INPUT TYPE=TEXT NAME="result"
   VALUE="click the button to see the result" SIZE=50></P>
<P><INPUT TYPE=SUBMIT VALUE="Replace" ONCLICK="demoReplaceClick()"></P>
</FORM>
```

31.4. Casos de Estudio

31.4.1. Secuencias de números de tamaño fijo

El siguiente problema y sus soluciones se describen en el libro de J.E.F. Friedl [8]. Supongamos que tenemos un texto conteniendo códigos que son números de tamaño fijo, digamos seis dígitos, todos pegados, sin separadores entre ellos, como sigue:

01234567890**123334**234567890123**125934**890123345126

El problema es encontrar los códigos que comienzan por 12. En negrita se han resaltado las soluciones. Son soluciones sólo aquellas que, comienzan por 12 en una posición múltiplo de seis. Una solución es:

```
@nums = grep {m/^12/} m/\d{6}/g;
```

que genera una lista con los números y luego selecciona los que comienzan por 12. Otra solución es:

```
@nums = grep { defined } m/(12\d{4})|\d{6}/g;
```

que aprovecha que la expresión regular devolverá una lista vacía cuando el número no empieza por 12:

```
DB<1> $x = '012345678901123334234567890123125934890123345126'
DB<2> x  ($x =~ m/(12\d{4})|\d{6}/g)
0  undef
1  undef
2  123334
3  undef
4  undef
5  125934
6  undef
7  undef
```

Obsérvese que se esta utilizando también que el operador | no es *greedy*.

¿Se puede resolver el problema usando sólamente una expresión regular? Obsérvese que esta solución "casi funciona":

```
 DB<3> x @nums = $x =~ m/(?:\d{6})*?(12\d{4})/g;
0  123334
1  125934
2  123345
```

recoge la secuencia mas corta de grupos de seis dígitos que no casan, seguida de una secuencia que casa. El problema que tiene esta solución es al final, cuando se han casado todas las soluciones, entonces la búsqueda exhaustiva hará que nos muestre soluciones que no comienzan en posiciones múltiplo de seis. Por eso encuentra 123345:

01234567890**123334**234567890123**125934**890**123345**126

Por eso, Friedl propone esta solución:

```
@nums = m/(?:\d{6})*?(12\d{4})(?:(?!12)\d{6})*/g;
```

Se asume que existe al menos un éxito en la entrada inicial. Que es un extraordinario ejemplo de como el uso de paréntesis de agrupamiento simplifica y mejora la legibilidad de la solución. Es fantástico también el uso del operador de predicción negativo.

Solución usando el ancla \ G

El ancla \G ha sido concebida para su uso con la opción /g. Casa con el punto en la cadena en el que terminó el último emparejamiento. Cuando se trata del primer intento o no se está usando /g, usar \G es lo mismo que usar \A.

Mediante el uso de este ancla es posible formular la siguiente solución al problema planteado:

```
pl@nereida:~/Lperltesting$ perl -wde 0
main::(-e:1):   0
DB<1> $_ = '012345678901123334234567890123125934890123345126'
DB<2> x m/\G(?:\d{6})*?(12\d{4})/g
0  123334
1  125934
```

Sustitución

Si lo que se quiere es sustituir las secuencias deseadas es poisble hacerlo con la siguiente expresión regular:

```
casiano@nereida:~/docs/curriculums/CV_MEC$ perl -wde 0
DB<1> x $x = '012345678901123334234567890123125934890123345126'
0  012345678901123334234567890123125934890123345126
DB<2> x  ($y = $x) =~ s/(12\d{4})|\d{6}/$1? "-$1-":$& /ge
0  8
DB<3> p $y
012345678901-123334-234567890123-125934-890123345126
```

31.4.2. Palabras Repetidas

Su jefe le pide una herramienta que compruebe la aparición de duplicaciones consecutivas en un texto texto (como esta esta y la anterior anterior). La solución debe cumplir las siguientes especificaciones:

- Aceptar cualquier número de ficheros. Resaltar las apariciones de duplicaciones. Cada línea del informe debe estar precedida del nombre del fichero.
- Funcionar no sólo cuando la duplicación ocurre en la misma línea.
- Funcionar independientemente del *case* y de los blancos usados en medio de ambas palabras.
- Las palabras en cuestión pueden estar separadas por *tags* HTML.

```
 1  #!/usr/bin/perl -w
 2  use strict;
 3  use Term::ANSIScreen qw/:constants/;
 4
 5  my $bold = BOLD();
 6  my $clear = CLEAR();
 7  my $line = 1;
 8
 9  # read paragraph
10  local $/ = ".\n";
11  while (my $par = <>) {
12    next unless $par =~ s{
13          \b                  # start word ...
14          ([a-z]+)            # grab word in $1 and \1
15          (                   # save the tags and spaces in $2
16          (\s|<[^>]+>)+       # spaces or HTML tags
```

```
17          )
18          (\1\b)              # repeated word in $4
19    }!$bold$1$clear$2$bold$4$clear!igx;
20
21    $par =~ s/^/"$ARGV(".$line++."): "/meg;   # insert filename and line number
22
23    print $par;
24  }
```

31.4.3. Análisis de cadenas con datos separados por comas

Supongamos que tenemos cierto texto en `$text` proveniente de un fichero `CSV` (*Comma Separated Values*). Esto es el fichero contiene líneas con el formato:

```
"earth",1,"moon",9.374
```

Esta línea representa cinco campos. Es razonable querer guardar esta información en un *array*, digamos `@field`, de manera que `$field[0] == 'earth'`, `$field[1] == '1'`, etc. Esto no sólo implica descomponer la cadena en campos sino también quitar las comillas de los campos entrecomillados. La primera solución que se nos ocurre es hacer uso de la función `split`:

```
@fields = split(/,/,$text);
```

Pero esta solución deja las comillas dobles en los campos entrecomillados. Peor aún, los campos entrecomillados pueden contener comas, en cuyo caso la división proporcionada por `split` sería errónea.

```
 1 #!/usr/bin/perl -w
 2 use Text::ParseWords;
 3
 4 sub parse_csv {
 5   my $text = shift;
 6   my @fields = (); # initialize @fields to be empty
 7
 8   while ($text =~
 9     m/"(([^"\\]|\\.)*)",? # quoted fields
10       |
11       ([^,]+),?           # $3 = non quoted fields
12       |
13       ,                   # allows empty fields
14     /gx
15     )
16   {
17     push(@fields, defined($1)? $1:$3); # add the just matched field
18   }
19   push(@fields, undef) if $text =~ m/,$/; #account for an empty last field
20   return @fields;
21 }
22
23 $test = '"earth",1,"a1, a2","moon",9.374';
24 print "string = \'$test\'\n";
25 print "Using parse_csv\n:";
26 @fields = parse_csv($test);
27 foreach $i (@fields) {
28   print "$i\n";
```

```
 29 }
 30
 31 print "Using Text::ParseWords\n:";
 32 #  @words = &quotewords($delim, $keep, @lines);
 33 #The $keep argument is a boolean flag.  If true, then the
 34 #tokens are split on the specified delimiter, but all other
 35 #characters (quotes, backslashes, etc.) are kept in the
 36 #tokens.  If $keep is false then the &*quotewords()
 37 #functions remove all quotes and backslashes that are not
 38 #themselves backslash-escaped or inside of single quotes
 39 #(i.e., &quotewords() tries to interpret these characters
 40 #just like the Bourne shell).
 41
 42 @fields = quotewords(',',0,$test);
 43 foreach $i (@fields) {
 44   print "$i\n";
 45 }
```

Las subrutinas en Perl reciben sus argumentos en el *array* `@_`. Si la lista de argumentos contiene listas, estas son "aplanadas" en una única lista. Si, como es el caso, la subrutina ha sido declarada antes de la llamada, los argumentos pueden escribirse sin paréntesis que les rodeen:

```
@fields = parse_csv $test;
```

Otro modo de llamar una subrutina es usando el prefijo `&`, pero sin proporcionar lista de argumentos.

```
@fields = &parse_csv;
```

En este caso se le pasa a la rutina el valor actual del *array* `@_`.

Los operadores `push` (usado en la línea 17) y `pop` trabajan sobre el final del *array*. De manera análoga los operadores `shift` y `unshift` lo hacen sobre el comienzo. El operador ternario `?` trabaja de manera análoga como lo hace en C.

El código del `push` podría sustituirse por este otro:

```
push(@fields, $+);
```

Puesto que la variable `$+` contiene la cadena que ha casado con el último paréntesis que haya casado en el ultimo "matching".

La segunda parte del código muestra que existe un módulo en Perl, el módulo `Text::Parsewords` que proporciona la rutina `quotewords` que hace la misma función que nuestra subrutina.

Sigue un ejemplo de ejecución:

```
> csv.pl
string = '"earth",1,"a1, a2","moon",9.374'
Using parse_csv
:earth
1
a1, a2
moon
9.374
Using Text::ParseWords
:earth
1
a1, a2
moon
9.374
```

31.4.4. Las Expresiones Regulares como Exploradores de un Árbol de Soluciones

Números Primos

El siguiente programa evalúa si un número es primo o no:

```
pl@nereida:~/Lperltesting$ cat -n isprime.pl
 1  #!/usr/bin/perl -w
 2  use strict;
 3
 4  my $num = shift;
 5  die "Usage: $0 integer\n" unless (defined($num) && $num =~ /^\d+$/);
 6
 7  if (("1" x $num) =~ /^(11+)\1+$/) {
 8    my $factor = length($1);
 9    print "$num is $factor x ".$num/$factor."\n";
10  }
11  else {
12    print "$num is prime\n";
13  }
```

Siguen varias ejecuciones:

```
pl@nereida:~/Lperltesting$ ./isprime.pl 35.32
Usage: ./isprime.pl integer
pl@nereida:~/Lperltesting$ ./isprime.pl 47
47 is prime
pl@nereida:~/Lperltesting$ ./isprime.pl 137
137 is prime
pl@nereida:~/Lperltesting$ ./isprime.pl 147
147 is 49 x 3
pl@nereida:~/Lperltesting$ ./isprime.pl 137
137 is prime
pl@nereida:~/Lperltesting$ ./isprime.pl 49
49 is 7 x 7
pl@nereida:~/Lperltesting$ ./isprime.pl 47
47 is prime
```

Ecuaciones Diofánticas: Una solución

Según dice la entrada Diophantine_equationen la wikipedia:

> *In mathematics, a Diophantine equation is an indeterminate polynomial equation that allows the variables to be integers only.*

La siguiente sesión con el depurador muestra como se puede resolver una ecuación lineal diofántica con coeficientes positivos usando una expresión regular:

```
  DB<1> # Resolvamos 3x + 2y + 5z = 40
DB<2> x ('a'x40) =~  /^((?:...)+)((?:..)+)((?:.....)+)$/
0  'aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa'
1  'aa'
2  'aaaaa'
DB<3> x map { length }  ('a'x40) =~  /^((?:...)+)((?:..)+)((?:.....)+)$/
0  33
1  2
2  5
```

```
DB<4> @c = (3, 2, 5)
DB<5> x map { length($_) / $c[$i++] }  ('a'x40) =~  /^((?:...)+)((?:..)+)((?:.....)+)$/
0  11
1  1
2  1
DB<6> p 3*11+2*1+5*1
40
```

Ecuaciones Diofánticas: Todas las soluciones

Usando el verbo (*FAIL) es posible obtener todas las soluciones:

```
main::(-e:1):   0
DB<1>  sub equ { my @c = @_; print "\t3*$c[0]+2*$c[1]+5*$c[2] = ",3*$c[0]+2*$c[1]+5*$c[2],"\n"
DB<2> sub f { my @c = ((length($1)/3), (length($2)/2), (length($3)/5)); equ(@c); }
DB<3> x ('a'x40) =~  /^((?:...)+)((?:..)+)((?:.....)+)$(?{ f() })(*FAIL)/x
        3*11+2*1+5*1 = 40
        3*9+2*4+5*1 = 40
        3*8+2*3+5*2 = 40
        3*7+2*7+5*1 = 40
        3*7+2*2+5*3 = 40
        3*6+2*6+5*2 = 40
        3*6+2*1+5*4 = 40
        3*5+2*10+5*1 = 40
        3*5+2*5+5*3 = 40
        3*4+2*9+5*2 = 40
        3*4+2*4+5*4 = 40
        3*3+2*13+5*1 = 40
        3*3+2*8+5*3 = 40
        3*3+2*3+5*5 = 40
        3*2+2*12+5*2 = 40
        3*2+2*7+5*4 = 40
        3*2+2*2+5*6 = 40
        3*1+2*16+5*1 = 40
        3*1+2*11+5*3 = 40
        3*1+2*6+5*5 = 40
        3*1+2*1+5*7 = 40
  empty array
DB<4>
```

Ecuaciones Diofánticas: Resolutor general

El siguiente programa recibe en línea de comandos los coeficientes y término inependeinte de una ecuación lineal diofántica con coeficientes positivos y muestra todas las soluciones. El algoritmo primero crea una cadena conteniendo el código Perl que contiene la expresión regular adecuada para pasar luego a evaluarlo:

```
pl@nereida:~/Lperltesting$ cat -n diophantinesolvergen.pl
 1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1 -w
 2  use v5.10;
 3  use strict;
 4
 5  # Writes a Perl solver for
 6  # a1 x1 + a2 x2 + ... + an xn = b
 7  # a_i and b integers > 0
 8  #
```

```
 9
10  my $b = pop;
11  my @a = @ARGV;
12  my $debug = 1;
13
14  my $b1 = '1'x$b;
15  my @a1 = map { '1'x$_ } @a;
16  my @index = map { 'length($'.$_.")/".$a[$_-1] } 1..(@a);
17  my $aux = join ",", @index;
18
19  my $regexp = '^';
20  $regexp .= "((?:$_)+)" for @a1;
21
22  $regexp .= '$(?{ f() })(*FAIL)';
23
24  my $solver = <<"SOLVER";
25  my \@stack;
26  sub f {
27    my \@s = ($aux);
28    push \@stack, [ \@s ];
29  }
30
31  q{$b1} =~ m{$regexp}x;
32
33  return \@stack;
34  SOLVER
35
36  print "Solver:\n--------\n$solver\n--------\n" if $debug;
37
38  my @stack = eval $solver;
39
40  say("@$_") for @stack
```

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lperltesting$ ./diophantinesolvergen.pl 3 2 5 40
Solver:
--------
my @stack;
sub f {
  my @s = (length($1)/3,length($2)/2,length($3)/5);
  push @stack, [ @s ];
}

q{111111111111111111111111111111111111111111} =~ m{^((?:111)+)((?:11)+)((?:11111)+)$(?{ f() })(*

return @stack;

--------
11 1 1
9 4 1
8 3 2
7 7 1
7 2 3
```

```
6 6 2
6 1 4
5 10 1
5 5 3
4 9 2
4 4 4
3 13 1
3 8 3
3 3 5
2 12 2
2 7 4
2 2 6
1 16 1
1 11 3
1 6 5
1 1 7
```

Las Tres Hijas

En la páginas de Retos Matemáticos de
DIVULGAMAT
puede encontrarse el siguiente problema:

Ejercicio 31.4.1. *Dos matemáticos se vieron en la calle después de muchos años sin coincidir.*

- *¡Hola!, ¿qué tal?, ¿te casaste?, y... ¿cuántos hijos tienes?*
- *Pues tengo tres hijas.*
- *¿y qué años tienen?*
- *¡A ver si lo adivinas!: el producto de las edades de las tres es 36, y su suma es el número del portal que ves enfrente...*
- *¡Me falta un dato!*
- *¡Ah, sí!, ¡la mayor toca el piano!*

¿Qué edad tendrán las tres hijas?

¿Podemos ayudarnos de una expresión regular para resolver el problema? Al ejecutar el siguiente programa:

```
pl@nereida:~/Lperltesting$ cat -n playspiano.pl
 1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1  -w
 2  use v5.10;
 3  use strict;
 4  use List::Util qw{sum};
 5
 6  local our %u;
 7  sub f {
 8    my @a = @_;
 9    @a = sort { $b <=> $a } (length($a[1]), length($a[0])/length($a[1]), 36/length($a[0]) );
10
11    local $" = ", ";
12    say "(@a)\t ".sum(@a) unless exists($u{"@a"});
13    $u{"@a"} = undef;
14  }
```

```
15
16  say "SOL\t\tNUMBER";
17  my @a =  ('1'x36) =~
18           /^((1+)\2+)(\1+)$
19                     (?{ f($1, $2, $3)
20                      })
21             (*FAIL)
22           /x;
```

obtenemos la salida:

```
pl@nereida:~/Lperltesting$ ./playspiano.pl
SOL             NUMBER
(9, 2, 2)        13
(6, 3, 2)        11
(4, 3, 3)        10
(18, 2, 1)       21
(12, 3, 1)       16
(9, 4, 1)        14
(6, 6, 1)        13
```

Explique el funcionamiento del programa. A la vista de la salida ¿Cuáles eran las edades de las hijas?

Mochila 0-1

Para una definición del problema vea la sección El Problema de la Mochila 0-1 en los apuntes de LHP

Ejercicio 31.4.2. *¿Sería capaz de resolver usando expresiones regulares el problema de la mochila 0-1?*

¡Si lo logra merece el premio a la solución mas freak que se haya encontrado para dicho problema!

Véase también

Véase también:

- Véase el nodo en PerlMonks *The Oldest Plays the Piano*
- Solving Algebraic Equations Using Regular Expressions

31.4.5. Número de substituciones realizadas

El operador de substitución devuelve el número de substituciones realizadas, que puede ser mayor que uno si se usa la opción `/g`. En cualquier otro caso retorna el valor falso.

```
1 #!/usr/bin/perl -w
2 undef($/);
3 $paragraph = <STDIN>;
4 $count = 0;
5 $count = ($paragraph =~ s/Mister\b/Mr./ig);
6 print "$paragraph";
7 print "\n$count\n";
```

El resultado de la ejecución es el siguiente:

```
> numsust.pl
Dear Mister Bean,
Is a pleasure for me and Mister Pluto
to invite you to the Opening Session
Official dinner that will be chaired by
Mister Goofy.

Yours sincerely
  Mister Mickey Mouse
Dear Mr. Bean,
Is a pleasure for me and Mr. Pluto
to invite you to the Opening Session
Official dinner that will be chaired by
Mr. Goofy.

Yours sincerely
  Mr. Mickey Mouse

4
```

31.4.6. Expandiendo y comprimiendo tabs

Este programa convierte los tabs en el número apropiado de blancos.

```
pl@nereida:~/Lperltesting$ cat -n expandtabs.pl
   1  #!/usr/bin/perl -w
   2  use strict;
   3
   4  my @string = <>;
   5
   6  for (@string) {
   7    while (s/\t+/' ' x (length($&)*8 - length($`)%8)/e) {}
   8    print $_;
   9  }
```

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lperltesting$ cat -nt tabs.in
     1  012345670123456701234567012345670
     2  one^Itwo^I^Ithree
     3  four^I^I^Ifive
     4  ^I^Itwo
pl@nereida:~/Lperltesting$ ./expandtabs.pl tabs.in | cat -tn
     1  012345670123456701234567012345670
     2  one     two             three
     3  four                            five
     4                  two
```

Ejercicio 31.4.3. *¿Funciona igual si se cambia el bucle* `while` *por una opción* `/g`*?*

```
pl@nereida:~/Lperltesting$ cat -n ./expandtabs2.pl
   1    #!/usr/bin/perl -w
   2    use strict;
   3
   4    my @string = <>;
   5
```

```
6    for (@string) {
7      s/\t+/' ' x (length($&)*8 - length($`)%8)/ge;
8      print $_;
9    }
```

¿Porqué?

31.4.7. Modificación de Múltiples Ficheros: one liner

Aunque no es la forma de uso habitual, Perl puede ser utilizado en "modo sed" para modificar el texto en múltiples ficheros:

`perl -e 's/nereida\.deioc\.ull\.es/miranda.deioc.ull.es/gi' -p -i.bak *.html`

Este programa sustituye la palabra original (g)lobalmente e i)gnorando el "case") en todos los ficheros `*.html` y para cada uno de ellos crea una copia de seguridad `*.html.bak`.

Otro ejemplo: la sustitución que sigue ocurre en todos los ficheros `info.txt` en todos los subdirectorios de los subdirectorios que comiencen por `alu`:

```
perl -e 's/\|hyperpage//gi' -p -i.bak  alu*/*/info.txt
```

Las *opciones de línea* de comandos significan lo siguiente:

-e puede usarse para definir el script en la línea de comandos. Multiples -e te permiten escribir un multi-script. Cuando se usa -e, perl no busca por un fichero de script entre la lista de argumentos.

-p La opción `-p` hace que perl incluya un bucle alrededor de tu "script" al estilo sed:

```
while (<>) {
        ...             # your script goes here
} continue {
        print;
}
```

-n Nótese que las líneas se imprimen automáticamente. Para suprimir la impresión usa la opción `-n`

-i[ext] La opción `-i` Expresa que los ficheros procesados serán modificados. Se renombra el fichero de entrada `file.in` a `file.in.ext`, abriendo el de salida con el mismo nombre del fichero de entrada `file.in`. Se selecciona dicho fichero como de salida por defecto para las sentencias `print`. Si se proporciona una extensión se hace una copia de seguridad. Si no, no se hace copia de seguridad.

En general las opciones pueden ponerse en la primera línea del "script", donde se indica el intérprete. Asi pues, decir

`perl -p -i.bak -e "s/foo/bar/;"`

es equivalente a usar el "script":

```
#!/usr/bin/perl -pi.bak
s/foo/bar/;
```

31.5. tr y split

El operador de traducción permite la conversión de unos caracteres por otros. Tiene la sintáxis:

`tr/SEARCHLIST/REPLACEMENTLIST/cds`

`y/SEARCHLIST/REPLACEMENTLIST/cds`

El operador permite el reemplazo carácter a carácter, por ejemplo:

```
$ perl -de 0
  DB<1> $a = 'fiboncacci'
  DB<2> $a =~ tr/aeiou/AEIOU/
  DB<3> print $a
fIbOncAccI
  DB<4> $a =~ y/fbnc/FBNC/
  DB<5> print $a
FIBONCACCI
```

El operador devuelve el número de carácteres reeemplazados o suprimidos.

```
$cnt = $sky =~ tr/*/*/; # count the stars in $sky
```

Si se especifica el modificador `/d`, cualquier carácter en `SEARCHLIST` que no figure en `REPLACEMENTLIST` es eliminado.

```
  DB<6> print $a
FIBONCACCI
  DB<7> $a =~ y/OA//d
  DB<8> print $a
FIBNCCCI
```

Si se especifica el modificador `/s`, las secuencias de carácteres consecutivos que serían traducidas al mismo carácter son comprimidas a una sola:

```
  DB<1> $b = 'aaghhh!'
  DB<2> $b =~ tr/ah//s
  DB<3> p $b
agh!
```

Observa que si la cadena `REPLACEMENTLIST` es vacía, no se introduce ninguna modificación.

Si se especifica el modificador `/c`, se complementa `SEARCHLIST`; esto es, se buscan los caracteres que no están en `SEARCHLIST`.

```
tr/a-zA-Z/ /cs; # change non-alphas to single space
```

Cuando se dan múltiples traducciones para un mismo carácter, solo la primera es utilizada:

```
tr/AAA/XYZ/
```

traducirá `A` por `X`.

El siguiente *script* busca una expresión regular en el fichero de `passwords` e imprime los *login* de los usuarios que casan con dicha cadena. Para evitar posibles confusiones con las vocales acentuadas se usa el operador `tr`.

```
 1 #!/usr/bin/perl -w
 2 $search = shift(@ARGV) or die("you must provide a regexpr\n");
 3 $search =~ y/ÁÉÍÓÚáéíóú/AEIOUaeiou/;
 4 open(FILE,"/etc/passwd");
 5 while ($line = <FILE>) {
 6   $line =~ y/ÁÉÍÓÚáéíóú/AEIOUaeiou/;
 7   if ($line =~ /$search/io) {
 8     @fields = split(":",$line);
 9     $login = $fields[0];
10     if ($line !~ /^#/) {
11       print "$login\n";
12     }
```

```
 13     else {
 14       print "#$login\n";
 15     }
 16   }
 17 }
 18
```

Ejecución (suponemos que el nombre del fichero anterior es `split.pl`):

```
> split.pl Rodriguez
##direccion
call
casiano
alu5
alu6
##doctorado
paco
falmeida
##ihiu07
```

Para familiarizarte con este operador, codifica y prueba el siguiente código:

```
 1 #!/usr/bin/perl -w
 2 $searchlist = shift @ARGV;
 3 $replacelist = shift @ARGV;
 4 $option = "";
 5 $option = shift @ARGV if @ARGV;
 6
 7 while (<>) {
 8   $num = eval "tr/$searchlist/$replacelist/$option";
 9   die $@ if $@;
10   print "$num: $_";
11 }
```

Perl construye la tabla de traducción en "tiempo de compilación". Por ello ni `SEARCHLIST` ni `REPLACEMENTLIST` son susceptibles de ser interpolados. Esto significa que si queremos usar variables tenemos que recurrir a la función `eval`.

La expresión pasada como parámetro a `eval` en la línea 8 es analizada y ejecutada como si se tratara de un pequeño programa Perl. Cualquier asignación a variables permanece después del `eval`, asi como cualquier definición de subrutina. El código dentro de `eval` se trata como si fuera un bloque, de manera que cualesquiera variables locales (declaradas con `my`) desaparecen al final del bloque.

La variable `$@` contiene el mensaje de error asociado con la última ejecución del comando `eval`. Si es nula es que el último comando se ejecuto correctamente. Aqui tienes un ejemplo de llamada:

```
> tr.pl 'a-z' 'A-Z' s
jose hernandez
13: JOSE HERNANDEZ
joosee hernnandez
16: JOSE HERNANDEZ
```

31.6. Pack y Unpack

El operador `pack` trabaja de forma parecida a `sprintf`. Su primer argumento es una cadena, seguida de una lista de valores a formatear y devuelve una cadena:

```
pack("CCC", 65, 66, 67, 68) # empaquetamos A B C D
```

el inverso es el operador `unpack`

```
unpack("CCC", "ABCD")
```

La cadena de formato es una lista de especificadores que indican el tipo del dato que se va a empaquetar/desempaquetar. Cada especificador puede opcionalmente seguirse de un contador de repetición que indica el número de elementos a formatear. Si se pone un asterisco (`*`) se indica que la especificación se aplica a todos los elementos restantes de la lista.

Formato	Descripción
A	Una cadena completada con blancos
a	Una cadena completada con ceros
B	Una cadena binaria en orden descendente
b	Una cadena binaria en orden ascendente
H	Una cadena hexadecimal, los nibble altos primero
h	Una cadena hexadecimal, los nibble bajos primero

Ejemplo de uso del formato `A`:

```
DB<1> $a = pack "A2A3", "Pea","rl"
DB<2> p $a
Perl
DB<3> @b = unpack "A2A3", "Perl"
DB<4> p "@b"
Pe rl
```

La variable `@b` tiene ahora dos cadenas. Una es `Pe` la otra es `rl`. Veamos un ejemplo con el formato `B`:

```
p ord('A')
65
DB<22> $x = pack "B8", "01000001"
DB<23> p $x
A
DB<24> @y = unpack "B8", "A"
DB<25> p "@y"
01000001
DB<26> $x = pack "b8", "10000010"
DB<27> p $x
```

31.7. Práctica: Un lenguaje para Componer Invitaciones

En el capítulo 6 (sección 6.4.2.2) del libro The LaTex Web Companion se define un lenguaje para componer textos para enviar invitaciones.

Para escribir una invitación en ese lenguaje escribiríamos algo así:

```
pl@nereida:~/Lpl0910/Practicas/161009/src$ cat -n invitation.xml
 1  <?xml version="1.0"?>
 2  <!DOCTYPE invitation SYSTEM "invitation.dtd">
 3  <invitation>
 4  <!-- ++++ The header part of the document ++++ -->
 5  <front>
 6  <to>Anna, Bernard, Didier, Johanna</to>
 7  <date>Next Friday Evening at 8 pm</date>
 8  <where>The Web Cafe</where>
 9  <why>My first XML baby</why>
10  </front>
```

```
11  <!-- +++++ The main part of the document +++++ -->
12  <body>
13  <par>
14  I would like to invite you all to celebrate
15  the birth of <emph>Invitation</emph>, my
16  first XML document child.
17  </par>
18  <par>
19  Please do your best to come and join me next Friday
20  evening. And, do not forget to bring your friends.
21  </par>
22  <par>
23  I <emph>really</emph> look forward to see you soon!
24  </par>
25  </body>
26  <!-- +++ The closing part of the document ++++ -->
27  <back>
28  <signature>Michel</signature>
29  </back>
30  </invitation>
```

La sintáxis del lenguaje queda reflejada en la siguiente *Document Type Definition* (*DTD*) que aparece en la sección 6.4.3 del libro de Goosens:

```
pl@nereida:~/Lpl0910/Practicas/161009/src$ cat -n invitation.dtd
 1  <!-- invitation DTD   -->
 2  <!-- May 26th 1998 mg -->
 3  <!ELEMENT invitation (front, body, back) >
 4  <!ELEMENT front      (to, date, where, why?) >
 5  <!ELEMENT date       (#PCDATA) >
 6  <!ELEMENT to         (#PCDATA) >
 7  <!ELEMENT where      (#PCDATA) >
 8  <!ELEMENT why        (#PCDATA) >
 9  <!ELEMENT body       (par+) >
10  <!ELEMENT par        (#PCDATA|emph)* >
11  <!ELEMENT emph       (#PCDATA) >
12  <!ELEMENT back       (signature) >
13  <!ELEMENT signature  (#PCDATA) >
```

El objetivo de esta práctica es escribir un programa Perl que usando las extensiones para expresiones regulares presentes en la versión 5.10 reconozca el lenguaje anterior.

Véase también:

- The LaTex Web Companion
- Examples from The LaTeX Web Companion (véanse los subdirectorios correspondietnes a los capítulos 6 y 7)

31.8. Analisis Sintáctico con Expresiones Regulares Perl

31.8.1. Introducción al Anaĺisis Sintáctico con Expresiones Regulares

Como se ha comentado en la sección 31.2.5 Perl 5.10 permite el reconocimiento de expresiones definidas mediante gramáticas recursivas, siempre que estas puedan ser analizadas por un analizador

recursivo descendente. Sin embargo, las expresiones regulares Perl 5.10 hace difícil construir una representación del árbol de análisis sintáctico abstracto. Además, la necesidad de explicitar en la regexp los blancos existentes entre los símbolos hace que la descripción sea menos robusta y menos legible.

Ejemplo: Traducción de expresiones aritméticas en infijo a postfijo

El siguiente ejemplo muestra una expresión regular que traduce expresiones de diferencias en infijo a postfijo.

Se usa una variable `$tran` para calcular la traducción de la subexpresión vista hasta el momento.

La gramática original que consideramos es recursiva a izquierdas:

```
exp ->   exp '-' digits
       | digits
```

aplicando las técnicas explicadas en 33.8.1 y en el nodo de perlmonks 553889 transformamos la gramática en:

```
exp ->   digits rest
rest ->   '-' rest
        | # empty
```

Sigue el código:

```
pl@nereida:~/Lperltesting$ cat -n infixtopostfix.pl
  1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
  2    use v5.10;
  3
  4    # Infix to postfix translator using 5.10 regexp
  5    # original grammar:
  6    #  exp ->   exp '-' digits
  7    #         | digits
  8    #
  9    # Applying left-recursion elimination we have:
 10    #  exp ->   digits rest
 11    #  rest ->   '-' rest
 12    #          | # empty
 13    #
 14    my $input;
 15    local our $tran = '';
 16
 17    my $regexp = qr{
 18        (?&exp)
 19
 20        (?(DEFINE)
 21            (?<exp>    ((?&digits)) \s* (?{ $tran .= "$^N "; say "tran=$tran"; }) (?&rest)
 22                          (?{
 23                             say "exp -> digits($^N) rest";
 24                          })
 25            )
 26
 27            (?<rest>     \s* - ((?&digits)) (?{ $tran .= "$^N - "; say "tran=$tran"; }) (?&
 28                            (?{
 29                               say "rest -> - digits($^N) rest";
 30                            })
 31                       |  # empty
 32                            (?{
```

```
33                                say "rest -> empty";
34                              })
35            )
36
37            (?<digits>  \s* (\d+)
38            )
39        )
40    }xms;
41
42    $input = <>;
43    chomp($input);
44    if ($input =~ $regexp) {
45      say "matches: $&\ntran=$tran";
46    }
47    else {
48      say "does not match";
49    }
```

La variable `$^N` contiene el valor que casó con el último paréntesis. Al ejecutar el código anterior obtenemos:

Véase la ejecución:

```
pl@nereida:~/Lperltesting$ ./infixtopostfix.pl
ab 5 - 3 -2 cd;
tran= 5
tran= 5  3 -
tran= 5  3 - 2 -
rest -> empty
rest -> - digits(2) rest
rest -> - digits( 3) rest
exp -> digits( 5) rest
matches:  5 - 3 -2
tran= 5  3 - 2 -
```

Como se ve, el recorrido primero profundo se traduce en la reconstrucción de una derivación a derechas.

Accediendo a los atributos de paréntesis anteriores mediante acciones intermedias

Es difícil extender el ejemplo anterior a lenguajes mas complejos debido a la limitación de que sólo se dispone de acceso al último paréntesis vía `$^N`. En muchos casos es necesario poder acceder a paréntesis/atributos anteriores.

El siguiente código considera el caso de expresiones con sumas, restas, multiplicaciones y divisiones. Utiliza la variable `op` y una acción intermedia (líneas 51-53) para almacenar el segundo paréntesis necesitado:

```
pl@nereida:~/Lperltesting$ cat -n ./calc510withactions3.pl
 1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
 2    use v5.10;
 3
 4    # Infix to postfix translator using 5.10 regexp
 5    # Original grammar:
 6
 7    #  exp ->   exp [-+] term
 8    #         | term
 9    #  term ->   term [*/] digits
```

```
10  #          | digits
11
12  # Applying left-recursion elimination we have:
13
14  #  exp  ->   term re
15  #  re   ->   [+-] term re
16  #          | # empty
17  #  term ->   digits rt
18  #  rt   ->   [*/] rt
19  #          | # empty
20
21
22  my $input;
23  my @stack;
24
25  local our $op = '';
26  my $regexp = qr{
27      (?&exp)
28
29      (?(DEFINE)
30          (?<exp>    (?&term) (?&re)
31                       (?{ say "exp -> term re" })
32          )
33
34          (?<re>     \s* ([+-]) (?&term) \s* (?{ push @stack, $^N }) (?&re)
35                       (?{ say "re -> [+-] term re" })
36                   | # empty
37                       (?{ say "re -> empty" })
38          )
39
40          (?<term>   ((?&digits))
41                        (?{ # intermediate action
42                            push @stack, $^N
43                        })
44                     (?&rt)
45                        (?{
46                            say "term-> digits($^N) rt";
47                        })
48          )
49
50          (?<rt>     \s*([*/])
51                             (?{ # intermediate action
52                                local $op = $^N;
53                             })
54                     ((?&digits)) \s*
55                             (?{  # intermediate action
56                                  push @stack, $^N, $op
57                              })
58                     (?&rt) # end of <rt> definition
59                             (?{
60                                  say "rt -> [*/] digits($^N) rt"
61                              })
62                    |  # empty
```

```
63                          (?{ say "rt -> empty" })
64            )
65
66            (?<digits>  \s* \d+
67            )
68        )
69    }xms;
70
71    $input = <>;
72    chomp($input);
73    if ($input =~ $regexp) {
74      say "matches: $&\nStack=(@stack)";
75    }
76    else {
77      say "does not match";
78    }
```

Sigue una ejecución:

```
pl@nereida:~/Lperltesting$ ./calc510withactions3.pl
5-8/4/2-1
rt -> empty
term-> digits(5) rt
rt -> empty
rt -> [*/] digits(2) rt
rt -> [*/] digits(4) rt
term-> digits(8) rt
rt -> empty
term-> digits(1) rt
re -> empty
re -> [+-] term re
re -> [+-] term re
exp -> term re
matches: 5-8/4/2-1
Stack=(5 8 4 / 2 / - 1 -)
```

Accediendo a los atributos de paréntesis anteriores mediante `@-`

Sigue una solución alternativa que obvia la necesidad de introducir incómodas acciones intermedias. Utilizamos las variables `@-` y `@+`:

> *Since Perl 5.6.1 the special variables* `@-` *and* `@+` *can functionally replace* `$‘`, `$&` *and* `$’`. *These arrays contain pointers to the beginning and end of each match (see* `perlvar` *for the full story), so they give you essentially the same information, but without the risk of excessive string copying.*

Véanse los párrafos en las páginas 165, 165) y 166 para mas información sobre `@-` y `@+`.

Nótese la función `rc` en las líneas 21-28. `rc(1)` nos retorna lo que casó con el último paréntesis, `rc(2)` lo que casó con el penúltimo, etc.

```
pl@nereida:~/Lperltesting$ cat -n calc510withactions4.pl
 1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
 2    use v5.10;
 3
 4    # Infix to postfix translator using 5.10 regexp
 5    # Original grammar:
```

```
 6
 7    #  exp ->   exp [-+] term
 8    #         | term
 9    #  term ->   term [*/] digits
10    #          | digits
11
12    # Applying left-recursion elimination we have:
13
14    #  exp  ->   term re
15    #  re   ->   [+-] term re
16    #          | # empty
17    #  term ->   digits rt
18    #  rt   ->   [*/] rt
19    #          | # empty
20
21    sub rc {
22      my $ofs = - shift;
23
24      # Number of parenthesis that matched
25      my $np = @-;
26      #      string, ofsset, length
27      substr($_, $-[$ofs], $+[$np+$ofs] - $-[$ofs])
28    }
29
30    my $input;
31    my @stack;
32
33    my $regexp = qr{
34        (?&exp)
35
36        (?(DEFINE)
37            (?<exp>    (?&term) (?&re)
38                         (?{ say "exp -> term re" })
39            )
40
41            (?<re>     \s* ([+-]) (?&term) \s* (?{ push @stack, rc(1) }) (?&re)
42                         (?{ say "re -> [+-] term re" })
43                     | # empty
44                         (?{ say "re -> empty" })
45            )
46
47            (?<term>   ((?&digits))
48                          (?{ # intermediate action
49                              push @stack, rc(1)
50                          })
51                       (?&rt)
52                          (?{
53                              say "term-> digits(".rc(1).") rt";
54                          })
55            )
56
57            (?<rt>     \s*([*/]) ((?&digits)) \s*
58                                  (?{  # intermediate action
```

```
59                                      push @stack, rc(1), rc(2)
60                                   })
61                          (?&rt) # end of <rt> definition
62                                  (?{
63                                      say "rt -> [*/] digits(".rc(1).") rt"
64                                   })
65                         |  # empty
66                            (?{ say "rt -> empty" })
67             )
68
69             (?<digits>  \s* \d+
70             )
71         )
72     }xms;
73
74     $input = <>;
75     chomp($input);
76     if ($input =~ $regexp) {
77       say "matches: $&\nStack=(@stack)";
78     }
79     else {
80       say "does not match";
81     }
```

Ahora accedemos a los atributos asociados con los dos paréntesis, en la regla de `<rt>` usando la función `rc`:

```
         (?<rt>     \s*([*/]) ((?&digits)) \s*
                            (?{  # intermediate action
                                 push @stack, rc(1), rc(2)
                             })
```

Sigue una ejecución del programa:

```
pl@nereida:~/Lperltesting$ ./calc510withactions4.pl
5-8/4/2-1
rt -> empty
term-> digits(5) rt
rt -> empty
rt -> [*/] digits(2) rt
rt -> [*/] digits(4) rt
term-> digits(8) rt
rt -> empty
term-> digits(1) rt
re -> empty
re -> [+-] term re
re -> [+-] term re
exp -> term re
matches: 5-8/4/2-1
Stack=(5 8 4 / 2 / - 1 -)
pl@nereida:~/Lperltesting$
```

Accediendo a los atributos de paréntesis anteriores mediante paréntesis con nombre

Una nueva solución: dar nombre a los paréntesis y acceder a los mismos:

```
47          (?<rt>     \s*(?<op>[*/]) (?<num>(?&digits)) \s*
48                              (?{  # intermediate action
49                                   push @stack, $+{num}, $+{op}
50                               })
```

Sigue el código completo:

```
pl@nereida:~/Lperltesting$ cat -n ./calc510withnamedpar.pl
 1    #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
 2    use v5.10;
 3
 4    # Infix to postfix translator using 5.10 regexp
 5    # Original grammar:
 6
 7    #  exp ->   exp [-+] term
 8    #         | term
 9    #  term ->   term [*/] digits
10    #          | digits
11
12    # Applying left-recursion elimination we have:
13
14    #  exp  ->   term re
15    #  re   ->   [+-] term re
16    #          | # empty
17    #  term ->   digits rt
18    #  rt   ->   [*/] rt
19    #          | # empty
20
21    my @stack;
22
23    my $regexp = qr{
24        (?&exp)
25
26        (?(DEFINE)
27            (?<exp>    (?&term) (?&re)
28                         (?{ say "exp -> term re" })
29            )
30
31            (?<re>     \s* ([+-]) (?&term) \s* (?{ push @stack, $^N }) (?&re)
32                         (?{ say "re -> [+-] term re" })
33                     | # empty
34                         (?{ say "re -> empty" })
35            )
36
37            (?<term>   ((?&digits))
38                          (?{ # intermediate action
39                              push @stack, $^N
40                          })
41                       (?&rt)
42                          (?{
43                              say "term-> digits($^N) rt";
44                          })
45            )
46
```

```
47            (?<rt>     \s*(?<op>[*/]) (?<num>(?&digits)) \s*
48                                (?{  # intermediate action
49                                     push @stack, $+{num}, $+{op}
50                                 })
51                        (?&rt) # end of <rt> definition
52                                (?{
53                                     say "rt -> [*/] digits($^N) rt"
54                                 })
55                       |  # empty
56                          (?{ say "rt -> empty" })
57            )
58
59            (?<digits>  \s* \d+
60            )
61        )
62    }xms;
63
64    my $input = <>;
65    chomp($input);
66    if ($input =~ $regexp) {
67      say "matches: $&\nStack=(@stack)";
68    }
69    else {
70      say "does not match";
71    }
```

Ejecución:

```
pl@nereida:~/Lperltesting$ ./calc510withnamedpar.pl
5-8/4/2-1
rt -> empty
term-> digits(5) rt
rt -> empty
rt -> [*/] digits(2) rt
rt -> [*/] digits(4) rt
term-> digits(8) rt
rt -> empty
term-> digits(1) rt
re -> empty
re -> [+-] term re
re -> [+-] term re
exp -> term re
matches: 5-8/4/2-1
Stack=(5 8 4 / 2 / - 1 -)
```

Véase También

- El nodo *Backreference variables in code embedded inside Perl 5.10 regexps* en PerlMonks
- El nodo *Strange behavior of @- and @+ in perl5.10 regexps* en PerlMonks

31.8.2. Construyendo el AST con Expresiones Regulares 5.10

Construiremos en esta sección un traductor de infijo a postfijo utilizando una aproximación general: construiremos una representación del Abstract Syntax Tree o AST (véase la sección 33.9 Árbol de Análisis Abstracto para una definición detallada de que es un árbol sintáctico).

Como la aplicación es un poco mas compleja la hemos dividido en varios ficheros. Esta es la estructura:

```
.
|-- ASTandtrans3.pl    # programa principal
|-- BinaryOp.pm        # clases para el manejo de los nodos del AST
|-- testreegxpparen.pl # prueba para Regexp::Paren
`-- Regexp
    `-- Paren.pm       # módulo de extensión de $^N
```

La salida del programa puede ser dividida en tres partes. La primera muestra una antiderivación a derechas inversa:

```
pl@nereida:~/Lperltesting$ ./ASTandtrans3.pl
2*(3-4)
factor -> NUM(2)
factor -> NUM(3)
rt -> empty
term-> factor rt
factor -> NUM(4)
rt -> empty
term-> factor rt
re -> empty
re -> [+-] term re
exp -> term re
factor -> ( exp )
rt -> empty
rt -> [*/] factor rt
term-> factor rt
re -> empty
exp -> term re
matches: 2*(3-4)
```

Que leída de abajo a arriba nos da una derivación a derechas de la cadena `2*(3-4)`:

```
exp => term re => term => factor rt =>
factor [*/](*) factor rt => factor [*/](*) factor =>
factor [*/](*) ( exp ) => factor [*/](*) ( term re ) =>
factor [*/](*) ( term [+-](-) term re ) =>
factor [*/](*) ( term [+-](-) term ) =>
factor [*/](*) ( term [+-](-) factor rt ) =>
factor [*/](*) ( term [+-](-) factor ) =>
factor [*/](*) ( term [+-](-) NUM(4) ) =>
factor [*/](*) ( factor rt [+-](-) NUM(4) ) =>
factor [*/](*) ( factor [+-](-) NUM(4) ) =>
factor [*/](*) ( NUM(3) [+-](-) NUM(4) )  =>
NUM(2) [*/](*) ( NUM(3) [+-](-) NUM(4) )
```

La segunda parte nos muestra la representación del AST para la entrada dada (`2*(3-4)`):

```
AST:
$VAR1 = bless( {
  'left' => bless( {
  'right' => bless( {
    'left' => bless(
    'right' => bless
    'op' => '-'
  }, 'ADD' ),
  'op' => '*'
}, 'MULT' );
```

La última parte de la salida nos muestra la traducción a postfijo de la expresión en infijo suministrada en la entrada (`2*(3-4)`):

```
2 3 4 - *
```

Programa Principal: usando la pila de atributos

La gramática original que consideramos es recursiva a izquierdas:

```
exp    ->   exp [-+] term
          | term
term   ->   term [*/] factor
          | factor
factor ->  \( exp \)
          | \d+
```

aplicando las técnicas explicadas en 33.8.2 es posible transformar la gramática en una no recursiva por la izquierda:

```
exp       ->   term restoexp
restoexp  ->   [-+] term restoexp
             | # vacío
term      ->   term restoterm
restoterm ->   [*/] factor restoterm
             | # vacío
factor    ->   \( exp \)
             | \d+
```

Ahora bien, no basta con transformar la gramática en una equivalente. Lo que tenemos como punto de partida no es una gramática sino un *esquema de traducción* (véase la sección 33.7) que construye el AST asociado con la expresión. Nuestro esquema de traducción conceptual es algo así:

```
exp    ->   exp ([-+]) term       { ADD->new(left => $exp, right => $term, op => $1) }
          | term                  { $term }
term   ->   term ([*/]) factor    { MULT->new(left => $exp, right => $term, op => $1) }
          | factor                { $factor }
factor ->  \( exp \)              { $exp }
          | (\d+)                 { NUM->new(val => $1) }
```

Lo que queremos conseguir un conjunto de acciones semánticas asociadas para gramática no recursiva que sea equivalente a este.

Este es el programa resultante una vez aplicadas las transformaciones. La implementación de la asociación entre símbolos y atributos la realizamos manualmente mediante una pila de atributos:

```
pl@nereida:~/Lperltesting$ cat -n ./ASTandtrans3.pl
 1  #!/usr/local/lib/perl/5.10.1/bin//perl5.10.1
 2  use v5.10;
 3  use strict;
 4  use Regexp::Paren qw{g};
 5  use BinaryOp;
 6
 7  use Data::Dumper;
 8  $Data::Dumper::Indent = 1;
 9
10  # Builds AST
11  my @stack;
12  my $regexp = qr{
13      (?&exp)
14
15      (?(DEFINE)
16          (?<exp>    (?&term) (?&re)
17                       (?{ say "exp -> term re" })
18          )
19
20          (?<re>     \s* ([+-]) (?&term)
21                        (?{  # intermediate action
22                            local our ($ch1, $term) = splice @stack, -2;
23
24                            push @stack, ADD->new( {left => $ch1, right => $term, op => g(1)
25                        })
26                     (?&re)
27                       (?{ say "re -> [+-] term re" })
28                   | # empty
29                       (?{ say "re -> empty" })
30          )
31
32          (?<term>   ((?&factor)) (?&rt)
33                        (?{
34                            say "term-> factor rt";
35                        })
36          )
37
38          (?<rt>     \s*([*/]) (?&factor)
39                         (?{  # intermediate action
40                              local our ($ch1, $ch2) = splice @stack, -2;
41
42                              push @stack, MULT->new({left => $ch1, right => $ch2, op => g(1
43                          })
44                     (?&rt) # end of <rt> definition
45                         (?{
46                              say "rt -> [*/] factor rt"
47                          })
48                   | # empty
49                         (?{ say "rt -> empty" })
50          )
51
52          (?<factor> \s* (\d+)
```

```
53                          (?{
54                             say "factor -> NUM($^N)";
55                             push @stack, bless { 'val' => g(1) }, 'NUM';
56                          })
57                     | \s* \( (?&exp) \s* \)
58                          (?{ say "factor -> ( exp )" })
59          )
60      )
61  }xms;
62
63  my $input = <>;
64  chomp($input);
65  if ($input =~ $regexp) {
66    say "matches: $&";
67    my $ast = pop @stack;
68    say "AST:\n", Dumper $ast;
69
70    say $ast->translate;
71  }
72  else {
73    say "does not match";
74  }
```

Las Clases representando a los AST

Cada nodo del AST es un objeto. La clase del nodo nos dice que tipo de nodo es. Así los nodos de la clase `MULT` agrupan a los nódos de multiplicación y división. Los nodos de la clase `ADD` agrupan a los nódos de suma y resta. El procedimiento general es asociar un método `translate` con cada clase de nodo. De esta forma se logra el polimorfismo necesario: cada clase de nodo sabe como traducirse y el método `translate` de cada clase puede escribirse como

- Obtener los resultados de llamar a `$child->translate` para cada uno de los nodos hijos `$child`. Por ejemplo, si el nodo fuera un nodo `IF_ELSE` de un hipotético lenguaje de programación, se llamaría a los métodos `translate` sobre sus tres hijos `boolexpr`, `ifstatement` y `elsestatement`.
- Combinar los resultados para producir la traducción adecuada del nodo actual.

Es esta combinación la que mas puede cambiar según el tipo de nodo. Así, en el caso de el nodo `IF_ELSE` el seudocódigo para la traducción sería algo parecido a esto:

```
my $self = shift;
my $etiqueta1 = generar_nueva_etiqueta;
my $etiqueta2 = generar_nueva_etiqueta;

my $boolexpr      = $self->boolexpr->translate;
my $ifstatement   = $self->ifstatement->translate,
my $elsestatement = $self->elsestatement->translate,
return << "ENDTRANS";
    $boolexpr
    JUMPZERO $etiqueta1:
    $ifstatement
    JUMP     $etiqueta2:
  $etiqueta1:
    $elsestatement
  $etiqueta2:
ENDTRANS
```

Siguiendo estas observaciones el código de `BinaryOp.pm` queda así:

```
pl@nereida:~/Lperltesting$ cat -n BinaryOp.pm
 1  package BinaryOp;
 2  use strict;
 3  use base qw(Class::Accessor);
 4
 5  BinaryOp->mk_accessors(qw{left right op});
 6
 7  sub translate {
 8    my $self = shift;
 9
10    return $self->left->translate." ".$self->right->translate." ".$self->op;
11  }
12
13  package ADD;
14  use base qw{BinaryOp};
15
16  package MULT;
17  use base qw{BinaryOp};
18
19  package NUM;
20
21  sub translate {
22    my $self = shift;
23
24    return $self->{val};
25  }
26
27  1;
```

Véase también:

- `Class::Accessor`

Accediendo a los paréntesis lejanos: El módulo `Regexp::Paren`

En esta solución utilizamos las variables `@-` y `@+` para construir una función que nos permite acceder a lo que casó con los últimos paréntesis con memoria:

> *Since Perl 5.6.1 the special variables* `@-` *and* `@+` *can functionally replace* `$‘`, `$&` *and* `$’`. *These arrays contain pointers to the beginning and end of each match (see* `perlvar` *for the full story), so they give you essentially the same information, but without the risk of excessive string copying.*

Véanse los párrafos en las páginas 165, 165) y 166 para mas información sobre `@-` y `@+`.
`g(1)` nos retorna lo que casó con el último paréntesis, `g(2)` lo que casó con el penúltimo, etc.

```
pl@nereida:~/Lperltesting$ cat -n Regexp/Paren.pm
 1  package Regexp::Paren;
 2  use strict;
 3
 4  use base qw{Exporter};
 5
 6  our @EXPORT_OK = qw{g};
 7
```

```
 8  sub g {
 9    die "Error in 'Regexp::Paren::g'. Not used inside (?{ code }) construct\n" unless define
10    my $ofs = - shift;
11
12    # Number of parenthesis that matched
13    my $np = @-;
14    die "Error. Illegal 'Regexp::Paren::g' ref inside (?{ code }) construct\n" unless ($np >
15    # $_ contains the string being matched
16    substr($_, $-[$ofs], $+[$np+$ofs] - $-[$ofs])
17  }
18
19  1;
20
21  =head1 NAME
22
23  Regexp::Paren - Extends $^N inside (?{ ... }) constructs
24
25  =head1 SYNOPSIS
26
27    use Regexp::Paren qw{g};
28
29    'abcde' =~ qr{(.)(.)(.)
30                          (?{ print g(1)." ".g(2)." ".g(3)."\n" })                    # c b a
31                  (.)     (?{ print g(1)." ".g(2)." ".g(3)." ".g(4)."\n" })           # d c b
32                  (.)     (?{ print g(1)." ".g(2)." ".g(3)." ".g(4)." ".g(5)."\n" }) # e d c
33                 }x;
34
35    print g(1)." ".g(2)." ".g(3)." ".g(4)." ".g(5)."\n"; # error!
36
37  =head1 DESCRIPTION
38
39  Inside a C<(?{ ... })> construct, C<g(1)> refers to what matched the last parenthesis
40  (like C<$^N>), C<g(2)> refers to the string that matched with the parenthesis before
41  the last, C<g(3)> refers to the string that matched with the parenthesis at distance 3,
42  etc.
43
44  =head1 SEE ALSO
45
46  =over 2
47
48  =item * L<perlre>
49
50  =item * L<perlretut>
51
52  =item * PerlMonks node I<Strange behavior o> C<@-> I<and> C<@+> I<in perl5.10 regexps> L<h
53
54  =item * PerlMonks node I<Backreference variables in code embedded inside Perl 5.10 regexps
55
56  =back
57
58  =head1 AUTHOR
59
60  Casiano Rodriguez-Leon (casiano@ull.es)
```

```
61
62  =head1 ACKNOWLEDGMENTS
63
64  This work has been supported by CEE (FEDER) and the Spanish Ministry of
65  I<Educacion y Ciencia> through I<Plan Nacional I+D+I> number TIN2005-08818-C04-04
66  (ULL::OPLINK project L<http://www.oplink.ull.es/>).
67  Support from Gobierno de Canarias was through GC02210601
68  (I<Grupos Consolidados>).
69  The University of La Laguna has also supported my work in many ways
70  and for many years.
71
72  =head1 LICENCE AND COPYRIGHT
73
74  Copyright (c) 2009- Casiano Rodriguez-Leon (casiano@ull.es). All rights reserved.
75
76  These modules are free software; you can redistribute it and/or
77  modify it under the same terms as Perl itself. See L<perlartistic>.
78
79  This program is distributed in the hope that it will be useful,
80  but WITHOUT ANY WARRANTY; without even the implied warranty of
81  MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

Al ejecutar `perldoc Regexp::Paren` podemos ver la documentación incluida (véase la documentación en `perlpod` y `perlpodspec` así como la sección La Documentación en Perl para mas detalles):

```
NAME
    Regexp::Paren - Extends $^N inside (?{ ... }) constructs

SYNOPSIS
      use Regexp::Paren qw{g};

      'abcde' =~ qr{(.)(.)(.)
                        (?{ print g(1)." ".g(2)." ".g(3)."\n" })                  # c
                  (.)   (?{ print g(1)." ".g(2)." ".g(3)." ".g(4)."\n" })         # d
                  (.)   (?{ print g(1)." ".g(2)." ".g(3)." ".g(4)." ".g(5)."\n" }) # e
                 }x;

      print g(1)." ".g(2)." ".g(3)." ".g(4)." ".g(5)."\n"; # error!

DESCRIPTION
    Inside a "(?{ ... })" construct, g(1) refers to what matched the last
    parenthesis (like $^N), g(2) refers to the string that matched with the
    parenthesis before the last, g(3) refers to the string that matched with
    the parenthesis at distance 3, etc.

SEE ALSO
    * perlre
    * perlretut
    * PerlMonks node *Strange behavior o* "@-" *and* "@+" *in perl5.10
      regexps* <http://www.perlmonks.org/?node_id=794736>
    * PerlMonks node *Backreference variables in code embedded inside Perl
      5.10 regexps* <http://www.perlmonks.org/?node_id=794424>

AUTHOR
    Casiano Rodriguez-Leon (casiano@ull.es)

ACKNOWLEDGMENTS
    This work has been supported by CEE (FEDER) and the Spanish Ministry of
    *Educacion y Ciencia* through *Plan Nacional I+D+I* number
    TIN2005-08818-C04-04 (ULL::OPLINK project <http://www.oplink.ull.es/>).
    Support from Gobierno de Canarias was through GC02210601 (*Grupos
    Consolidados*). The University of La Laguna has also supported my work
    in many ways and for many years.

LICENCE AND COPYRIGHT
    Copyright (c) 2009- Casiano Rodriguez-Leon (casiano@ull.es). All rights
```

31.9. Práctica: Traducción de `invitation` a HTML

Esta práctica es continuación de la práctica *un lenguaje para componer invitaciones* especificada en la sección 31.7.

El objetivo es traducir la entrada escrita en el lenguaje de invitaciones a HTML. La traducción del

ejemplo anterior debería ser parecida a esta:

```
pl@nereida:~/Lpl0910/Practicas/161009/src$ cat -n invitat
 1  <?xml version="1.0"?>
 2  <!DOCTYPE invitation SYSTEM "invitation.dtd">
 3  <invitation>
 4  <!-- ++++ The header part of the document ++++ -->
 5  <front>
 6  <to>Anna, Bernard, Didier, Johanna</to>
 7  <date>Next Friday Evening at 8 pm</date>
 8  <where>The Web Cafe</where>
 9  <why>My first XML baby</why>
10  </front>
11  <!-- +++++ The main part of the document +++++ -->
12  <body>
13  <par>
14  I would like to invite you all to celebrate
15  the birth of <emph>Invitation</emph>, my
16  first XML document child.
17  </par>
18  <par>
19  Please do your best to come and join me next Friday
20  evening. And, do not forget to bring your friends.
21  </par>
22  <par>
23  I <emph>really</emph> look forward to see you soon!
24  </par>
25  </body>
26  <!-- +++ The closing part of the document ++++ -->
27  <back>
28  <signature>Michel</signature>
29  </back>
30  </invitation>
```

Para ver el resultado en su navegador visite el fichero invitation.html

Su programa deberá producir un Abstract Syntax Tree. Los nodos serán objetos. Cada clase (`FRONT`, `TO`, etc.) deberá de disponer de un método `translate`.

Para simplificar el proceso de traducción a `HTML` se sugiere utilizar una hoja de estilo parecida a la siguiente (tomada de la seción 7.4.4 del citado libro de Goosens):

```
pl@nereida:~/Lpl0910/Practicas/161009/src$ cat -n invit.css
 1  /* CSS stylesheet for invitation1 in HTML */
 2  BODY {margin-top: 1em;     /* global page parameters */
 3        margin-bottom: 1em;
 4        margin-left: 1em;
 5        margin-right: 1em;
 6        font-family: serif;
 7        line-height: 1.1;
 8        color: black;
 9  }
```

```
10  H1   {text-align: center;  /* for global title   */
11        font-size: x-large;
12  }
13  P    {text-align: justify; /* paragraphs in body */
14        margin-top: 1em;
15  }
16  TABLE { border-width: 0pt }
17  TBODY { border-width: 0pt }
18  TD[class="front"] {        /* table data in front matter */
19        text-align: left;
20        font-weight: bold;
21  }
22  TD.front {        /* table data in front matter */
23        text-align: left;
24        font-weight: bold;
25  }
26  EM   {font-style: italic;  /* emphasis in body   */
27  }
28  P.signature {     /* signature          */
29        text-align: right;
30        font-weight: bold;
31  }
```

Véase también:

- The LaTex Web Companion
- Examples from The LaTeX Web Companion (véanse los subdirectorios correspondietnes a los capítulos 6 y 7)
- CSS Tutorial
- Edición extremadamente simple de HTML
- Perl-XML Frequently Asked Questions

31.10. Análisis Sintáctico con `Regexp::Grammars`

El módulo `Regexp::Grammars` escrito por Damian Conway extiende las expresiones regulares Perl con la capacidad de generar representaciones del árbol de análisis sintáctico abstracto y obviando la necesidad de explicitar los blancos. El módulo necesita para funcionar una versión de Perl superior o igual a la 5.10.

31.10.1. Introducción

El Problema

La documentación de `Regexp::Grammars` establece cual es el problema que aborda el módulo:

> *...Perl5.10 makes possible to use regexes to recognize complex, hierarchical–and even recursive–textual structures. The problem is that Perl 5.10 doesn't provide any support for extracting that hierarchical data into nested data structures. In other words, using Perl 5.10 you can match complex data, but not parse it into an internally useful form.*
>
> *An additional problem when using Perl 5.10 regexes to match complex data formats is that you have to make sure you remember to insert whitespace- matching constructs (such as* `\s*`*) at every possible position where the data might contain ignorable whitespace. This reduces the readability of such patterns, and increases the chance of errors (typically caused by overlooking a location where whitespace might appear).*

Una solución: `Regexp::Grammars`

The `Regexp::Grammars` *module solves both those problems.*

If you import the module into a particular lexical scope, it preprocesses any regex in that scope, so as to implement a number of extensions to the standard Perl 5.10 regex syntax. These extensions simplify the task of defining and calling subrules within a grammar, and allow those subrule calls to capture and retain the components of they match in a proper hierarchical manner.

La sintaxis de una expresión regular `Regexp::Grammars`

Las expresiones regulares `Regexp::Grammars` aumentan las regexp Perl 5.10. La sintáxis se expande y se modifica:

A `Regexp::Grammars` *specification consists of a pattern (which may include both standard Perl 5.10 regex syntax, as well as special* `Regexp::Grammars` *directives), followed by one or more rule or token definitions.*

Sigue un ejemplo:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n balanced_brackets.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5
     6  my $rbb = do {
     7      use Regexp::Grammars;
     8      qr{
     9        (<pp>)
    10
    11        <rule: pp>   \( (?: [^()]*+ | <escape> | <pp> )* \)
    12
    13        <token: escape> \\.
    14
    15      }xs;
    16  };
    17
    18  while (my $input = <>) {
    19      while ($input =~ m{$rbb}g) {
    20          say("matches: <$&>");
    21          say Dumper \%/;
    22      }
    23  }
```

Note that there is no need to explicitly place `\s*` *subpatterns throughout the rules; that is taken care of automatically.*

. . .

The initial pattern (`(<pp>)`*) acts like the* top *rule of the grammar, and must be matched completely for the grammar to match.*

The rules and tokens are declarations only and they are not directly matched. Instead, they act like subroutines, and are invoked by name from the initial pattern (or from within a rule or token).

Each rule or token extends from the directive that introduces it up to either the next rule or token directive, or (in the case of the final rule or token) to the end of the grammar.

El hash %/: Una representación del AST Al ejecutar el programa anterior con entrada `(2*(3+5))*4+(2-3)` produce:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 balanced_brackets.pl
(2*(3+5))*4+(2-3)
matches: <(2*(3+5))>
$VAR1 = {
          '' => '(2*(3+5))',
          'pp' => {
                    '' => '(2*(3+5))',
                    'pp' => '(3+5)'
                  }
        };

matches: <(2-3)>
$VAR1 = {
          '' => '(2-3)',
          'pp' => '(2-3)'
        };
```

> *Each rule calls the subrules specified within it, and then return a hash containing whatever result each of those subrules returned, with each result indexed by the subrule's name.*
>
> *In this way, each level of the hierarchical regex can generate hashes recording everything its own subrules matched, so when the entire pattern matches, it produces a tree of nested hashes that represent the structured data the pattern matched.*
>
> *...*
>
> *In addition each result-hash has one extra key: the empty string. The value for this key is whatever string the entire subrule call matched.*

Diferencias entre `token` y `rule`

> *The difference between a token and a rule is that a token treats any whitespace within it exactly as a normal Perl regular expression would. That is, a sequence of whitespace in a token is ignored if the `/x` modifier is in effect, or else matches the same literal sequence of whitespace characters (if `/x` is not in effect).*

En el ejemplo anterior el comportamiento es el mismo si se reescribe la regla para el token `escape` como:

```
    13        <rule: escape> \\.
```

En este otro ejemplo mostramos que la diferencia entre token y rule es significativa:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n tokenvsrule.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5
     6  my $rbb = do {
     7      use Regexp::Grammars;
     8      qr{
     9        <s>
    10
    11        <rule: s> <a> <c>
```

```
12
13        <rule: c>  c d
14
15        <token: a>  a b
16
17      }xs;
18  };
19
20  while (my $input = <>) {
21      if ($input =~ m{$rbb}) {
22          say("matches: <$&>");
23          say Dumper \%/;
24      }
25      else {
26          say "Does not match";
27      }
28  }
```

Al ejecutar este programa vemos la diferencia en la interpretación de los blancos:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 tokenvsrule.pl
ab c d
matches: <ab c d>
$VAR1 = {
          '' => 'ab c d',
          's' => {
                   '' => 'ab c d',
                   'c' => 'c d',
                   'a' => 'ab'
                 }
        };

a b c d
Does not match
ab cd
matches: <ab cd>
$VAR1 = {
          '' => 'ab cd',
          's' => {
                   '' => 'ab cd',
                   'c' => 'cd',
                   'a' => 'ab'
                 }
        };
```

Obsérvese como la entrada `a b c d` es rechazada mientras que la entrada `ab c d` es aceptada.

Redefinición de los espacios en blanco

> *In a rule, any sequence of whitespace (except those at the very start and the very end of the rule) is treated as matching the implicit subrule* `<.ws>`*, which is automatically predefined to match optional whitespace (i.e.* `\s*`*).*
>
> *You can explicitly define a* `<ws>` *token to change that default behaviour. For example, you could alter the definition of whitespace to include Perlish comments, by adding an explicit* `<token: ws>`*:*

```
<token: ws>
   (?: \s+ | #[^\n]* )*
```

But be careful not to define `<ws>` *as a rule, as this will lead to all kinds of infinitely recursive unpleasantness.*

El siguiente ejemplo ilustra como redefinir `<ws>`:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n tokenvsruleandws.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8      no warnings 'uninitialized';
 9      qr{
10        <s>
11
12        <token: ws> (?: \s+ | /\* .*? \*/)*+
13
14        <rule: s> <a> <c>
15
16        <rule: c>  c d
17
18        <token: a>  a b
19
20      }xs;
21  };
22
23  while (my $input = <>) {
24      if ($input =~ m{$rbb}) {
25          say Dumper \%/;
26      }
27      else {
28          say "Does not match";
29      }
30  }
```

Ahora podemos introducir comentarios en la entrada:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 -w tokenvsruleandws.pl
ab /* 1 */ c d
$VAR1 = {
          '' => 'ab /* 1 */ c d',
          's' => {
                   '' => 'ab /* 1 */ c d',
                   'c' => 'c d',
                   'a' => 'ab'
                 }
        };
```

Llamando a las subreglas

To invoke a rule to match at any point, just enclose the rule's name in angle brackets (like in Perl 6). There must be no space between the opening bracket and the rulename. For example:

```
qr{
    file:             # Match literal sequence 'f' 'i' 'l' 'e' ':'
    <name>            # Call <rule: name>
    <options>?        # Call <rule: options> (it's okay if it fails)

    <rule: name>
        # etc.
}x;
```

If you need to match a literal pattern that would otherwise look like a subrule call, just backslash-escape the leading angle:

```
qr{
    file:             # Match literal sequence 'f' 'i' 'l' 'e' ':'
    \<name>           # Match literal sequence '<' 'n' 'a' 'm' 'e' '>'
    <options>?        # Call <rule: options> (it's okay if it fails)

    <rule: name>
        # etc.
}x;
```

El siguiente programa ilustra algunos puntos discutidos en la cita anterior:

```
casiano@millo:~/src/perl/regexp-grammar-examples$ cat -n badbracket.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8      qr{
 9        (<pp>)
10
11        <rule: pp>   \( (?: <b  > | \< | < escape> | <pp> )* \)
12
13        <token: b  > b
14
15        <token: escape> \\.
16
17      }xs;
18  };
19
20  while (my $input = <>) {
21      while ($input =~ m{$rbb}g) {
22          say("matches: <$&>");
23          say Dumper \%/;
24      }
25  }
```

Obsérvense los blancos en `< escape>` y en `<token: b  > b`. Pese a ello el programa funciona:

```
casiano@millo:~/src/perl/regexp-grammar-examples$ perl5.10.1 badbracket.pl
(\(\))
matches: <(\(\))>
$VAR1 = {
          '' => '(\\(\\))',
          'pp' => {
                    '' => '(\\(\\))',
                    'escape' => '\\)'
                  }
        };

(b)
matches: <(b)>
$VAR1 = {
          '' => '(b)',
          'pp' => {
                    '' => '(b)',
                    'b' => 'b'
                  }
        };

(<)
matches: <(<)>
$VAR1 = {
          '' => '(<)',
          'pp' => '(<)'
        };

(c)

casiano@millo:
```

Eliminación del anidamiento de ramas unarias en %/

> *. . . Note, however, that if the result-hash at any level contains only the empty-string key (i.e. the subrule did not call any sub-subrules or save any of their nested result-hashes), then the hash is unpacked and just the matched substring itself if returned.*
>
> *For example, if* `<rule: sentence>` *had been defined:*
>
> ```
> <rule: sentence>
>     I see dead people
> ```
>
> *then a successful call to the rule would only add:*
>
> ```
> sentence => 'I see dead people'
> ```
>
> *to the current result-hash.*
>
> *This is a useful feature because it prevents a series of nested subrule calls from producing very unwieldy data structures. For example, without this automatic unpacking, even the simple earlier example:*
>
> ```
> <rule: sentence>
>     <noun> <verb> <object>
> ```

would produce something needlessly complex, such as:

```
sentence => {
    ""     => 'I saw a dog',
    noun   => {
        "" => 'I',
    },
    verb   => {
        "" => 'saw',
    },
    object => {
        ""      => 'a dog',
        article => {
            "" => 'a',
        },
        noun    => {
            "" => 'dog',
        },
    },
}
```

El siguiente ejemplo ilustra este punto:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n unaryproductions.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8      qr{
 9        <s>
10
11        <rule: s> <noun> <verb> <object>
12
13        <token: noun> he | she | Peter | Jane
14
15        <token: verb> saw | sees
16
17        <token: object> a\s+dog | a\s+cat
18
19      }x;
20  };
21
22  while (my $input = <>) {
23      while ($input =~ m{$rbb}g) {
24          say("matches: <$&>");
25          say Dumper \%/;
26      }
27  }
```

Sigue una ejecución del programa anterior:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 unaryproductions.pl
```

```
he saw a dog
matches: <he saw a dog>
$VAR1 = {
          '' => 'he saw a dog',
          's' => {
                    '' => 'he saw a dog',
                    'object' => 'a dog',
                    'verb' => 'saw',
                    'noun' => 'he'
                  }
        };

Jane sees a cat
matches: <Jane sees a cat>
$VAR1 = {
          '' => 'Jane sees a cat',
          's' => {
                    '' => 'Jane sees a cat',
                    'object' => 'a cat',
                    'verb' => 'sees',
                    'noun' => 'Jane'
                  }
        };
```

Ámbito de uso de `Regexp::Grammars`

Cuando se usa `Regexp::Grammars` como parte de un programa que utiliza otras regexes hay que evitar que `Regexp::Grammars` procese las mismas. `Regexp::Grammars` reescribe las expresiones regulares durante la fase de preproceso. Esta por ello presenta las mismas limitaciones que cualquier otra forma de 'source filtering' (véase `perlfilter`). Por ello es una buena idea declarar la gramática en un bloque `do` restringiendo de esta forma el ámbito de acción del módulo.

```
 5  my $calculator = do{
 6      use Regexp::Grammars;
 7      qr{
 .          ........
28      }xms
29  };
```

31.10.2. Objetos

> *When a grammar has parsed successfully, the* `%/` *variable will contain a series of nested hashes (and possibly arrays) representing the hierarchical structure of the parsed data.*
>
> *Typically, the next step is to walk that tree, extracting or converting or otherwise processing that information. If the tree has nodes of many different types, it can be difficult to build a recursive subroutine that can navigate it easily.*
>
> *A much cleaner solution is possible if the nodes of the tree are proper objects. In that case, you just define a* `trasnlate()` *method for each of the classes, and have every node call that method on each of its children. The chain of* `translate()` *calls would cascade down the nodes of the tree, each one invoking the appropriate* `translate()` *method according to the type of node encountered.*
>
> *The only problem is that, by default,* `Regexp::Grammars` *returns a tree of plain-old hashes, not* `Class::Whatever` *objects. Fortunately, it's easy to request that the result has-*

hes be automatically blessed into the appropriate classes, using the `<objrule:...>` *and* `<objtoken:...>` *directives.*

These directives are identical to the `<rule:...>` *and* `<token:...>` *directives (respectively), except that the rule or token they create will also bless the hash it normally returns, converting it to an object of a class whose name is the same as the rule or token itself.*

For example:

```
<objrule: Element>
    # ...Defines a rule that can be called as <Element>
    # ...and which returns a hash-based Element object
```

The `IDENTIFIER` *of the rule or token may also be fully qualified. In such cases, the rule or token is defined using only the final* short name*, but the result object is blessed using the fully qualified* long name*. For example:*

```
<objrule: LaTeX::Element>
    # ...Defines a rule that can be called as <Element>
    # ...and which returns a hash-based LaTeX::Element object
```

This can be useful to ensure that returned objects don't collide with other namespaces in your program.

Note that you can freely mix object-returning and plain-old-hash-returning rules and tokens within a single grammar, though you have to be careful not to subsequently try to call a method on any of the unblessed nodes.

31.10.3. Renombrando los resultados de una subregla

Nombre de la regla versus Nombre del Resultado

No siempre el nombre de la regla es el mas apropiado para ser el nombre del resultado:

It is not always convenient to have subrule results stored under the same name as the rule itself. Rule names should be optimized for understanding the behaviour of the parser, whereas result names should be optimized for understanding the structure of the data. Often those two goals are identical, but not always; sometimes rule names need to describe what the data looks like, while result names need to describe what the data means.

Colisión de nombres de reglas

For example, sometimes you need to call the same rule twice, to match two syntactically identical components whose positions give then semantically distinct meanings:

```
<rule: copy_cmd>
    copy <file> <file>
```

The problem here is that, if the second call to `<file>` *succeeds, its result-hash will be stored under the key* `file`*, clobbering the data that was returned from the first call to* `<file>`*.*

Aliasing

To avoid such problems, `Regexp::Grammars` *allows you to alias any subrule call, so that it is still invoked by the original name, but its result-hash is stored under a different key. The syntax for that is:* `<alias=rulename>`*. For example:*

```
<rule: copy_cmd>
    copy <from=file> <to=file>
```

Here, `<rule: file>` *is called twice, with the first result-hash being stored under the key* `from`*, and the second result-hash being stored under the key* `to`*.*

Note, however, that the alias before the `=` *must be a proper identifier (i.e. a letter or underscore, followed by letters, digits, and/or underscores). Aliases that start with an underscore and aliases named* `MATCH` *have special meaning.*

Normalización de los resultados mediante aliasing

Aliases can also be useful for normalizing data that may appear in different formats and sequences. For example:

```
<rule: copy_cmd>
    copy <from=file>        <to=file>
  | dup    <to=file>  as  <from=file>
  |      <from=file>  ->    <to=file>
  |        <to=file>  <-  <from=file>
```

Here, regardless of which order the old and new files are specified, the result-hash always gets:

```
copy_cmd => {
    from => 'oldfile',
      to => 'newfile',
}
```

Ejemplo

El siguiente programa ilustra los comentarios de la documentación:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n copygrammar.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5
     6  my $rbb = do {
     7      use Regexp::Grammars;
     8      qr{
     9        <copy_cmd>
    10
    11        <rule: copy_cmd>
    12              copy <from=file> <to=file>
    13          |   <from=file> ->   <to=file>
    14          |   <to=file>   <- <from=file>
    15
    16        <token: file> [\w./\\]+
    17      }x;
    18  };
    19
    20  while (my $input = <>) {
    21      while ($input =~ m{$rbb}g) {
    22          say("matches: <$&>");
    23          say Dumper \%/;
    24      }
    25  }
```

Cuando lo ejecutamos obtenemos:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 copygrammar.pl
copy a b
matches: <copy a b>
$VAR1 = {
          '' => 'copy a b',
          'copy_cmd' => {
                          '' => 'copy a b',
                          'to' => 'b',
                          'from' => 'a'
                        }
        };

b <- a
matches: <b <- a>
$VAR1 = {
          '' => 'b <- a',
          'copy_cmd' => {
                          '' => 'b <- a',
                          'to' => 'b',
                          'from' => 'a'
                        }
        };

a -> b
matches: <a -> b>
$VAR1 = {
          '' => 'a -> b',
          'copy_cmd' => {
                          '' => 'a -> b',
                          'to' => 'b',
                          'from' => 'a'
                        }
        };
```

31.10.4. Listas

El operador de cierre positivo

If a subrule call is quantified with a repetition specifier:

```
<rule: file_sequence>
    <file>+
```

then each repeated match overwrites the corresponding entry in the surrounding rule's result-hash, so only the result of the final repetition will be retained. That is, if the above example matched the string `foo.pl bar.py baz.php`*, then the result-hash would contain:*

```
file_sequence {
    ""   => 'foo.pl bar.py baz.php',
    file => 'baz.php',
}
```

Operadores de listas y espacios en blanco

Existe un caveat con el uso de los operadores de repetición y el manejo de los blancos. Véase el siguiente programa:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n numbers3.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8
 9      qr{
10        <numbers>
11
12        <rule: numbers>
13          (<number>)+
14
15        <token: number> \s*\d+
16      }xms;
17  };
18
19  while (my $input = <>) {
20      if ($input =~ m{$rbb}) {
21          say("matches: <$&>");
22          say Dumper \%/;
23      }
24  }
```

Obsérvese el uso explícito de espacios `\s*\d+` en la definición de `number`.

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5_10_1 numbers3.pl
1 2 3 4
matches: <1 2 3 4>
$VAR1 = {
          '' => '1 2 3 4',
          'numbers' => {
                         '' => '1 2 3 4',
                         'number' => ' 4'
                       }
        };
```

Si se eliminan los blancos de la definición de `number`:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n numbers.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5
     6  my $rbb = do {
     7      use Regexp::Grammars;
     8
```

```
 9      qr{
10        <numbers>
11
12        <rule: numbers>
13          (<number>)+
14
15        <token: number> \d+
16      }xms;
17  };
18
19  while (my $input = <>) {
20      if ($input =~ m{$rbb}) {
21          say("matches: <$&>");
22          say Dumper \%/;
23      }
24  }
```

se obtiene una conducta que puede sorprender:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 numbers.pl
12 34 56
matches: <12>
$VAR1 = {
          '' => '12',
          'numbers' => {
                         '' => '12',
                         'number' => '12'
                       }
        };
```

La explicación está en la documentación: véase la sección Grammar Syntax:

> `<rule: IDENTIFIER>`
> *Define a rule whose name is specified by the supplied identifier.*
> *Everything following the* `<rule:...>` *directive (up to the next* `<rule:...>` *or* `<token:...>` *directive) is treated as part of the rule being defined.*
> Any whitespace in the rule is replaced by a call to the `<.ws>` *subrule (which defaults to matching* `\s*`*, but may be explicitly redefined).*

También podríamos haber resuelto el problema introduciendo un blanco explícito dentro del cierre positivo:

```
<rule: numbers>
  (<number> )+

<token: number> \d+
```

Una Solución al problema de recordar los resultados de una lista: El uso de brackets

> *Usually, that's not the desired outcome, so* `Regexp::Grammars` *provides another mechanism by which to call a subrule; one that saves all repetitions of its results.*
> *A regular subrule call consists of the rule's name surrounded by angle brackets. If, instead, you surround the rule's name with* `<[...]>` *(angle and square brackets) like so:*
>
> ```
> <rule: file_sequence>
>     <[file]>+
> ```

then the rule is invoked in exactly the same way, but the result of that submatch is pushed onto an array nested inside the appropriate result-hash entry. In other words, if the above example matched the same `foo.pl bar.py baz.php` *string, the result-hash would contain:*

```
file_sequence {
    ""   => 'foo.pl bar.py baz.php',
    file => [ 'foo.pl', 'bar.py', 'baz.php' ],
}
```

Teniendo en cuenta lo dicho anteriormente sobre los blancos dentro de los cuantificadores, es necesario introducir blancos dentro del operador de repetición:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n numbers4.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5
     6  my $rbb = do {
     7      use Regexp::Grammars;
     8
     9      qr{
    10        <numbers>
    11
    12        <rule: numbers>
    13          (?:  <[number]> )+
    14
    15        <token: number> \d+
    16      }xms;
    17  };
    18
    19  while (my $input = <>) {
    20      if ($input =~ m{$rbb}) {
    21          say("matches: <$&>");
    22          say Dumper \%/;
    23      }
    24  }
```

Al ejecutar este programa obtenemos:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5_10_1 numbers4.pl
1 2 3 4
matches: <1 2 3 4
>
$VAR1 = {
          '' => '1 2 3 4
',
          'numbers' => {
                         '' => '1 2 3 4
',
                         'number' => [ '1', '2', '3', '4' ]
                       }
        };
```

Otra forma de resolver las colisiones de nombres: salvarlos en una lista

> *This* listifying subrule call *can also be useful for non-repeated subrule calls, if the same subrule is invoked in several places in a grammar. For example if a cmdline option could be given either one or two values, you might parse it:*
>
> ```
> <rule: size_option>
>     -size <[size]> (?: x <[size]> )?
> ```
>
> *The result-hash entry for* `size` *would then always contain an array, with either one or two elements, depending on the input being parsed.*

Sigue un ejemplo:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n sizes.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8
 9      qr{
10        <command>
11
12        <rule: command> ls <size_option>
13
14        <rule: size_option>
15            -size <[size]> (?: x <[size]> )?
16
17        <token: size> \d+
18      }x;
19  };
20
21  while (my $input = <>) {
22      while ($input =~ m{$rbb}g) {
23          say("matches: <$&>");
24          say Dumper \%/;
25      }
26  }
```

Veamos su comportamiento con diferentes entradas:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 sizes.pl
ls -size 4
matches: <ls -size 4
>
$VAR1 = {
          '' => 'ls -size 4
',
          'command' => {
                         'size_option' => {
                                            '' => '-size 4
',
```

```
                                          'size' => [ '4' ]
                                        },
                        '' => 'ls -size 4
'
                      }
        };

ls -size 2x8
matches: <ls -size 2x8
>
$VAR1 = {
          '' => 'ls -size 2x8
',
          'command' => {
                        'size_option' => {
                                          '' => '-size 2x8
',
                                          'size' => [ '2', '8' ]
                                        },
                        '' => 'ls -size 2x8
'
                      }
        };
```

Aliasing de listas

Listifying subrules can also be given aliases, just like ordinary subrules. The alias is always specified inside the square brackets:

```
 <rule: size_option>
     -size <[size=pos_integer]> (?: x <[size=pos_integer]> )?
```

Here, the sizes are parsed using the `pos_integer` *rule, but saved in the result-hash in an array under the key* `size`*.*

Sigue un ejemplo:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n aliasedsizes.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8
 9      qr{
10        <command>
11
12        <rule: command> ls <size_option>
13
14        <rule: size_option>
15            -size <[size=int]> (?: x <[size=int]> )?
16
```

```
17        <token: int> \d+
18      }x;
19  };
20
21  while (my $input = <>) {
22      while ($input =~ m{$rbb}g) {
23          say("matches: <$&>");
24          say Dumper \%/;
25      }
26  }
```

Veamos el resultado de una ejecución:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 aliasedsizes.pl
ls -size 2x4
matches: <ls -size 2x4
>
$VAR1 = {
          '' => 'ls -size 2x4
',
          'command' => {
                         'size_option' => {
                                            '' => '-size 2x4
',
                                            'size' => [
                                                        '2',
                                                        '4'
                                                      ]
                                          },
                         '' => 'ls -size 2x4
'
                       }
        };
```

Caveat: Cierres y Warnings

En este ejemplo aparece `<number>+` sin corchetes ni paréntesis:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n numbers5.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5
     6  my $rbb = do {
     7      use Regexp::Grammars;
     8
     9      qr{
    10        <numbers>
    11
    12        <rule: numbers>
    13          <number>+
    14
    15        <token: number> \d+
    16      }xms;
```

```
17  };
18
19  while (my $input = <>) {
20      if ($input =~ m{$rbb}) {
21          say("matches: <$&>");
22          say Dumper \%/;
23      }
24  }
```

Este programa produce un mensaje de advertencia:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 numbers5.pl
  warn | Repeated subrule <number>+ will only capture its final match
       | (Did you mean <[number]>+ instead?)
       |
```

Si se quiere evitar el mensaje y se está dispuesto a asumir la pérdida de los valores asociados con los elementos de la lista se deberán poner el operando entre paréntesis (con o sin memoria).

Esto es lo que dice la documentación sobre este warning:

> `Repeated subrule <rule> will only capture its final match`
> *You specified a subrule call with a repetition qualifier, such as:*
>
> `<ListElem>*`
>
> *or:*
>
> `<ListElem>+`
>
> *Because each subrule call saves its result in a hash entry of the same name, each repeated match will overwrite the previous ones, so only the last match will ultimately be saved. If you want to save all the matches, you need to tell* `Regexp::Grammars` *to save the sequence of results as a nested array within the hash entry, like so:*
>
> `<[ListElem]>*`
>
> *or:*
>
> `<[ListElem]>+`
>
> *If you really did intend to throw away every result but the final one, you can silence the warning by placing the subrule call inside any kind of parentheses. For example:*
>
> `(<ListElem>)*`
>
> *or:*
>
> `(?: <ListElem> )+`

31.10.5. Pseudo sub-reglas

Subpatrones

Aliases can also be given to standard Perl subpatterns, as well as to code blocks within a regex. The syntax for subpatterns is:

```
<ALIAS= (SUBPATTERN) >
```

In other words, the syntax is exactly like an aliased subrule call, except that the rule name is replaced with a set of parentheses containing the subpattern. Any parentheses–capturing or non-capturing–will do.

The effect of aliasing a standard subpattern is to cause whatever that subpattern matches to be saved in the result-hash, using the alias as its key. For example:

```
<rule: file_command>

    <cmd=(mv|cp|ln)>  <from=file>  <to=file>
```

Here, the `<cmd=(mv|cp|ln)>` *is treated exactly like a regular* `(mv|cp|ln)`*, but whatever substring it matches is saved in the result-hash under the key* `'cmd'`*.*

Sigue un ejemplo:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n subpattern.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8
 9      qr{
10          <file_command>
11
12          <rule: file_command>
13
14          <cmd=(mv|cp|ln)>  <from=([\w./]+)>  <to=([\w./]+)>
15
16      }x;
17  };
18
19  while (my $input = <>) {
20      while ($input =~ m{$rbb}g) {
21          say("matches: <$&>");
22          say Dumper \%/;
23      }
24  }
```

y una ejecución:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 subpattern.pl
mv a b
matches: <mv a b>
$VAR1 = {
```

```
          '' => 'mv a b',
          'file_command' => {
                               '' => 'mv a b',
                               'to' => 'b',
                               'cmd' => 'mv',
                               'from' => 'a'
                             }
        };

cp c d
matches: <cp c d>
$VAR1 = {
          '' => 'cp c d',
          'file_command' => {
                               '' => 'cp c d',
                               'to' => 'd',
                               'cmd' => 'cp',
                               'from' => 'c'
                             }
        }
```

Bloques de código

The syntax for aliasing code blocks is:

```
<ALIAS= (?{ your($code->here) }) >
```

Note, however, that the code block must be specified in the standard Perl 5.10 regex notation: `(?{...})`. *A common mistake is to write:*

```
<ALIAS= { your($code->here } >
```

instead, which will attempt to interpolate `$code` *before the regex is even compiled, as such variables are only* protected *from interpolation inside a* `(?{...})`.

When correctly specified, this construct executes the code in the block and saves the result of that execution in the result-hash, using the alias as its key. Aliased code blocks are useful for adding semantic information based on which branch of a rule is executed. For example, consider the `copy_cmd` *alternatives shown earlier:*

```
<rule: copy_cmd>
    copy <from=file>        <to=file>
  | dup    <to=file>  as  <from=file>
  |      <from=file>  ->    <to=file>
  |        <to=file>  <-  <from=file>
```

Using aliased code blocks, you could add an extra field to the result- hash to describe which form of the command was detected, like so:

```
<rule: copy_cmd>
    copy <from=file>        <to=file>  <type=(?{ 'std' })>
  | dup    <to=file>  as  <from=file>  <type=(?{ 'rev' })>
  |      <from=file>  ->    <to=file>  <type=(?{ 'fwd' })>
  |        <to=file>  <-  <from=file>  <type=(?{ 'bwd' })>
```

Now, if the rule matched, the result-hash would contain something like:

```
        copy_cmd => {
            from => 'oldfile',
              to => 'newfile',
            type => 'fwd',
        }
```

El siguiente ejemplo ilustra lo dicho en la documentación. En la línea 15 hemos introducido una regla para el control de errores[8]:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n aliasedcodeblock2.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5
 6  my $rbb = do {
 7      use Regexp::Grammars;
 8      qr{
 9        <copy_cmd>
10
11        <rule: copy_cmd>
12              copy (<from=file>) (<to=file>) <type=(?{ 'std' })>
13          |   <from=file> ->   <to=file> <type=(?{ 'fwd' })>
14          |   <to=file>   <- <from=file> <type=(?{ 'bwd' })>
15          |   .+ (?{ die "Syntax error!\n" })
16
17        <token: file> [\w./\\]+
18      }x;
19  };
20
21  while (my $input = <>) {
22      while ($input =~ m{$rbb}g) {
23          say("matches: <$&>");
24          say Dumper \%/;
25      }
26  }
```

La ejecución muestra el comportamiento del programa con tres entradas válidas y una errónea:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 aliasedcodeblock2.pl
copy a b
matches: <copy a b
>
$VAR1 = {
          '' => 'copy a b
',
          'copy_cmd' => {
                          '' => 'copy a b
',
                          'to' => 'b',
                          'from' => 'a',
                          'type' => 'std'
                        }
```

[8] Versión de `Grammar.pm` obtenida por email con las correcciones de Damian

```
        };

b <- a
matches: <b <- a
>
$VAR1 = {
          '' => 'b <- a
',
          'copy_cmd' => {
                          '' => 'b <- a
',
                          'to' => 'b',
                          'from' => 'a',
                          'type' => 'bwd'
                        }
        };

a -> b
matches: <a -> b
>
$VAR1 = {
          '' => 'a -> b
',
          'copy_cmd' => {
                          '' => 'a -> b
',
                          'to' => 'b',
                          'from' => 'a',
                          'type' => 'fwd'
                        }
        };

cp a b
Syntax error!
```

Pseudo subreglas y depuración

Note that, in addition to the semantics described above, aliased subpatterns and code blocks also become visible to `Regexp::Grammars` *integrated debugger (see Debugging).*

31.10.6. Llamadas a subreglas desmemoriadas

By default, every subrule call saves its result into the result-hash, either under its own name, or under an alias.

However, sometimes you may want to refactor some literal part of a rule into one or more subrules, without having those submatches added to the result-hash. The syntax for calling a subrule, but ignoring its return value is:

```
<.SUBRULE>
```

(which is stolen directly from Perl 6).

For example, you may prefer to rewrite a rule such as:

```
<rule: paren_pair>
```

```
\(
    (?: <escape> | <paren_pair> | <brace_pair> | [^()] )*
\)
```

without any literal matching, like so:

```
<rule: paren_pair>

    <.left_paren>
        (?: <escape> | <paren_pair> | <brace_pair> | <.non_paren> )*
    <.right_paren>

<token: left_paren>   \(
<token: right_paren>  \)
<token: non_paren>    [^()]
```

Moreover, as the individual components inside the parentheses probably aren't being captured for any useful purpose either, you could further optimize that to:

```
<rule: paren_pair>

    <.left_paren>
        (?: <.escape> | <.paren_pair> | <.brace_pair> | <.non_paren> )*
    <.right_paren>
```

Note that you can also use the dot modifier on an aliased subpattern:

```
<.Alias= (SUBPATTERN) >
```

This seemingly contradictory behaviour (of giving a subpattern a name, then deliberately ignoring that name) actually does make sense in one situation. Providing the alias makes the subpattern visible to the debugger, while using the dot stops it from affecting the result-hash. See Debugging non-grammars *for an example of this usage.*

Ejemplo: Números entre comas

Por ejemplo, queremos reconocer listas de números separados por comas. Supongamos también que queremos darle un nombre a la expresión regular de separación. Quizá, aunque no es el caso, porque la expresión regular de separación sea suficientemente compleja. Si no usamos la notación *punto* la coma aparecerá en la estructura:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n numberscomma.pl
     1  use strict;
     2  use warnings;
     3  use 5.010;
     4  use Data::Dumper;
     5  $Data::Dumper::Indent = 1;
     6
     7  my $rbb = do {
     8      use Regexp::Grammars;
     9
    10      qr{
    11        <numbers>
    12
    13        <objrule: numbers>
```

```
    14          <[number]> (<comma> <[number]>)*
    15
    16        <objtoken: number> \s*\d+
    17        <token: comma>  \s*,
    18      }xms;
    19  };
    20
    21  while (my $input = <>) {
    22      if ($input =~ m{$rbb}) {
    23          say("matches: <$&>");
    24          say Dumper \%/;
    25      }
    26  }
```

En efecto, aparece la clave `comma`:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 numberscomma.pl
2, 3, 4
matches: <2, 3, 4>
$VAR1 = {
  '' => '2, 3, 4',
  'numbers' => bless( {
    '' => '2, 3, 4',
    'number' => [
      bless( { '' => '2' }, 'number' ),
      bless( { '' => '3' }, 'number' ),
      bless( { '' => '4' }, 'number' )
    ],
    'comma' => ','
  }, 'numbers' )
};
```

Si cambiamos la llamada a la regla `<comma>` por `<.comma>`

```
pl@nereida:~/Lregexpgrammars/demo$ diff numberscomma.pl numberscomma2.pl
14c14
<         <[number]> (<comma> <[number]>)*
---
>         <[number]> (<.comma> <[number]>)*
```

eliminamos la aparición de la innecesaria clave:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 numberscomma2.pl
2, 3, 4
matches: <2, 3, 4>
$VAR1 = {
  '' => '2, 3, 4',
  'numbers' => bless( {
    '' => '2, 3, 4',
    'number' => [
      bless( { '' => '2' }, 'number' ),
      bless( { '' => '3' }, 'number' ),
      bless( { '' => '4' }, 'number' )
    ]
  }, 'numbers' )
};
```

31.10.7. Destilación del resultado

Destilación manual

`Regexp::Grammars` *also offers full manual control over the distillation process. If you use the reserved word* `MATCH` *as the alias for a subrule call:*

```
<MATCH=filename>
```

or a subpattern match:

```
<MATCH=( \w+ )>
```

or a code block:

```
<MATCH=(?{ 42 })>
```

then the current rule will treat the return value of that subrule, pattern, or code block as its complete result, and return that value instead of the usual result-hash it constructs. This is the case even if the result has other entries that would normally also be returned.

For example, in a rule like:

```
<rule: term>
      <MATCH=literal>
    | <left_paren> <MATCH=expr> <right_paren>
```

The use of `MATCH` *aliases causes the rule to return either whatever* `<literal>` *returns, or whatever* `<expr>` *returns (provided it's between left and right parentheses).*

Note that, in this second case, even though `<left_paren>` *and* `<right_paren>` *are captured to the result-hash, they are not returned, because the* `MATCH` *alias overrides the normal* return the result-hash *semantics and returns only what its associated subrule (i.e.* `<expr>`*) produces.*

El siguiente ejemplo ilustra el uso del alias `MATCH`:

```
$ cat -n demo_calc.pl
 1  #!/usr/local/lib/perl/5.10.1/bin/perl5.10.1
 2  use v5.10;
 3  use warnings;
 4
 5  my $calculator = do{
 6      use Regexp::Grammars;
 7      qr{
 8          <Answer>
 9
10          <rule: Answer>
11              <X=Mult> <Op=([+-])> <Y=Answer>
12            | <MATCH=Mult>
13
14          <rule: Mult>
15              <X=Pow> <Op=([*/%])> <Y=Mult>
16            | <MATCH=Pow>
17
18          <rule: Pow>
19              <X=Term> <Op=(\^)> <Y=Pow>
20            | <MATCH=Term>
```

```
21
22          <rule: Term>
23                 <MATCH=Literal>
24            | \( <MATCH=Answer> \)
25
26          <token: Literal>
27              <MATCH=( [+-]? \d++ (?: \. \d++ )?+ )>
28      }xms
29  };
30
31  while (my $input = <>) {
32      if ($input =~ $calculator) {
33          use Data::Dumper 'Dumper';
34          warn Dumper \%/;
35      }
36  }
```

Veamos una ejecución:

```
$ ./demo_calc.pl
2+3*5
$VAR1 = {
          '' => '2+3*5',
          'Answer' => {
                        '' => '2+3*5',
                        'Op' => '+',
                        'X' => '2',
                        'Y' => {
                                 '' => '3*5',
                                 'Op' => '*',
                                 'X' => '3',
                                 'Y' => '5'
                               }
                      }
        };
4-5-2
$VAR1 = {
          '' => '4-5-2',
          'Answer' => {
                        '' => '4-5-2',
                        'Op' => '-',
                        'X' => '4',
                        'Y' => {
                                 '' => '5-2',
                                 'Op' => '-',
                                 'X' => '5',
                                 'Y' => '2'
                               }
                      }
        };
```

Obsérvese como el árbol construido para la expresión `4-5-2` se hunde a derechas dando lugar a una jerarquía errónea. Para arreglar el problema sería necesario eliminar la recursividad por la izquierda en las reglas correspondientes.

Destilación en el programa

It's also possible to control what a rule returns from within a code block. `Regexp::Grammars` *provides a set of reserved variables that give direct access to the result-hash.*

The result-hash itself can be accessed as `%MATCH` *within any code block inside a rule. For example:*

```
<rule: sum>
    <X=product> \+ <Y=product>
        <MATCH=(?{ $MATCH{X} + $MATCH{Y} })>
```

Here, the rule matches a product (aliased `'X'` *in the result-hash), then a literal* `'+'`*, then another product (aliased to* `'Y'` *in the result-hash). The rule then executes the code block, which accesses the two saved values (as* `$MATCH{X}` *and* `$MATCH{Y}`*), adding them together. Because the block is itself aliased to* `MATCH`*, the sum produced by the block becomes the (only) result of the rule.*

It is also possible to set the rule result from within a code block (instead of aliasing it). The special override *return value is represented by the special variable* `$MATCH`*. So the previous example could be rewritten:*

```
<rule: sum>
    <X=product> \+ <Y=product>
        (?{ $MATCH = $MATCH{X} + $MATCH{Y} })
```

Both forms are identical in effect. Any assignment to `$MATCH` *overrides the normal* return all subrule results *behaviour.*

Assigning to `$MATCH` *directly is particularly handy if the result may not always be* distillable*, for example:*

```
<rule: sum>
    <X=product> \+ <Y=product>
        (?{ if (!ref $MATCH{X} && !ref $MATCH{Y}) {
                # Reduce to sum, if both terms are simple scalars...
                $MATCH = $MATCH{X} + $MATCH{Y};
            }
            else {
                # Return full syntax tree for non-simple case...
                $MATCH{op} = '+';
            }
        })
```

Note that you can also partially override the subrule return behaviour. Normally, the subrule returns the complete text it matched under the empty key *of its result-hash. That is, of course,* `$MATCH{""}`*, so you can override just that behaviour by directly assigning to that entry.*

For example, if you have a rule that matches key/value pairs from a configuration file, you might prefer that any trailing comments not be included in the matched text entry of the rule's result-hash. You could hide such comments like so:

```
<rule: config_line>
    <key> : <value>  <comment>?
        (?{
            # Edit trailing comments out of "matched text" entry...
            $MATCH = "$MATCH{key} : $MATCH{value}";
        })
```

Some more examples of the uses of `$MATCH`*:*

```
<rule: FuncDecl>
  # Keyword  Name               Keep return the name (as a string)...
    func     <Identifier> ;     (?{ $MATCH = $MATCH{'Identifier'} })


<rule: NumList>
  # Numbers in square brackets...
    \[
        ( \d+ (?: , \d+)* )
    \]

  # Return only the numbers...
    (?{ $MATCH = $CAPTURE })


<token: Cmd>
  # Match standard variants then standardize the keyword...
    (?: mv | move | rename )      (?{ $MATCH = 'mv'; })
```

`$CAPTURE` *and* `$CONTEXT` *are both aliases for the built-in read-only* `$^N` *variable, which always contains the substring matched by the nearest preceding* `(...)` *capture.* `$^N` *still works perfectly well, but these are provided to improve the readability of code blocks and error messages respectively.*

El siguiente código implementa una calculadora usando destilación en el código:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n demo_calc_inline.pl
 1  use v5.10;
 2  use warnings;
 3
 4  my $calculator = do{
 5      use Regexp::Grammars;
 6      qr{
 7          <Answer>
 8
 9          <rule: Answer>
10              <X=Mult> \+ <Y=Answer>
11                  (?{ $MATCH = $MATCH{X} + $MATCH{Y}; })
12            | <X=Mult> - <Y=Answer>
13                  (?{ $MATCH = $MATCH{X} - $MATCH{Y}; })
14            | <MATCH=Mult>
15
16          <rule: Mult>
17              <X=Pow> \* <Y=Mult>
18                  (?{ $MATCH = $MATCH{X} * $MATCH{Y}; })
19            | <X=Pow>  / <Y=Mult>
20                  (?{ $MATCH = $MATCH{X} / $MATCH{Y}; })
21            | <X=Pow>  % <Y=Mult>
22                  (?{ $MATCH = $MATCH{X} % $MATCH{Y}; })
23            | <MATCH=Pow>
24
25          <rule: Pow>
```

```
26              <X=Term> \^ <Y=Pow>
27                  (?{ $MATCH = $MATCH{X} ** $MATCH{Y}; })
28            | <MATCH=Term>
29
30          <rule: Term>
31                 <MATCH=Literal>
32            | \( <MATCH=Answer> \)
33
34          <token: Literal>
35              <MATCH=( [+-]? \d++ (?: \. \d++ )?+ )>
36      }xms
37  };
38
39  while (my $input = <>) {
40      if ($input =~ $calculator) {
41          say '--> ', $/{Answer};
42      }
43  }
```

Ejercicio 31.10.1. *Cual es la salida del programa anterior para las entradas:*

- `4-2-2`
- `8/4/2`
- `2^2^3`

31.10.8. Llamadas privadas a subreglas y subreglas privadas

If a rule name (or an alias) begins with an underscore:

```
  <_RULENAME>        <_ALIAS=RULENAME>
 <[_RULENAME]>      <[_ALIAS=RULENAME]>
```

then matching proceeds as normal, and any result that is returned is stored in the current result-hash in the usual way.

However, when any rule finishes (and just before it returns) it first filters its result-hash, removing any entries whose keys begin with an underscore. This means that any subrule with an underscored name (or with an underscored alias) remembers its result, but only until the end of the current rule. Its results are effectively private to the current rule.

This is especially useful in conjunction with result distillation.

31.10.9. Mas sobre listas

Reconocimiento manual de listas

Analizando listas manualmente

El siguiente ejemplo muestra como construir un reconocedor de listas (posiblemente vacías) de números:

```
casiano@millo:~/Lregexp-grammar-examples$ cat -n simple_list.pl
     1  #!/soft/perl5lib/bin/perl5.10.1
     2  use v5.10;
     3
     4  use Regexp::Grammars;
     5
```

```
 6  my $list = qr{
 7      <List>
 8
 9      <rule: List>
10           <digit> <List>
11         | # empty
12
13      <rule: digit>
14          <MATCH=(\d+)>
15
16  }xms;
17
18  while (my $input = <>) {
19      chomp $input;
20      if ($input =~ $list) {
21          use Data::Dumper 'Dumper';
22          warn Dumper \%/;
23      }
24      else {
25        warn "Does not match\n"
26      }
27  }
```

Sigue una ejecución:

```
casiano@millo:~/Lregexp-grammar-examples$ ./simple_list.pl
2 3 4
$VAR1 = {
          '' => '2 3 4',
          'List' => {
                      '' => '2 3 4',
                      'digit' => '2'
                      'List' => {
                                  '' => '3 4',
                                  'digit' => '3'
                                  'List' => {
                                              '' => '4',
                                              'digit' => '4'
                                              'List' => '',
                                            },
                                },
                    }
        };
```

Influencia del orden en el lenguaje reconocido

Tenga en cuenta que el orden de las reglas influye en el lenguaje reconocido. Véase lo que ocurre si cambiamos en el ejemplo anterior el orden de las reglas:

```
casiano@millo:~/Lregexp-grammar-examples$ cat -n simple_list_empty_first.pl
     1  #!/soft/perl5lib/bin/perl5.10.1
     2  use v5.10;
     3
     4  use Regexp::Grammars;
     5
     6  my $list = qr{
```

```
 7      <List>
 8
 9      <rule: List>
10           # empty
11         | <digit> <List>
12
13      <rule: digit>
14          <MATCH=(\d+)>
15
16  }xms;
17
18  while (my $input = <>) {
19      chomp $input;
20      if ($input =~ $list) {
21          use Data::Dumper 'Dumper';
22          warn Dumper \%/;
23      }
24      else {
25        warn "Does not match\n"
26      }
27  }
```

Al ejecutar se obtiene:

```
casiano@millo:~/Lregexp-grammar-examples$ ./simple_list_empty_first.pl
2 3 4
$VAR1 = {
          '' => '',
          'List' => ''
        };
```

Por supuesto basta poner anclas en el patrón a buscar para forzar a que se reconozca la lista completa:

```
pl@nereida:~/Lregexpgrammars/demo$ diff simple_list_empty_first.pl simple_list_empty_first_wit
7c7
<      <List>
---
>      ^<List>$
```

En efecto, la nueva versión reconoce la lista:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 simple_list_empty_first_with_anchors.pl
2 3 4
$VAR1 = {
          '' => '2 3 4',
          'List' => {
                      'List' => {
                                  'List' => {
                                              'List' => {
                                                          'List' => '',
                                                          '' => '4',
                                                          'digit' => '4'
                                                        },
                                              '' => '3 4',
                                              'digit' => '3'
                                            },
```

```
                    '' => '2 3 4',
                    'digit' => '2'
                  }
        };
```

Si se quiere mantener la producción vacía en primer lugar pero forzar el reconocimiento de la lista completa, se puede hacer uso de un lookahead negativo:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n simple_list_empty_first_with_lookahead.pl
     1  #!/soft/perl5lib/bin/perl5.10.1
     2  use v5.10;
     3
     4  use strict;
     5  use Regexp::Grammars;
     6
     7  my $list = qr{
     8      <List>
     9
    10      <rule: List>
    11           (?! <digit> ) # still empty production
    12         | <digit> <List>
    13
    14      <rule: digit>
    15          <MATCH=(\d+)>
    16
    17  }xms;
    18
    19  while (my $input = <>) {
    20      chomp $input;
    21      if ($input =~ $list) {
    22          use Data::Dumper 'Dumper';
    23          warn Dumper \%/;
    24      }
    25      else {
    26        warn "Does not match\n"
    27      }
    28  }
```

Así, sólo se reducirá por la regla vacía si el siguiente token no es un número. Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 simple_list_empty_first_with_lookahead.pl
2 3 4
$VAR1 = {
          '' => '2 3 4',
          'List' => {
                      'List' => {
                                  'List' => {
                                              'List' => '',
                                              '' => '4',
                                              'digit' => '4'
                                            },
                                  '' => '3 4',
                                  'digit' => '3'
                                },
```

```
                    '' => '2 3 4',
                    'digit' => '2'
                  }
        };
```

Aplanamiento manual de listas

¿Cómo podemos hacer que la estructura retornada por el reconocedor sea una lista?. Podemos añadir acciones como sigue:

```
casiano@millo:~/Lregexp-grammar-examples$ cat -n simple_list_action.pl
     1  #!/soft/perl5lib/bin/perl5.10.1
     2  use v5.10;
     3
     4  use Regexp::Grammars;
     5
     6  my $list = qr{
     7      <List>
     8
     9      <rule: List>
    10           <digit> <X=List> <MATCH= (?{ unshift @{$MATCH{X}}, $MATCH{digit}; $MATCH{X} }
    11         | # empty
    12           <MATCH= (?{ [] })>
    13
    14      <rule: digit>
    15          <MATCH=(\d+)>
    16
    17  }xms;
    18
    19  while (my $input = <>) {
    20      chomp $input;
    21      if ($input =~ $list) {
    22          use Data::Dumper 'Dumper';
    23          warn Dumper \%/;
    24      }
    25      else {
    26        warn "Does not match\n"
    27      }
    28  }
```

Al ejecutarse este programa produce una salida como:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 simple_list_action.pl
2 3 4
$VAR1 = {
          '' => '2 3 4',
          'List' => [ '2', '3', '4' ]
        };
```

Los operadores de repetición

Los operadores de repetición como `*`, `+`, etc. permiten simplificar el análisis de lenguajes de listas:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n simple_list_star.pl
 1  #!/soft/perl5lib/bin/perl5.10.1
```

```
 2  use v5.10;
 3
 4  use Regexp::Grammars;
 5
 6  my $list = qr{
 7      <List>
 8
 9      <rule: List>
10          (?: <[digit]>)*
11
12      <rule: digit>
13          <MATCH=(\d+)>
14
15  }xms;
16
17  while (my $input = <>) {
18      chomp $input;
19      if ($input =~ $list) {
20          use Data::Dumper 'Dumper';
21          warn Dumper \%/;
22      }
23      else {
24        warn "Does not match\n"
25      }
26  }
```

Los corchetes alrededor de `digit` hacen que el valor asociado con el patrón sea la lista de números. Si no los ponemos el valor asociado sería el último valor de la lista.

Listas separadas por Algo

> *One of the commonest tasks in text parsing is to match a list of unspecified length, in which items are separated by a fixed token. Things like:*
>
> ```
> 1, 2, 3 , 4 ,13, 91        # Numbers separated by commas and spaces
>
> g-c-a-g-t-t-a-c-a          # Bases separated by dashes
>
> /usr/local/bin             # Names separated by directory markers
>
> /usr:/usr/local:bin        # Directories separated by colons
> ```
>
> *The usual construct required to parse these kinds of structures is either:*
>
> ```
> <rule: list>
>
>     <item> <separator> <list              # recursive definition
>   | <item>                                # base case
> ```
>
> *Or, more efficiently, but less prettily:*
>
> ```
> <rule: list>
>
>     <[item]> (?: <separator> <[item]> )*   # iterative definition
> ```

Because this is such a common requirement, `Regexp::Grammars` *provides a cleaner way to specify the iterative version. The syntax is taken from Perl 6:*

```
<rule: list>

    <[item]> ** <separator>                # iterative definition
```

This is a repetition specifier on the first subrule (hence the use of `**` *as the marker, to reflect the repetitive behaviour of* `*`*). However, the number of repetitions is controlled by the second subrule: the first subrule will be repeatedly matched for as long as the second subrule matches immediately after it.*

So, for example, you can match a sequence of numbers separated by commas with:

```
<[number]> ** <comma>

<token: number>  \d+
<token: comma>   \s* , \s*
```

Note that it's important to use the `<[...]>` *form for the items being matched, so that all of them are saved in the result hash. You can also save all the separators (if that's important):*

```
<[number]> ** <[comma]>
```

The repeated item must be specified as a subrule call fo some kind, but the separators may be specified either as a subrule or a bracketed pattern. For example:

```
<[number]> ** ( , )
```

The separator must always be specified in matched delimiters of some kind: either matching `<...>` *or matching* `(...)`*. A common error is to write:*

```
<[number]> ** ,
```

You can also use a pattern as the item matcher, but it must be aliased into a subrule:

```
<[item=(\d+)]> ** ( , )
```

Ejemplo: Listas de números separados por comas

Veamos un ejemplo sencillo:

```
casiano@millo:~/src/perl/regexp-grammar-examples$ cat -n demo_list.pl
 1  #!/soft/perl5lib/bin/perl5.10.1
 2  use v5.10;
 3
 4  use Regexp::Grammars;
 5
 6  my $list_nonempty = qr{
 7      <List>
 8
 9      <rule: List>
10          \(  <[Value]> ** (,)  \)
11
12      <token: Value>
13          \d+
```

```
14  }xms;
15
16  my $list_empty = qr{
17      <List>
18
19      <rule: List>
20          \(  (?: <[Value]> ** <_Sep=(,)> )?  \)
21
22      <token: Value>
23          \d+
24  }xms;
25
26  use Smart::Comments;
27
28
29  while (my $input = <>) {
30      my $input2 = $input;
31      if ($input =~ $list_nonempty) {
32          ### nonempty: $/{List}
33      }
34      if ($input2 =~ $list_empty) {
35          ### empty: $/{List}
36      }
37  }
```

Sigue un ejemplo de ejecución:

```
casiano@millo:~/src/perl/regexp-grammar-examples$ ./demo_list.pl
(3,4,5)

### nonempty: {
###             '' => '(3,4,5)',
###             Value => [
###                         '3',
###                         '4',
###                         '5'
###                       ]
###           }

### empty: {
###           '' => '(3,4,5)',
###           Value => [
###                       '3',
###                       '4',
###                       '5'
###                     ]
###         }
()

### empty: '()'
```

Ejemplo: AST para las expresiones aritméticas

Las expresiones aritméticas puede definirse como una jerarquía de listas como sigue:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n calcaslist.pl
 1  use strict;
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5  $Data::Dumper::Indent = 1;
 6
 7  my $rbb = do {
 8      use Regexp::Grammars;
 9
10      qr{
11        \A<expr>\z
12
13        <objrule: expr>      <[operands=term]> ** <[operators=addop]>
14
15        <objrule: term>      <[operands=uneg]> ** <[operators=mulop]>
16
17        <objrule: uneg>      <[operators=minus]>* <[operands=power]>
18
19        <objrule: power>     <[operands=factorial]> ** <[operators=powerop]>
20
21        <objrule: factorial> <[operands=factor]>  <[operators=(!)]>*
22
23        <objrule: factor>    <val=([+-]?\d+(?:\.\d*)?)>
24                           | \( <MATCH=expr> \)
25
26        <token: addop>        [+-]
27
28        <token: mulop>        [*/]
29
30        <token: powerop>      \*\*|\^
31
32        <token: minus>        - <MATCH=(?{ 'NEG' })>
33
34      }x;
35  };
36
37  while (my $input = <>) {
38      chomp($input);
39      if ($input =~ m{$rbb}) {
40          my $tree = $/{expr};
41          say Dumper $tree;
42
43      }
44      else {
45          say("does not match");
46      }
47  }
```

Obsérvese el árbol generado para la expresión `4-2-2`:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 calcaslist.pl
4-2-2
$VAR1 = bless( {
```

```
'operands' => [
  bless( {
    'operands' => [
      bless( {
        'operands' => [
          bless( {
            'operands' => [
              bless( {
                'operands' => [
                  bless( { '' => '4', 'val' => '4' }, 'factor' )
                ],
                '' => '4'
              }, 'factorial' )
            ],
            '' => '4'
          }, 'power' )
        ],
        '' => '4'
      }, 'uneg' )
    ],
    '' => '4'
  }, 'term' ),
  bless( {
    'operands' => [
      bless( {
        'operands' => [
          bless( {
            'operands' => [
              bless( {
                'operands' => [
                  bless( { '' => '2', 'val' => '2' }, 'factor' )
                ],
                '' => '2'
              }, 'factorial' )
            ],
            '' => '2'
          }, 'power' )
        ],
        '' => '2'
      }, 'uneg' )
    ],
    '' => '2'
  }, 'term' ),
  bless( {
    'operands' => [
      bless( {
        'operands' => [
          bless( {
            'operands' => [
              bless( {
                'operands' => [
                  bless( { '' => '2', 'val' => '2' }, 'factor' )
                ],
```

```
                '' => '2'
              }, 'factorial' )
            ],
            '' => '2'
          }, 'power' )
        ],
        '' => '2'
      }, 'uneg' )
    ],
    '' => '2'
  }, 'term' )
 ],
 '' => '4-2-2',
 'operators' => [
   '-',
   '-'
 ]
}, 'expr' );
```

31.10.10. La directiva `require`

La directiva `require` es similar en su funcionamiento al paréntesis 5.10 (`??{ Código Perl }`) el cuál hace que el `Código Perl` sea evaluado durante el tiempo de matching. El resultado de la evaluación se trata como una expresión regular con la que deberá casarse. (véase la sección 31.2.9 para mas detalles).

La sintáxis de la directiva `<require:>` es

```
<require: (?{ CODE }) >
```

The code block is executed and if its final value is true, matching continues from the same position. If the block's final value is false, the match fails at that point and starts backtracking.

The `<require:...>` *directive is useful for testing conditions that it's not easy (or even possible) to check within the syntax of the the regex itself. For example:*

```
<rule: IPV4_Octet_Decimal>
    # Up three digits...
    <MATCH= ( \d{1,3}+ )>

    # ...but less that 256...
    <require: (?{ $MATCH <= 255 })>
```

A require expects a regex codeblock as its argument and succeeds if the final value of that codeblock is true. If the final value is false, the directive fails and the rule starts backtracking.

Note, in this example that the digits are matched with `\d{1,3}+` . The trailing `+` *prevents the* `{1,3}` *repetition from backtracking to a smaller number of digits if the* `<require:...>` *fails.*

El programa `demo_IP4.pl` ilustra el uso de la directiva:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n ./demo_IP4.pl
 1  #!/usr//bin/env perl5.10.1
 2  use v5.10;
 3  use warnings;
```

```
 4
 5  use Regexp::Grammars;
 6
 7  my $grammar = qr{
 8      \A <IP4_addr> \Z
 9
10      <token: quad>
11          <MATCH=(\d{1,3})>
12          <require: (?{ $MATCH < 256 })>
13
14      <token: IP4_addr>
15          <[MATCH=quad]>**(\.)
16          <require: (?{ @$MATCH == 4 })>
17  }xms;
18
19  while (my $line = <>) {
20      if ($line =~ $grammar) {
21          use Data::Dumper 'Dumper';
22          say Dumper \%/;
23      }
24      else {
25          say 'Does not match'
26      }
27  }
```

Las condiciones usadas en el `require` obligan a que cada quad[9] sea menor que 256 y a que existan sólo cuatro quads.

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lregexpgrammars/demo$ ./demo_IP4.pl
123 . 145 . 105 . 252
Does not match
pl@nereida:~/Lregexpgrammars/demo$ ./demo_IP4.pl
123.145.105.252
$VAR1 = {
          '' => '123.145.105.252',
          'IP4_addr' => [
                          123,
                          145,
                          105,
                          252
                        ]
        };
pl@nereida:~/Lregexpgrammars/demo$ ./demo_IP4.pl
148.257.128.128
Does not match
0.0.0.299
Does not match
pl@nereida:~/Lregexpgrammars/demo$  ./demo_IP4.pl
123.145.105.242.193
```

[9] A quad (pronounced KWAHD) is a unit in a set of something that comes in four units. The term is sometimes used to describe each of the four numbers that constitute an Internet Protocol (IP) address. Thus, an Internet address in its numeric form (which is also sometimes called a dot address) consists of four quads separated by "dots"(periods).

A quad also means *a quarter* in some usages. (A quarter as a U.S. coin or monetary unit means *a quarter of a dollar,* and in slang is sometimes called *two bits.* However, this usage does not mean two binary bits as used in computers.)

```
Does not match
```

Obsérvese como no se aceptan blancos entre los puntos en esta versión. ¿Sabría explicar la causa?

31.10.11. Casando con las claves de un hash

In some situations a grammar may need a rule that matches dozens, hundreds, or even thousands of one-word alternatives. For example, when matching command names, or valid userids, or English words. In such cases it is often impractical (and always inefficient) to list all the alternatives between `|` *alterators:*

```
<rule: shell_cmd>
    a2p | ac | apply | ar | automake | awk | ...
    # ...and 400 lines later
    ... | zdiff | zgrep | zip | zmore | zsh

<rule: valid_word>
    a | aa | aal | aalii | aam | aardvark | aardwolf | aba | ...
    # ...and 40,000 lines later...
    ... | zymotize | zymotoxic | zymurgy | zythem | zythum
```

To simplify such cases, `Regexp::Grammars` *provides a special construct that allows you to specify all the alternatives as the keys of a normal hash. The syntax for that construct is simply to put the hash name inside angle brackets (with no space between the angles and the hash name).*

Which means that the rules in the previous example could also be written:

```
<rule: shell_cmd>
    <%cmds>

<rule: valid_word>
    <%dict>
```

provided that the two hashes (`%cmds` *and* `%dict`*) are visible in the scope where the grammar is created.*

Internally, the construct is converted to something equivalent to:

```
<rule: shell_cmd>
    (<.hk>)  <require: exists $cmds{$CAPTURE}>

<rule: valid_word>
    (<.hk>)  <require: exists $dict{$CAPTURE}>
```

The special `<hk>` *rule is created automatically, and defaults to* `\S+`*, but you can also define it explicitly to handle other kinds of keys. For example:*

```
<rule: hk>
    .+            # Key may be any number of chars on a single line

<rule: hk>
    [ACGT]{10,}   # Key is a base sequence of at least 10 pairs
```

Matching a hash key in this way is typically significantly faster than matching a full set of alternations. Specifically, it is O(length of longest potential key), instead of O(number of keys).

Ejemplo de uso de la directiva hash

Sigue un ejemplo:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n hash.pl
 1  #!/usr/bin/env perl5.10.1
 2  use strict;
 3  use warnings;
 4  use 5.010;
 5  use Data::Dumper;
 6  $Data::Dumper::Deparse = 1;
 7
 8  my %cmd = map { ($_ => undef ) } qw( uname pwd date );
 9
10  my $rbb = do {
11      use Regexp::Grammars;
12
13      qr{
14        ^<command>$
15
16        <rule: command>
17          <cmd=%cmd> (?: <[arg]> )*
18
19        <token: arg> [^\s<>‘&]+
20      }xms;
21  };
22
23  while (my $input = <>) {
24      chomp($input);
25      if ($input =~ m{$rbb}) {
26          say("matches: <$&>");
27          say Dumper \%/;
28          system $/{''}
29      }
30      else {
31          say("does not match");
32      }
33  }
```

Sigue un ejemplo de ejecución:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 hash.pl
a2p f1 f2
matches: <a2p f1 f2>
$VAR1 = {
          '' => 'a2p f1 f2',
          'command' => {
                         '' => 'a2p f1 f2',
                         'cmd' => 'a2p',
                         'arg' => [
                                    'f1',
                                    'f2'
                                  ]
                       }
        };
```

```
pocho 2 5
does not match
```

31.10.12. Depuración

`Regexp::Grammars` *provides a number of features specifically designed to help debug both grammars and the data they parse.*

All debugging messages are written to a log file (which, by default, is just `STDERR`*). However, you can specify a disk file explicitly by placing a* `"<logfile:...>"` *directive* at the start of your grammar[10]*:*

```
$grammar = qr{

    <logfile: LaTeX_parser_log >

    \A <LaTeX_file> \Z    # Pattern to match

    <rule: LaTeX_file>
        # etc.
}x;
```

You can also explicitly specify that messages go to the terminal:

```
    <logfile: - >
```

Debugging grammar creation

Whenever a log file has been directly specified, `Regexp::Grammars` *automatically does verbose static analysis of your grammar. That is, whenever it compiles a grammar containing an explicit* `"<logfile:...>"` *directive it logs a series of messages explaining how it has interpreted the various components of that grammar. For example, the following grammar:*

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n log.pl
     1  #!/usr/bin/env perl5.10.1
     2  use strict;
     3  use warnings;
     4  use 5.010;
     5  use Data::Dumper;
     6
     7  my $rbb = do {
     8      use Regexp::Grammars;
     9
    10      qr{
    11        <logfile: ->
    12
    13        <numbers>
    14
    15        <rule: numbers>
    16          <number> ** <.comma>
    17
    18        <token: number> \d+
    19
```

[10] no funcionará si no se pone al principio de la gramática

```
    20        <token: comma>   ,
    21      }xms;
    22  };
    23
    24  while (my $input = <>) {
    25      if ($input =~ m{$rbb}) {
    26          say("matches: <$&>");
    27          say Dumper \%/;
    28      }
    29  }
```

would produce the following analysis in the terminal:

```
pl@nereida:~/Lregexpgrammars/demo$ ./log.pl
  warn | Repeated subrule <number>* will only capture its final match
       | (Did you mean <[number]>* instead?)
       |
  info | Processing the main regex before any rule definitions
       |    |
       |    |...Treating <numbers> as:
       |    |       |  match the subrule <numbers>
       |    |       \ saving the match in $MATCH{'numbers'}
       |    |
       |     \___End of main regex
       |
       | Defining a rule: <numbers>
       |    |...Returns: a hash
       |    |
       |    |...Treating <number> as:
       |    |       |  match the subrule <number>
       |    |       \ saving the match in $MATCH{'number'}
       |    |
       |    |...Treating <.comma> as:
       |    |       |  match the subrule <comma>
       |    |       \ but don't save anything
       |    |
       |    |...Treating <number> ** <.comma> as:
       |    |       |  repeatedly match the subrule <number>
       |    |       \ as long as the matches are separated by matches of <.comma>
       |    |
       |     \___End of rule definition
       |
       | Defining a rule: <number>
       |    |...Returns: a hash
       |    |
       |    |...Treating '\d' as:
       |    |       \ normal Perl regex syntax
       |    |
       |    |...Treating '+ ' as:
       |    |       \ normal Perl regex syntax
       |    |
       |     \___End of rule definition
       |
       | Defining a rule: <comma>
```

```
        |    |...Returns: a hash
        |    |
        |    |...Treating ', ' as:
        |    |       \ normal Perl regex syntax
        |    |
        |     \___End of rule definition
        |
2, 3, 4
matches: <2, 3, 4>
$VAR1 = {
          '' => '2, 3, 4',
          'numbers' => {
                         '' => '2, 3, 4',
                         'number' => '4'
                       }
        };
```

This kind of static analysis is a useful starting point in debugging a miscreant grammar[11]*, because it enables you to see what you actually specified (as opposed to what you* thought *you'd specified).*

Debugging grammar execution

`Regexp::Grammars` *also provides a simple interactive debugger, with which you can observe the process of parsing and the data being collected in any result-hash.*

To initiate debugging, place a `<debug:...>` *directive anywhere in your grammar. When parsing reaches that directive the debugger will be activated, and the command specified in the directive immediately executed. The available commands are:*

```
<debug: on>    - Enable debugging, stop when entire grammar matches
<debug: match> - Enable debugging, stope when a rule matches
<debug: try>   - Enable debugging, stope when a rule is tried
<debug: off>   - Disable debugging and continue parsing silently

<debug: continue> - Synonym for <debug: on>
<debug: run>      - Synonym for <debug: on>
<debug: step>     - Synonym for <debug: try>
```

These directives can be placed anywhere within a grammar and take effect when that point is reached in the parsing. Hence, adding a `<debug:step>` *directive is very much like setting a breakpoint at that point in the grammar. Indeed, a common debugging strategy is to turn debugging on and off only around a suspect part of the grammar:*

```
<rule: tricky>   # This is where we think the problem is...
    <debug:step>
    <preamble> <text> <postscript>
    <debug:off>
```

Once the debugger is active, it steps through the parse, reporting rules that are tried, matches and failures, backtracking and restarts, and the parser's location within both the grammar and the text being matched. That report looks like this:

[11] miscreant - *One who has behaved badly, or illegally; One not restrained by moral principles; an unscrupulous villain; One who holds an incorrect religious belief; an unbeliever; Lacking in conscience or moral principles; unscrupulous; Holding an incorrect religious belief.*

```
===============> Trying <grammar> from position 0
> cp file1 file2 |...Trying <cmd>
                 |   |...Trying <cmd=(cp)>
                 |   |    \FAIL <cmd=(cp)>
                 |    \FAIL <cmd>
                  \FAIL <grammar>
===============> Trying <grammar> from position 1
 cp file1 file2  |...Trying <cmd>
                 |   |...Trying <cmd=(cp)>
 file1 file2     |   |    \_____<cmd=(cp)> matched 'cp'
file1 file2      |   |...Trying <[file]>+
 file2           |   |    \_____<[file]>+ matched 'file1'
                 |   |...Trying <[file]>+
[eos]            |   |    \_____<[file]>+ matched ' file2'
                 |   |...Trying <[file]>+
                 |   |    \FAIL <[file]>+
                 |   |...Trying <target>
                 |   |   |...Trying <file>
                 |   |   |    \FAIL <file>
                 |   |    \FAIL <target>
 <~~~~~~~~~~~~~~ |   |...Backtracking 5 chars and trying new match
file2            |   |...Trying <target>
                 |   |   |...Trying <file>
                 |   |   |    \____ <file> matched 'file2'
[eos]            |   |    \_____<target> matched 'file2'
                 |    \_____<cmd> matched ' cp file1 file2'
                  \_____<grammar> matched ' cp file1 file2'
```

The first column indicates the point in the input at which the parser is trying to match, as well as any backtracking or forward searching it may need to do. The remainder of the columns track the parser's hierarchical traversal of the grammar, indicating which rules are tried, which succeed, and what they match.

Provided the logfile is a terminal (as it is by default), the debugger also pauses at various points in the parsing process–before trying a rule, after a rule succeeds, or at the end of the parse–according to the most recent command issued. When it pauses, you can issue a new command by entering a single letter:

```
m       - to continue until the next subrule matches
t or s  - to continue until the next subrule is tried
r or c  - to continue to the end of the grammar
o       - to switch off debugging
```

Note that these are the first letters of the corresponding `<debug:...>` *commands, listed earlier. Just hitting ENTER while the debugger is paused repeats the previous command.*

While the debugger is paused you can also type a `d`*, which will display the result-hash for the current rule. This can be useful for detecting which rule isn't returning the data you expected.*

Veamos un ejemplo. El siguiente programa activa el depurador:

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n demo_debug.pl
     1  #!/usr/bin/env perl5.10.1
     2  use 5.010;
     3  use warnings;
     4
```

```
 5      use Regexp::Grammars;
 6
 7      my $balanced_brackets = qr{
 8          <debug:on>
 9
10          <left_delim=(  \( )>
11          (?:
12              <[escape=(  \\ )]>
13          |   <recurse=( (?R) )>
14          |   <[simple=(  .  )]>
15          )*
16          <right_delim=( \) )>
17      }xms;
18
19      while (<>) {
20          if (/$balanced_brackets/) {
21              say 'matched:';
22              use Data::Dumper 'Dumper';
23              warn Dumper \%/;
24          }
25      }
```

Al ejecutar obtenemos

```
pl@nereida:~/Lregexpgrammars/demo$ ./demo_debug.pl
(a)
=====> Trying <grammar> from position 0
(a)\n  |...Trying <left_delim=(  \( )>
a)\n   |    \_____<left_delim=(  \( )> matched '('      c
       |...Trying <[escape=(  \ )]>
       |    \FAIL <[escape=(  \ )]>
       |...Trying <recurse=( (?R) )>
=====> Trying <grammar> from position 1
a)\n   |   |...Trying <left_delim=(  \( )>
       |   |    \FAIL <left_delim=(  \( )>
        \FAIL <grammar>
       |...Trying <[simple=(  .  )]>
)\n    |    \_____<[simple=(  .  )]> matched 'a'
       |...Trying <[escape=(  \ )]>
       |    \FAIL <[escape=(  \ )]>
       |...Trying <recurse=( (?R) )>
=====> Trying <grammar> from position 2
)\n    |   |...Trying <left_delim=(  \( )>
       |   |    \FAIL <left_delim=(  \( )>
        \FAIL <grammar>
       |...Trying <[simple=(  .  )]>
\n     |    \_____<[simple=(  .  )]> matched ')'
       |...Trying <[escape=(  \ )]>
       |    \FAIL <[escape=(  \ )]>
       |...Trying <recurse=( (?R) )>
=====> Trying <grammar> from position 3
\n     |   |...Trying <left_delim=(  \( )>
       |   |    \FAIL <left_delim=(  \( )>
        \FAIL <grammar>
```

```
       |...Trying <[simple=(  .  )]>
[eos]  |    \_____<[simple=(  .  )]> matched ''
       |...Trying <[escape=(  \ )]>
       |    \FAIL <[escape=(  \ )]>
       |...Trying <recurse=( (?R) )>
=====> Trying <grammar> from position 4
[eos]  |   |...Trying <left_delim=(  \( )>
       |   |    \FAIL <left_delim=(  \( )>
        \FAIL <grammar>
       |...Trying <[simple=(  .  )]>
       |    \FAIL <[simple=(  .  )]>
       |...Trying <right_delim=( \) )>
       |    \FAIL <right_delim=( \) )>
 <~~~~ |...Backtracking 1 char and trying new match
\n     |...Trying <right_delim=( \) )>
       |    \FAIL <right_delim=( \) )>
 <~~~~ |...Backtracking 1 char and trying new match
)\n    |...Trying <right_delim=( \) )>
\n     |    \_____<right_delim=( \) )> matched ')'
        \_____<grammar> matched '(a)'   d
              :         {
              :           '' => '(a)',
              :           'left_delim' => '(',
              :           'simple' => [
              :                         'a'
              :                       ],
              :           'right_delim' => ')'
              :         };      o
matched:
$VAR1 = {
          '' => '(a)',
          'left_delim' => '(',
          'simple' => [
                        'a'
                      ],
          'right_delim' => ')'
        };
```

31.10.13. Mensajes de `log` del usuario

Both static and interactive debugging send a series of predefined log messages to whatever log file you have specified. It is also possible to send additional, user-defined messages to the log, using the `"<log:...>"` *directive.*

This directive expects either a simple text or a codeblock as its single argument. If the argument is a code block, that code is expected to return the text of the message; if the argument is anything else, that something else is *the literal message. For example:*

```
<rule: ListElem>

    <Elem=   ( [a-z]\d+) >
        <log: Checking for a suffix, too...>

    <Suffix= ( : \d+   ) >?
        <log: (?{ "ListElem: $MATCH{Elem} and $MATCH{Suffix}" })>
```

User-defined log messages implemented using a codeblock can also specify a severity level. If the codeblock of a `<log:...>` *directive returns two or more values, the first is treated as a log message severity indicator, and the remaining values as separate lines of text to be logged. For example:*

```
<rule: ListElem>
    <Elem=   ( [a-z]\d+) >
    <Suffix= ( : \d+   ) >?

        <log: (?{
            warn => "Elem was: $MATCH{Elem}",
                    "Suffix was $MATCH{Suffix}",
        })>
```

When they are encountered, user-defined log messages are interspersed between any automatic log messages (i.e. from the debugger), at the correct level of nesting for the current rule.

31.10.14. Depuración de Regexps

It is possible to use `Regexp::Grammars` *without creating* any *subrule definitions, simply to debug a recalcitrant regex. For example, if the following regex wasn't working as expected:*

```
my $balanced_brackets = qr{
    \(              # left delim
    (?:
        \\          # escape or
    |   (?R)        # recurse or
    |   .           # whatever
    )*
    \)              # right delim
}xms;
```

you could instrument it with aliased subpatterns and then debug it step-by-step, using `Regexp::Grammars`*:*

```
use Regexp::Grammars;

my $balanced_brackets = qr{
    <debug:step>

    <.left_delim=  (  \(  )>
    (?:
        <.escape=  (  \\  )>
    |   <.recurse= ( (?R) )>
    |   <.whatever=(  .   )>
    )*
    <.right_delim= (  \)  )>
}xms;

while (<>) {
    say 'matched' if /$balanced_brackets/;
}
```

Note the use of amnesiac aliased subpatterns to avoid needlessly building a result-hash. Alternatively, you could use listifying aliases to preserve the matching structure as an additional debugging aid:

```
use Regexp::Grammars;

my $balanced_brackets = qr{
    <debug:step>

    <[left_delim=  (  \(  )]>
    (?:
        <[escape=  (  \\  )]>
    |   <[recurse= ( (?R) )]>
    |   <[whatever=(  .   )]>
    )*
    <[right_delim= (  \)  )]>
}xms;

if ( '(a(bc)d)' =~ /$balanced_brackets/) {
    use Data::Dumper 'Dumper';
    warn Dumper \%/;
}
```

31.10.15. Manejo y recuperación de errores

En este punto debo decir que no he podido reproducir el comportamiento de las directivas `<error:>` y `<warning:>` tal y como las describe Conway en el manual de `Regexp::Grammars`.

El siguiente ejemplo ilustra un conjunto de técnicas de gestión de errores que son independientes del soprote dado por `Regexp::Grammars`.

Se trata de la misma calculadora explicada en la sección 31.10.18.

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ cat -n calculatorwitherrmanagement.pl
 1  #!/usr/bin/env perl5.10.1
 2  use strict;
 3  use warnings;
 4  use 5.010;
 5  use Lingua::EN::Inflect qw(PL);
 6  use Scalar::Util qw{blessed};
 7
 8  my $rbb = do {
 9      my ($warnings, $errors);    # closure
10      sub warnings { $warnings }  # accessor
11      sub errors { $errors }      # accessor
12
13      use Regexp::Grammars;
14      qr{
15        (?{
16            $warnings = 0;
17            $errors = 0;
18        })
19        \A<expr>
20        (?:   \z
21             |
22               (.*) (?{
```

```
23                      # Accept the string but emit a warning
24                      $warnings++;
25                      local our $expr = \$MATCH{expr}{''};
26                      local our $endlegal = length($$expr) > 4? "... ".substr($$expr, -4)
27                      warn "Warning: Unexpected '". substr($^N, 0, 10)."' after '$endlegal
28                  })
29      )
30
31      <objrule: expr>      <[operands=term]> ** <[operators=addop]>
32
33      <objrule: term>      <[operands=uneg]> ** <[operators=mulop]>
34
35      <objrule: uneg>      <[operators=minus]>* <[operands=power]>
36
37      <objrule: power>     <[operands=factorial]> ** <[operators=powerop]>
38
39      <objrule: factorial> <[operands=factor]>  <[operators=(!)]>*
40
41      <objrule: factor>    (<val=([+-]?\d+(?:\.\d*)?)>)
42                         | \( <MATCH=expr> \)
43                         | ([^-+(0-9]+) (?{
44                                         # is + and not * to avoid infinite recursion
45                                         warn "Error: expecting a number or a open parent
46                                         $warnings++;
47                                         $errors++;
48                                     }) <MATCH=factor>
49
50      <token: addop>        [+-]
51
52      <token: mulop>        [*/]
53
54      <token: powerop>      \*\*|\^
55
56      <token: minus>        - <MATCH=(?{ 'NEG' })>
57
58    }x;
59 };
60
61 sub test_calc {
62   my $prompt = shift;
63
64   print $prompt;
65   while (my $input = <>) {
66       chomp($input);
67
68       local %/;
69       $input =~ m{$rbb};
70
71       say warnings." ".PL('warning',warnings) if warnings;
72       say errors." ".PL('error',errors)       if errors;
73
74       my $tree = $/{expr};
75       if (blessed($tree)) {
```

```
76            do "PostfixCalc.pm";
77            say "postfix: ".$tree->ceval;
78
79            do "EvalCalc.pm";
80            say "result: ".$tree->ceval;
81        }
82        print $prompt;
83    }
84    say "Bye!"
85  }
86
87  ########## main
88  test_calc(
89    'Parsing infix arithmetic expressions (CTRL-D to end in unix) ',
90  );
```

Veamos algunas ejecuciones que incluyen entradas erróneas:

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ ./calculatorwitherrmanagement.pl
Parsing infix arithmetic expressions (CTRL-D to end in unix) 2+3
postfix: 2 3 +
result: 5
Parsing infix arithmetic expressions (CTRL-D to end in unix) 2*(3+#)
Error: expecting a number or a open parenthesis, found: '#)'
Error: expecting a number or a open parenthesis, found: '#'
Error: expecting a number or a open parenthesis, found: ')'
Warning: Unexpected '*(3+#)' after '2'
4 warnings
3 errors
postfix: 2
result: 2
Parsing infix arithmetic expressions (CTRL-D to end in unix) 2+#*4
Error: expecting a number or a open parenthesis, found: '#*'
1 warning
1 error
postfix: 2 4 +
result: 6
Parsing infix arithmetic expressions (CTRL-D to end in unix) Bye!
```

Obsérvese los mensajes de error repetidos para la entrada `2*(3+#)`. Ellos son debidos a los reiterados intentos de casar `<factor>` en la regla de recuperación de errores:

```
41        <objrule: factor>    (<val=([+-]?\d+(?:\.\d*)?)>)
42                           | \( <MATCH=expr> \)
43                           | ([^-+(0-9]+) (?{
44                                             # is + and not * to avoid infinite recursion
45                                             warn "Error: expecting a number or a open parent
46                                             $warnings++;
47                                             $errors++;
48                                         }) <MATCH=factor>
```

en este caso resulta imposible encontrar un factor. Se puede cambiar la conducta indicando un `(* COMMIT)` antes de la llamada a `<MATCH=factor>`:

```
 41       <objrule: factor>    (<val=([+-]?\d+(?:\.\d*)?)>)
 42                          | \( <MATCH=expr> \)
```

```
 43                          | ([^-+(0-9]+) (?{
 44                                          # is + and not * to avoid infinite recursion
 45                                          warn "Error: expecting a number or a open parent
 46                                          $warnings++;
 47                                          $errors++;
 48                                      }) (*COMMIT) <MATCH=factor>
```

en este caso la conducta es abandonar en el caso de que no se pueda encontrar un `<factor>`:

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ ./calculatorwitherrmanagement.pl
Parsing infix arithmetic expressions (CTRL-D to end in unix) 2*(3+#)
Error: expecting a number or a open parenthesis, found: '#)'
1 warning
1 error
Parsing infix arithmetic expressions (CTRL-D to end in unix) 2*3
postfix: 2 3 *
result: 6
Parsing infix arithmetic expressions (CTRL-D to end in unix) @
Error: expecting a number or a open parenthesis, found: '@'
1 warning
1 error
Parsing infix arithmetic expressions (CTRL-D to end in unix) Bye!
```

31.10.16. Mensajes de Warning

Sometimes, you want to detect problems, but not invalidate the entire parse as a result. For those occasions, the module provides a less stringent *form of error reporting: the* `<warning:...>` *directive.*

This directive is exactly the same as an `<error:...>` *in every respect except that it does not induce a failure to match at the point it appears.*

The directive is, therefore, useful for reporting non-fatal *problems in a parse. For example:*

```
qr{ \A            # ...Match only at start of input
    <ArithExpr>   # ...Match a valid arithmetic expression

    (?:
        # Should be at end of input...
        \s* \Z
      |
        # If not, report the fact but don't fail...
        <warning: Expected end-of-input>
        <warning: (?{ "Extra junk at index $INDEX: $CONTEXT" })>
    )

    # Rule definitions here...
}xms;
```

Note that, because they do not induce failure, two or more `<warning:...>` *directives can be "stackedïn sequence, as in the previous example.*

31.10.17. Simplificando el AST

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n exprdamian.pl
     1  use strict;
```

```
 2  use warnings;
 3  use 5.010;
 4  use Data::Dumper;
 5  $Data::Dumper::Indent = 1;
 6
 7  my $rbb = do {
 8      use Regexp::Grammars;
 9
10      qr{
11        \A<expr>\z
12
13        <objrule: expr>    <MATCH=term> (?! <addop> )                    # bypass
14                         | <[operands=term]> ** <[operators=addop]>
15
16        <objrule: term>    <MATCH=factor> (?! <mulop> )                  # bypass
17                         | <[operands=factor]> ** <[operators=mulop]>
18
19        <objrule: factor>    <val=([+-]?\d+(?:\.\d*)?)>
20                         | \( <MATCH=expr> \)
21
22        <token: addop> [+-]
23
24        <token: mulop> [*/]
25
26      }x;
27  };
28
29  while (my $input = <>) {
30      chomp($input);
31      if ($input =~ m{$rbb}) {
32          my $tree = $/{expr};
33          say Dumper $tree;
34          say $tree->ceval;
35
36      }
37      else {
38          say("does not match");
39      }
40  }
41
42  BEGIN {
43
44    package LeftBinaryOp;
45    use strict;
46    use base qw(Class::Accessor);
47
48    LeftBinaryOp->mk_accessors(qw{operators operands});
49
50    my %f = (
51      '+' => sub { shift() + shift() },
52      '-' => sub { shift() - shift() },
53      '*' => sub { shift() * shift() },
54      '/' => sub { shift() / shift() },
```

```
55    );
56
57    sub ceval {
58      my $self = shift;
59
60      # recursively evaluate the children first
61      my @operands = map { $_->ceval } @{$self->operands};
62
63      # then combine them
64      my $s = shift @operands;
65      for (@{$self->operators}) {
66        $s = $f{$_}->($s, shift @operands);
67      }
68      return $s;
69    }
70
71    package term;
72    use base qw{LeftBinaryOp};
73
74    package expr;
75    use base qw{LeftBinaryOp};
76
77    package factor;
78
79    sub ceval {
80      my $self = shift;
81
82      return $self->{val};
83    }
84
85    1;
86  }
```

Ejecuciones:

```
pl@nereida:~/Lregexpgrammars/demo$ perl5.10.1 exprdamian.pl
4-2-2
$VAR1 = bless( {
  'operands' => [
    bless( {
      '' => '4',
      'val' => '4'
    }, 'factor' ),
    bless( {
      '' => '2',
      'val' => '2'
    }, 'factor' ),
    bless( {
      '' => '2',
      'val' => '2'
    }, 'factor' )
  ],
  '' => '4-2-2',
  'operators' => [
```

```
    '-',
    '-'
  ]
}, 'expr' );

0
8/4/2
$VAR1 = bless( {
  'operands' => [
    bless( {
      '' => '8',
      'val' => '8'
    }, 'factor' ),
    bless( {
      '' => '4',
      'val' => '4'
    }, 'factor' ),
    bless( {
      '' => '2',
      'val' => '2'
    }, 'factor' )
  ],
  '' => '8/4/2',
  'operators' => [
    '/',
    '/'
  ]
}, 'term' );

1
3
$VAR1 = bless( {
  '' => '3',
  'val' => '3'
}, 'factor' );

3
2*(3+4)
$VAR1 = bless( {
  'operands' => [
    bless( {
      '' => '2',
      'val' => '2'
    }, 'factor' ),
    bless( {
      'operands' => [
        bless( {
          '' => '3',
          'val' => '3'
        }, 'factor' ),
        bless( {
          '' => '4',
          'val' => '4'
```

```
        }, 'factor' )
      ],
      '' => '3+4',
      'operators' => [
        '+'
      ]
    }, 'expr' )
  ],
  '' => '2*(3+4)',
  'operators' => [
    '*'
  ]
}, 'term' );

14
```

31.10.18. Reciclando una `Regexp::Grammar`

Ejecución

El siguiente programa `calculator.pl` recibe como entrada una expresión en infijo.

La ejecución consta de dos bucles. En la primera parte se inyecta a la jerarquía de clases de los AST generados para las expresiones en infijo una semántica que permite evaluar la expresión:

```
58  require EvalCalc;
59
60  test_calc(
61    'Evaluating infix arithmetic expressions (CTRL-D to end in unix) ',
62    sub { print &Data::Dumper::Dumper(shift()) },
63  );
```

En esta primera parte mostraremos además el AST construido para la expresión infija de entrada.

```
pl@nereida:~/Lregexpgrammars/demo$ ./calculator.pl
Evaluating infix arithmetic expressions (CTRL-D to end in unix)
8-4-2
$VAR1 = bless( {
  'operands' => [
    bless( {
      'operands' => [
        bless( {
          'operands' => [
            bless( {
              'operands' => [
                bless( {
                  'operands' => [
                    bless( { '' => '8', 'val' => '8' }, 'factor' )
                  ],
                  '' => '8'
                }, 'factorial' )
              ],
              '' => '8'
            }, 'power' )
          ],
          '' => '8'
        }, 'uneg' )
```

```
      ],
      '' => '8'
    }, 'term' ),
    bless( {
      'operands' => [
        bless( {
          'operands' => [
            bless( {
              'operands' => [
                bless( {
                  'operands' => [
                    bless( { '' => '4', 'val' => '4' }, 'factor' )
                  ],
                  '' => '4'
                }, 'factorial' )
              ],
              '' => '4'
            }, 'power' )
          ],
          '' => '4'
        }, 'uneg' )
      ],
      '' => '4'
    }, 'term' ),
    bless( {
      'operands' => [
        bless( {
          'operands' => [
            bless( {
              'operands' => [
                bless( {
                  'operands' => [
                    bless( { '' => '2', 'val' => '2' }, 'factor' )
                  ],
                  '' => '2'
                }, 'factorial' )
              ],
              '' => '2'
            }, 'power' )
          ],
          '' => '2'
        }, 'uneg' )
      ],
      '' => '2'
    }, 'term' )
  ],
  '' => '8-4-2',
  'operators' => [
    '-',
    '-'
  ]
}, 'expr' );
2
```

Observamos que la asociatividad es la correcta. El 2 final es el resultado de la evaluación de 8-4-2.

La estructura del árbol se corresponde con la de la gramática:

```
 8  my $rbb = do {
 9      use Regexp::Grammars;
10
11      qr{
12        \A<expr>\z
13
14        <objrule: expr>      <[operands=term]> ** <[operators=addop]>
15
16        <objrule: term>      <[operands=uneg]> ** <[operators=mulop]>
17
18        <objrule: uneg>      <[operators=minus]>* <[operands=power]>
19
20        <objrule: power>     <[operands=factorial]> ** <[operators=powerop]>
21
22        <objrule: factorial> <[operands=factor]>  <[operators=(!)]>*
23
24        <objrule: factor>    <val=([+-]?\d+(?:\.\d*)?)>
25                           | \( <MATCH=expr> \)
26
27        <token: addop>        [+-]
28
29        <token: mulop>        [*/]
30
31        <token: powerop>      \*\*|\^
32
33        <token: minus>        - <MATCH=(?{ 'NEG' })>
34
35      }x;
36  };
```

Ahora, en una segunda parte sobreescribimos los métodos `sem` que describen la semántica para producir una traducción de infijo a postfijo:

```
 66  require PostfixCalc;
 67  test_calc('Translating expressions to postfix (CTRL-D to end in unix) ');
```

Ahora al proporcionar la entrada `6--3!` obtenemos:

```
Translating expressions to postfix (CTRL-D to end in unix)
6--3!
6 3 ! ~ -
```

Aquí `~` es el operador de negación unaria y `!` es el operador factorial.

Estructura de la aplicación

Estos son los ficheros que integran la aplicación:

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ tree
.
|-- EvalCalc.pm            # Soporte para la evaluación de la expresión: sem
|-- Operator.pm            # Soporte a las clases nodo: recorridos
|-- PostfixCalc.pm         # Soporte para la traducción a postfijo: sem
`-- calculator.pl          # programa principal
```

Programa principal

En el programa principal definimos la gramática y escribimos una subrutina `test_calc` que realiza el parsing.

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ cat -n calculator.pl
 1    #!/usr/bin/env perl5.10.1
 2    use strict;
 3    use warnings;
 4    use 5.010;
 5    use Data::Dumper;
 6    $Data::Dumper::Indent = 1;
 7
 8    my $rbb = do {
 9        use Regexp::Grammars;
10
11        qr{
12          \A<expr>\z
13
14          <objrule: expr>      <[operands=term]> ** <[operators=addop]>
15
16          <objrule: term>      <[operands=uneg]> ** <[operators=mulop]>
17
18          <objrule: uneg>      <[operators=minus]>* <[operands=power]>
19
20          <objrule: power>     <[operands=factorial]> ** <[operators=powerop]>
21
22          <objrule: factorial> <[operands=factor]>  <[operators=(!)]>*
23
24          <objrule: factor>    <val=([+-]?\d+(?:\.\d*)?)>
25                               | \( <MATCH=expr> \)
26
27          <token: addop>        [+-]
28
29          <token: mulop>        [*/]
30
31          <token: powerop>      \*\*|\^
32
33          <token: minus>        - <MATCH=(?{ 'NEG' })>
34
35        }x;
36    };
37
38    sub test_calc {
39      my $prompt = shift;
40      my $handler = shift;
41
42      say $prompt;
43      while (my $input = <>) {
44          chomp($input);
45          if ($input =~ m{$rbb}) {
46              my $tree = $/{expr};
47              $handler->($tree) if $handler;
48
49              say $tree->ceval;
```

```
50
51          }
52          else {
53              say("does not match");
54          }
55      }
56    }
57
58    require EvalCalc;
59
60    test_calc(
61      'Evaluating infix arithmetic expressions (CTRL-D to end in unix) ',
62      sub { print &Data::Dumper::Dumper(shift()) },
63    );
64
65
66    require PostfixCalc;
67    test_calc('Translating expressions to postfix (CTRL-D to end in unix) ');
```

Los nodos del AST poseen un método `ceval` que se encarga de realizar la traducción del nodo.

Las Clases de nodos del AST

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ cat -n Operator.pm
   1  #   Class hierarchy diagram:
   2  #  $ vgg -t 'Operator(LeftBinaryOp(expr,term),RightBinaryOp(power),PreUnaryOp(uneg),Post
   3  #                            +---------+
   4  #                            |Operator|
   5  #                            +---------+
   6  #          .---------------.----^--------.-------------.
   7  #    +-------------+ +--------------+ +-----------+ +------------+
   8  #    |LeftBinaryOp| |RightBinaryOp| |PreUnaryOp| |PostUnaryOp|
   9  #    +-------------+ +--------------+ +-----------+ +------------+
  10  #      .---^--.        |               |               |
  11  #    +----+ +----+  +-----+         +----+       +----------+
  12  #    |expr| |term|  |power|         |uneg|       |factorial|
  13  #    +----+ +----+  +-----+         +----+       +----------+
  14  #
  15  #
  16  # NOTE: package "factor" actually implements numbers and is
  17  #       outside this hierarchy
  18  #
  19  package Operator;
  20  use strict;
  21  use Carp;
  22
  23  sub Operands {
  24    my $self = shift;
  25
  26    return () unless exists $self->{operands};
  27    return @{$self->{operands}};
  28  }
  29
  30  sub Operators {
```

```
31    my $self = shift;
32
33    return () unless exists $self->{operators};
34    return @{$self->{operators}};
35  }
36
37  sub sem {
38    confess "not defined sem";
39  }
40
41  sub make_sem {
42    my $class = shift;
43    my %semdesc = @_;
44
45    for my $class (keys %semdesc) {
46      my %sem = %{$semdesc{$class}};
47
48      # Install 'sem' method in $class
49      no strict 'refs';
50      no warnings 'redefine';
51      *{$class."::sem"} = sub {
52        my ($self, $op) = @_;
53        $sem{$op}
54      };
55    }
56  }
57
58  package LeftBinaryOp;
59  use base qw{Operator};
60
61  sub ceval {
62    my $self = shift;
63
64    # recursively evaluate the children first
65    my @operands = map { $_->ceval } $self->Operands;
66
67    # then combine them
68    my $s = shift @operands;
69    for ($self->Operators) {
70      $s = $self->sem($_)->($s, shift @operands);
71    }
72    return $s;
73  }
74
75  package RightBinaryOp;
76  use base qw{Operator};
77
78  sub ceval {
79    my $self = shift;
80
81    # recursively evaluate the children first
82    my @operands = map { $_->ceval } $self->Operands;
83
```

```
 84    # then combine them
 85    my $s = pop @operands;
 86    for (reverse $self->Operators) {
 87      $s = $self->sem($_)->(pop @operands, $s);
 88    }
 89    return $s;
 90  }
 91
 92  package PreUnaryOp;
 93  use base qw{Operator};
 94
 95  sub ceval {
 96    my $self = shift;
 97
 98    # recursively evaluate the children first
 99    my @operands = map { $_->ceval } $self->Operands;
100
101    # then combine them
102    my $s = shift @operands;
103    for (reverse $self->Operators) {
104      $s = $self->sem($_)->($s);
105    }
106    return $s;
107  }
108
109  package PostUnaryOp;
110  use base qw{Operator};
111
112  sub ceval {
113    my $self = shift;
114
115    # recursively evaluate the children first
116    my @operands = map { $_->ceval } $self->Operands;
117
118    # then combine them
119    my $s = shift @operands;
120    for ($self->Operators) {
121      $s = $self->sem($_)->($s);
122    }
123    return $s;
124  }
125
126  package term;
127  use base qw{LeftBinaryOp};
128
129  package expr;
130  use base qw{LeftBinaryOp};
131
132  package power;
133  use base qw{RightBinaryOp};
134
135  package uneg;
136  use base qw{PreUnaryOp};
```

```
137
138  package factorial;
139  use base qw{PostUnaryOp};
140
141  package factor;
142
143  sub ceval {
144    my $self = shift;
145
146    return $self->{val};
147  }
148
149  1;
```

Definiendo sem para la evaluación de la expresión

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ cat -n EvalCalc.pm
 1  package EvalCalc;
 2  use strict;
 3  use Carp;
 4
 5  use Operator;
 6
 7  ####
 8  sub f {
 9    $_[0]>1?$_[0]*f($_[0]-1):1;
10  }
11
12  sub fac {
13    my $n = shift;
14
15    confess "Not valid number" unless $n =~ /^\d+$/;
16    f($n);
17  };
18
19  my $s = sub { shift() **  shift() };
20
21  Operator->make_sem(
22     expr => {
23        '+' => sub { shift()  +  shift() },
24        '-' => sub { shift()  -  shift() },
25     },
26     term => {
27       '*' => sub { shift()  *  shift() },
28       '/' => sub { shift()  /  shift() },
29     },
30     power => {
31        '^'  => $s,
32        '**' => $s,
33     },
34     uneg => {
35        'NEG' => sub { -shift() },
36     },
37     factorial => {
```

```
38        '!' => \&fac,
39     },
40  );
41
42  1;
```

Definiendo `sem` para la traducción a postfijo

```
pl@nereida:~/Lregexpgrammars/demo/calculator$ cat -n PostfixCalc.pm
 1  package PostfixCalc;
 2  use strict;
 3
 4  use Operator;
 5
 6  # Modify semantics: now translate to postfix
 7  my $powers = sub { shift().' '.shift().' **' };
 8
 9  Operator->make_sem(
10     expr => {
11        '+' => sub { shift().' '.shift().' +'  },
12        '-' => sub { shift().' '.shift().' -' },
13     },
14     term => {
15        '*' => sub { shift().' '.shift().' *'  },
16        '/' => sub { shift().' '.shift().' /' },
17     },
18     power => {
19        '^'  => $powers,
20        '**' => $powers,
21     },
22     uneg => {
23        # use ~ for unary minus
24        'NEG' => sub { shift().' ~' },
25     },
26     factorial => {
27        '!' => sub { shift().' !'},
28     },
29  );
30
31  1;
```

Ejercicio 31.10.2.

- *Explique el significado de la primera línea del programa principal*

```
pl@nereida:~/Lregexpgrammars/demo$ cat -n calculator.pl
   1  #!/usr/bin/env perl5.10.1
```

- *Explique el significado de* `$handler` *en* `test_calc`*:*

```
42  sub test_calc {
43    my $prompt = shift;
44    my $handler = shift;
45
46    say $prompt;
47    while (my $input = <>) {
```

```
48        chomp($input);
49        if ($input =~ m{$rbb}) {
50            my $tree = $/{expr};
51            $handler->($tree) if $handler;
52
53            say $tree->ceval;
54
55        }
56        else {
57            say("does not match");
58        }
59    }
60  }
```

- *Aisle las funciones relacionadas con la creación de semántica como* `make_sem`*,* `fac` *y las llamadas a* `make_sem` *en un módulo* `Calculator::Semantics` *aparte.*
- *Añada un traductor de infijo a prefijo al código presentado en esta sección. Una expresión como* `2*3+4` *se traducirá como* `+ * 2 3 4`

31.10.19. Práctica: Calculadora con `Regexp::Grammars`

- Reforme la estructura del ejemplo para que tenga una jerarquía de desarrollo de acuerdo a los estándares de Perl. Use `h2xs` o bien `Module::Starter`. Use el espacio de nombres `Calculator`. Mueva el módulo `Operator` a `Calculator::Operator`. Lea el capítulo Modulos de los apuntes de LHP.
- Defina el conjunto de pruebas que deberá pasar su traductor. Añádalas como pruebas `TODO`. Cuando la funcionalidad a comprobar esté operativa cambie su estatus.
- Añada variables y la expresión de asignación:

  ```
  b = a = 4*2
  ```

  que será traducida a postfijo como:

  ```
  4 2 * a = b =
  ```

  El operador de asignación es asociativo a derechas. El valor devuelto por una expresión de asignación es el valor asignado.

  Use un hash para implantar la relación nombre-valor en el caso de la evaluación
- Introduzca la expresión bloque:

  ```
  c = { a = 4; b = 2*a }
  ```

  Los bloques son listas entre llaves de expresiones separadas por punto y coma. El valor retornado por una expresión bloque es el último evaluado en el bloque.

  El símbolo de arranque de la gramática (esto es, el patrón regular contra el que hay que casar) será la expresión bloque.
- Introduzca las expresiones de comparación `<`, `>`, `<=`, `>=`, `==` y `!=` con la prioridad adecuada. Tenga en cuenta que una expresión como:

```
a = b+2 > c*4
```

deberá entenderse como

```
a = ((b+2) > (c*4))
```

Esto es, se traducirá como:

```
b 2 + c 4 * > a =
```

- Introduzca la expresión `if ... then ... else`. La parte del `else` será opcional:

```
c = if a > 0 then { a = a -1; 2*a } else { b + 2 };
d = if a > 0 then { a = b -1; 2*b };
```

un `else` casa con el `if` mas cercano. La sentencia:

```
if (a > 0) then if (b > 0) then {5} else {6}
```

se interpreta como:

```
if (a > 0) then (if (b > 0) then {5} else {6})
```

y no como:

```
if (a > 0) then (if (b > 0) then {5}) else {6}
```

Se traducirá como:

```
        a
        0
        >
        jz endif124
        b
        0
        >
        jz else125
        5
        j endif126
:else125
        6
:endif124
:endif125
        ...
```

- Escriba un intérprete de la máquina orientada a pila definida en los apartados anteriores. El código generado debería poder ejecutarse correctamente en el intérprete.

Capítulo 32

Análisis Sintáctico Descendente en JavaScript

32.1. Ejemplo Simple de Intérprete: Una Calculadora

1. How to write a simple interpreter in JavaScript

32.2. Análisis Top Down Usando Precedencia de Operadores

1. Véase el libro [9] Beautiful Code: Leading Programmers Explain How They Think, Capítulo 9.
2. Top Down Operator Precedence por Douglas Crockford
3. Top Down Operator Precedence demo por Douglas Crockford
4. jslint

32.2.1. Gramática de JavaScript

1. Especificación de JavaScript 1997
2. NQLL(1) grammar (Not Quite LL(1)) for JavaScrip 1997
3. Postscript con la especificación de JavaScript 1997
4. Mozilla JavaScript Language Resources
5. JavaScript 1.4 LR(1) Grammar 1999.
6. Apple JavaScript Core Specifications
7. Creating a JavaScript Parser Una implementación de ECAMScript 5.1 usando Jison disponible en GitHub en https://github.com/cjihrig/jsparser.

Capítulo 33

Análisis Sintáctico Descendente en Perl

Este capítulo tiene por objeto darte una visión global de la estructura de un compilador e introducirte en las técnicas básicas de la construcción de compiladores usando Perl.

Puesto que no todos los alumnos que se incorporan en este capítulo han leído los anteriores y no necesariamente conocen Perl, en la sección 33.1 comenzamos haciendo un breve repaso a como construir un módulo en Perl. Si quieres tener un conocimiento mas profundo lee el capítulo sobre módulos en [10].

La sección 33.2 describe las fases en las que -al menos conceptualmente- se divide un compilador. A continuación la sección 33.3 presenta la primera de dichas fases, el análisis léxico. En la sección 34.1 repasamos conceptos de análisis sintáctico que deberían ser familiares a cualquiera que haya seguido un curso en teoría de autómatas y lenguajes formales. Antes de comenzar a traducir es conveniente tener un *esquema* o estrategia de traducción para cada constructo sintáctico. La sección 33.7 introduce el concepto de esquema de traducción. La fase de análisis sintáctico consiste en la construcción del *árbol de análisis* a partir de la *secuencia de unidades léxicas*. Existen diversas estrategias para resolver esta fase. En la sección 33.6 introducimos la que posiblemente sea la mas sencilla de todas: el análisis descendente predictivo recursivo. En la sección 33.8 abordamos una estrategia para transformar ciertas gramáticas para las que dicho método no funciona.

Un analizador sintáctico implícitamente construye el árbol de análisis concreto. En muchas ocasiones resulta mas rentable trabajar con una forma simplificada (*abstracta*) del árbol que contiene la misma información que aquél. La sección 33.9 trata de la construcción de los árboles de análisis abstractos.

33.1. Las Bases

Puesto que no todos los alumnos que están interesados en esta sección tienen conocimientos previos de Perl, en esta sección comenzamos haciendo un breve repaso a como construir un módulo en Perl y al mismo tiempo repasamos las características usadas del lenguaje. Si quieres tener un conocimiento mas profundo de como construir un módulo, lee el capítulo sobre módulos en [10].

Version

El comportamiento de Perl puede variar ligeramente si la versión que tenemos instalada es antigua. Para ver la versión de Perl podemos hacer.

```
lhp@nereida:~/Lperl/src/topdown/PL0506$ perl -v

This is perl, v5.8.4 built for i386-linux-thread-multi

Copyright 1987-2004, Larry Wall

Perl may be copied only under the terms of either the Artistic License or the
GNU General Public License, which may be found in the Perl 5 source kit.
```

```
Complete documentation for Perl, including FAQ lists, should be found on
this system using 'man perl' or 'perldoc perl'.  If you have access to the
Internet, point your browser at http://www.perl.com/, the Perl Home Page.
```

h2xs

En primer lugar, construimos la estructura para nuestro proyecto de mini-lenguaje. La mejor forma de comenzar a escribir un módulo es usando la herramienta Perl `h2xs`. Supongamos que queremos construir un módulo `PL::Tutu`. Los nombres de los módulos siguen un esquema de identificadores separados por una pareja de `:`. Para saber más sobre el esquema de nombres de los módulos y la forma en la que estos se asigna a ficheros del sistema operativo, lea la sección sobre introducción a los módulos [10].

```
lhp@nereida:~/Lperl/src/topdown/PL0506$ h2xs -XA -n PL::Tutu
Defaulting to backwards compatibility with perl 5.8.4
If you intend this module to be compatible with earlier perl versions, please
specify a minimum perl version with the -b option.

Writing PL-Tutu/lib/PL/Tutu.pm
Writing PL-Tutu/Makefile.PL
Writing PL-Tutu/README
Writing PL-Tutu/t/PL-Tutu.t
Writing PL-Tutu/Changes
Writing PL-Tutu/MANIFEST
```

La herramienta `h2xs` fué concebida para ayudar en la transformación de ficheros de cabecera de C en código Perl. La opción `-X` hace que se omita la creación de subrutinas externas (`XS`) La opción `-A` implica que el módulo no hará uso del `AutoLoader`. La opción `-n` proporciona el nombre del módulo. La llamada a `h2xs` crea la siguiente estructura de directorios y ficheros:

```
lhp@nereida:~/Lperl/src/topdown/PL0506$ tree
.
'-- PL-Tutu
    |-- Changes
    |-- MANIFEST
    |-- Makefile.PL
    |-- README
    |-- lib
    |   '-- PL
    |       '-- Tutu.pm
    '-- t
        '-- PL-Tutu.t

4 directories, 6 files
```

Generación del Makefile

Después de esto tenemos un módulo "funcional" que no hace nada. Lo podríamos instalar como si lo hubieramos descargado desde CPAN (Véase [10]).

Después cambiamos al directorio `PL-Tutu/` y hacemos `perl Makefile.PL`.

```
lhp@nereida:~/Lperl/src/topdown/PL0506$ cd PL-Tutu/
lhp@nereida:~/Lperl/src/topdown/PL0506/PL-Tutu$ perl Makefile.PL
Checking if your kit is complete...
Looks good
Writing Makefile for PL::Tutu
```

Esto crea el fichero `Makefile` necesario para actualizar nuestra aplicación. Para saber más sobre `perl Makefile.PL` lea [10].

Documentación

Pasamos ahora a trabajar en el módulo. Primero escribimos la parte relativa a la documentación. Para ello editamos `Tutu.pm`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/PL-Tutu/lib/PL$ pwd
/home/lhp/Lperl/src/topdown/PL0506/PL-Tutu/lib/PL
lhp@nereida:~/Lperl/src/topdown/PL0506/PL-Tutu/lib/PL$ ls -l
total 4
-rw-r--r--  1 lhp lhp 2343 2005-09-28 11:16 Tutu.pm
```

y al final del mismo insertamos la documentación. Para saber más sobre el lenguajes de marcas de Perl (*pod* por *plain old documentation*) lea [10]. En este caso escribimos:

```
1;
__END__

=head1 NOMBRE

PL::Tutu - Compilador para un lenguaje sencillo denominado
          "Tutu" que usaremos en la asignatura PL

=head1 SINOPSIS

  use PL::Tutu;

  La subrutina PL::Tutu::compiler recibe dos argumentos: el
  nombre del fichero de entrada (fuente.tutu) y el nombre del fichero de
  salida (código ensamblador para una especie de P-máquina).

=head1 DESCRIPCIÓN

Este módulo tiene dos objetivos: aprender a hacer un pequeño compilador
y aprender a programar modularmente en Perl, usando un buen número
de los recursos que este lenguaje ofrece.

El siguiente es un ejemplo de código fuente tutu:

  int a,b;
  string c;
  a = 2+3;
  b = 3*4;
  c = "hola";
  p c;
  c = "mundo";
  p c;
  p 9+2;
  p a+1;
  p b+1

supuesto que está guardado en el fichero "test2.tutu", podemos
escribir un programa Perl "main.pl" para compilarlo:
```

```
    $ cat main.pl
    #!/usr/bin/perl -w -I..
    #use PL::Tutu;
    use Tutu;

    PL::Tutu::compiler(@ARGV);

al ejecutar "main.pl":

    $ ./main.pl test2.tutu test2.ok

obtenemos el fichero "test2.ok" con el ensamblador:
    $ cat test2.ok
    DATA holamundo
    PUSH 5
    PUSHADDR 0
    STORE_INT
    PUSH 12
    PUSHADDR 1
    STORE_INT
    PUSHSTR 0 4
    PUSHADDR 2
    STORE_STRING
    LOAD_STRING 2
    PRINT_STR
    PUSHSTR 4 5
    PUSHADDR 2
    STORE_STRING
    LOAD_STRING 2
    PRINT_STR
    PUSH 11
    PRINT_INT
    LOAD 0
    INC
    PRINT_INT
    LOAD 1
    INC
    PRINT_INT

Para mas información consulta la página de la asignatura.
¡Buena suerte!

=head2 EXPORT

No se exporta nada al espacio de nombres del cliente.


=head1 AUTOR

Casiano Rodríguez León, E<lt>casiano@ull.esE<gt>

=head1 VÉASE TAMBIÉN
```

```
L<perl>.

=cut
```

La documentación puede ser mostrada utilizando el comando `perldoc`.
Veáse la figura 33.1.

Figura 33.1: El resultado de usar `perldoc Tutu`

La forma en la que se controla la calidad de un módulo es mediante el desarrollo de pruebas. Las pruebas son programas perl que se sitúan en el directorio `t/` y que tienen la extensión `.t`. Para ejecutar las pruebas se escribe:

```
make test
```

Vease la sección [10] de los apuntes de LHP para ams detalles.

33.1.1. Repaso: Las Bases

Responda a las siguientes preguntas:

1. ¿Cómo puedo saber con que versión de Perl estoy trabajando?
2. Cuando el intérprete Perl encuentra una sentencia

   ```
   use Este::Modulo;
   ```

   ¿Donde busca el fichero `Modulo.pm`?
3. ¿Con que opción debo usar Perl para ejecutar un programa en la línea de comandos?
4. ¿Cómo se llama el programa que me permite crear el esqueleto para una distribución de un módulo? ¿Con que opciones debo llamarlo?
5. Cuando se crea con `h2xs` el esqueleto para `PL::Tutu`: ¿En que subdirectorio queda el fichero conteniendo el esqueleto del módulo creado `Tutu.pm`?
6. ¿Cuál es la función de `MANIFEST`?
7. ¿Qué es `Makefile.PL`? ¿Cuál es su función? ¿Que significa la frase *looks good*?
8. ¿Con que comando se crea el `Makefile` para trabajar en la plataforma actual?
9. ¿Cómo se puede ver la documentación de un módulo?
10. ¿Que hacen los siguientes comandos `pod`? Repase [10] si tiene dudas.

    ```
    =head1 cabecera
    =head2 cabecera
    =item texto
    =over N
    =back
    =cut
    =pod
    =for X
    =begin X
    =end X
    ```

11. ¿Que secuencia de comandos conocida como *mantra de instalación* es necesario ejecutar para instalar un módulo?

12. ¿Cual es la función del directorio `t`?

13. ¿Que tipo deben tener los programas de prueba para que `make test` los reconozca como pruebas?

33.1.2. Práctica: Crear y documentar el Módulo `PL::Tutu`

Reproduzca los pasos explicados en la sección 33.1 creando el módulo `PL::Tutu` y documentándolo.

- Compruebe que la documentación se muestra correctamente
- Compruebe que puede crear una distribución haciendo `make dist`
- Compruebe que la distribución creada puede instalarse correctamente siguiendo las instrucciones en [10].

33.2. Las Fases de un Compilador

La estructura del compilador, descompuesto en fases, queda explicitada en el código de la subrutina `compile`:

Esquema del Compilador

```
 1 package PL::Tutu;
 2 use 5.008004;   # Versión mínima de Perl 5.84
 3 use strict;     # Variables deben ser declaradas, etc.
 4 use warnings;   # Enviar warnings
 5 use IO::File;
 6 use Carp;       # Provee alternativas a "die" and "warn"
 7
 8 require Exporter;
 9
10 our @ISA = qw(Exporter);   # Heredamos los métodos de la clase Exporter
11 our @EXPORT = qw( compile compile_from_file); # Estas funciones serán exportadas
12 our $VERSION = '0.01';     # Variable que define la versión del módulo
13
14 our %symbol_table;         # La tabla de símbolos $symbol_table{x} contiene
15 our $data;                 # la información asociada con el objeto 'x'
16 our $target;               # tipo, dirección, etc.
17 our @tokens;               # La lista de terminales
18 our $errorflag;
19 our ($lookahead, $value);  # Token actual y su atributo
20 our $tree;                 # referencia al objeto que contiene
21 our $global_address;       # el árbol sintáctico
22
23 # Lexical analyzer
24 package Lexical::Analysis;
25 sub scanner {
26 }
27
28 package Syntax::Analysis;
29 sub parser {
30 }
```

```
31
32 package Machine::Independent::Optimization;
33 sub Optimize {
34 }
35
36 package Code::Generation;
37 sub code_generator {
38 }
39
40 package Peephole::Optimization;
41 sub transform {
42 }
43
44 package PL::Tutu;
45 sub compile {
46   my ($input) = @_; # Observe el contexto!
47   local %symbol_table = ();
48   local $data = ""; # Contiene todas las cadenas en el programa fuente
49   local $target = ""; # target code
50   local @tokens =();    # "local" salva el valor que serÃ¡ recuperado al finalizar
51   local $errorflag = 0; # el Ã¡mbito
52   local ($lookahead, $value) = ();
53   local $tree = undef; # Referencia al Ã¡rbol sintÃctico abstracto
54   local $global_address = 0; # Usado para guardar la Ãltima direcciÃn ocupada
55
56   ########lexical analysis
57   &Lexical::Analysis::scanner($input);
58
59   ########syntax (and semantic) analysis
60   $tree = &Syntax::Analysis::parser;
61
62   ########machine independent optimizations
63   &Machine::Independent::Optimization::Optimize;
64
65   ########code generation
66   &Code::Generation::code_generator;
67
68   ########peephole optimization
69   &Peephole::Optimization::transform($target);
70
71   return \$target; #retornamos una referencia a $target
72 }
73
74 sub compile_from_file {
75   my ($input_name, $output_name) = @_; # Nombres de ficheros
76   my $fhi;                             # de entrada y de salida
77   my $targetref;
78
79   if (defined($input_name) and (-r $input_name)) {
80     $fhi = IO::File->new("< $input_name");
81   }
82   else { $fhi = 'STDIN'; }
83   my $input;
```

```
84   { # leer todo el fichero
85     local $/ = undef; # localizamos para evitar efectos laterales
86     $input = <$fhi>;
87   }
88   $targetref = compile($input);
89
90   ########code output
91   my $fh = defined($output_name)? IO::File->new("> $output_name") : 'STDOUT';
92   $fh->print($$targetref);
93   $fh->close;
94   1; # El Ã°ltimo valor evaluado es el valor retornado
95 }
96
97 1;  # El 1 indica que la fase de carga termina con éxito
98 # Sigue la documentación ...
```

Añadiendo Ejecutables

Vamos a añadir un *script* que use el módulo `PL::Tutu` para asi poder ejecutar nuestro compilador:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/$ mkdir scripts
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ cd scripts/
```

A continuación creamos dos versiones del compilador `tutu.pl` y `tutu` y un programa de prueba `test01.tutu`:

```
... # despues de crear los ficheros
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ ls
test01.tutu  tutu  tutu.pl
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ cat tutu.pl
#!/usr/bin/perl -w -I../lib/
use PL::Tutu;

PL::Tutu::compile_from_file(@ARGV);
```

Búsqueda de Librerías en Tiempo de Desarrollo

El programa `tutu` ilustra otra forma de conseguir que el intérprete Perl busque por la librería que está siendo desarrollada, mediante el uso de `use lib`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ cat tutu
#!/usr/bin/perl -w
use lib ('../lib');
use PL::Tutu;

&PL::Tutu::compile_from_file(@ARGV);
```

Una tercera forma (la que recomiendo):

```
$ export PERL5LIB=~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/lib
$ perl -MPL::Tutu -e 'PL::Tutu::compile_from_file("test01.tutu")'
```

Añadiendo Los Ejecutables al `MANIFEST`

Ahora tenemos que añadir estos ficheros en `MANIFEST` para que formen parte del proyecto. En vez de eso lo que podemos hacer es crear un fichero `MANIFEST.SKIP`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ cat MANIFEST.SKIP
\.o$
^\.cvsignore$
/\.cvsignore$
\.cvsignore$
CVS/[^/]+$
\.svn\b
^Makefile$
/Makefile$
^blib/
\.swp$
\.bak$
\.pdf$
\.ps$
\.sal$
pm_to_blib
\.pdf$
\.tar.gz$
\.tgz$
^META.yml$
```

Ahora al hacer

```
make manifest
```

se crea un fichero `MANIFEST` que contiene los caminos relativos de todos los ficheros en la jerarquía cuyos nombres no casan con una de las expresiones regulares en `MANIFEST.SKIP`.

Para saber mas sobre `MANIFEST` léa [10].

No recomiendo el uso de `MANIFEST.SKIP`. Prefiero un control manual de los ficheros que integran la aplicacion.

Indicando al Sistema de Distribución que los Ficheros son Ejecutables

Es necesario indicarle a Perl que los ficheros añadidos son ejecutables. Esto se hace mediante el parámetro `EXE_FILES` de `WriteMakefile`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ cat Makefile.PL
use 5.008004;
use ExtUtils::MakeMaker;
# See lib/ExtUtils/MakeMaker.pm for details of how to influence
# the contents of the Makefile that is written.
WriteMakefile(
    NAME              => 'PL::Tutu',
    VERSION_FROM      => 'lib/PL/Tutu.pm', # finds $VERSION
    PREREQ_PM         => {}, # e.g., Module::Name => 1.1
    EXE_FILES         => [ 'scripts/tutu.pl', 'scripts/tutu' ],
    ($] >= 5.005 ?     ## Add these new keywords supported since 5.005
      (ABSTRACT_FROM  => 'lib/PL/Tutu.pm', # retrieve abstract from module
       AUTHOR         => 'Lenguajes y Herramientas de Programacion <lhp@>') : ()),
);
```

Perl utilizará esa información durante la fase de instalación para instalar los ejecutables en el *path* de búsqueda.

Reconstrucción de la aplicación

A continuación hay que rehacer el `Makefile`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ perl Makefile.PL
Writing Makefile for PL::Tutu
```

Para crear una versión funcional hacemos `make`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ make
cp scripts/tutu blib/script/tutu
/usr/bin/perl "-MExtUtils::MY" -e "MY->fixin(shift)" blib/script/tutu
cp scripts/tutu.pl blib/script/tutu.pl
/usr/bin/perl "-MExtUtils::MY" -e "MY->fixin(shift)" blib/script/tutu.pl
Manifying blib/man3/PL::Tutu.3pm
```

Para crear el `MANIFEST` hacemos `make manifest`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ make manifest
/usr/bin/perl "-MExtUtils::Manifest=mkmanifest" -e mkmanifest
Added to MANIFEST: scripts/tutu
```

Comprobemos que el test de prueba generado automáticamente por `h2xs` se pasa correctamente:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ make test
PERL_DL_NONLAZY=1 /usr/bin/perl "-MExtUtils::Command::MM" "-e" "test_harness(0, 'blib/lib', 'b
t/PL-Tutu....ok
All tests successful.
Files=1, Tests=1,  0 wallclock secs ( 0.08 cusr +  0.00 csys =  0.08 CPU)
```

Ejecución

Podemos ahora ejecutar los guiones:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ cd scripts/
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ ls -l
total 12
-rw-r--r--  1 lhp lhp 15 2005-09-29 12:56 test01.tutu
-rwxr-xr-x  1 lhp lhp 92 2005-09-29 13:29 tutu
-rwxr-xr-x  1 lhp lhp 80 2005-09-29 12:58 tutu.pl
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ tutu test01.tutu test01.sal
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ ls -l
total 12
-rw-r--r--  1 lhp lhp  0 2005-09-29 13:53 test01.sal
-rw-r--r--  1 lhp lhp 15 2005-09-29 12:56 test01.tutu
-rwxr-xr-x  1 lhp lhp 92 2005-09-29 13:29 tutu
-rwxr-xr-x  1 lhp lhp 80 2005-09-29 12:58 tutu.pl
```

Veamos los contenidos del programa fuente `test01.tutu` que usaremos para hacer una prueba:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu/scripts$ cat test01.tutu
int a,b;
a = 4
```

Construcción de una Distribución

Para hacer una distribución instalable hacemos `make dist`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ make dist
rm -rf PL-Tutu-0.01
/usr/bin/perl "-MExtUtils::Manifest=manicopy,maniread" \
        -e "manicopy(maniread(),'PL-Tutu-0.01', 'best');"
mkdir PL-Tutu-0.01
```

```
mkdir PL-Tutu-0.01/scripts
mkdir PL-Tutu-0.01/lib
mkdir PL-Tutu-0.01/lib/PL
mkdir PL-Tutu-0.01/t
tar cvf PL-Tutu-0.01.tar PL-Tutu-0.01
PL-Tutu-0.01/
PL-Tutu-0.01/scripts/
PL-Tutu-0.01/scripts/test01.tutu
PL-Tutu-0.01/scripts/tutu
PL-Tutu-0.01/scripts/tutu.pl
PL-Tutu-0.01/META.yml
PL-Tutu-0.01/Changes
PL-Tutu-0.01/MANIFEST
PL-Tutu-0.01/lib/
PL-Tutu-0.01/lib/PL/
PL-Tutu-0.01/lib/PL/Tutu.pm
PL-Tutu-0.01/MANIFEST.SKIP
PL-Tutu-0.01/t/
PL-Tutu-0.01/t/PL-Tutu.t
PL-Tutu-0.01/Makefile.PL
PL-Tutu-0.01/README
rm -rf PL-Tutu-0.01
gzip --best PL-Tutu-0.01.tar
```

Después de esto tenemos en el directorio de trabajo el fichero `PL-Tutu-0.01.tar.gz` con la distribución:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/02fases/PL-Tutu$ ls -ltr
total 72
drwxr-xr-x  2 lhp lhp  4096 2005-09-29 12:01 t
-rw-r--r--  1 lhp lhp  1196 2005-09-29 12:01 README
drwxr-xr-x  3 lhp lhp  4096 2005-09-29 12:01 lib
-rw-r--r--  1 lhp lhp   152 2005-09-29 12:01 Changes
-rw-r--r--  1 lhp lhp   167 2005-09-29 13:23 MANIFEST.SKIP
-rw-r--r--  1 lhp lhp     0 2005-09-29 13:23 pm_to_blib
drwxr-xr-x  6 lhp lhp  4096 2005-09-29 13:23 blib
-rw-r--r--  1 lhp lhp   113 2005-09-29 13:23 MANIFEST.bak
drwxr-xr-x  2 lhp lhp  4096 2005-09-29 13:29 scripts
-rw-r--r--  1 lhp lhp   616 2005-09-29 13:49 Makefile.PL
-rw-r--r--  1 lhp lhp 20509 2005-09-29 13:51 Makefile
-rw-r--r--  1 lhp lhp  3654 2005-09-29 16:34 PL-Tutu-0.01.tar.gz
-rw-r--r--  1 lhp lhp   298 2005-09-29 16:34 META.yml
-rw-r--r--  1 lhp lhp   205 2005-09-29 16:34 MANIFEST
```

33.2.1. Repaso: Fases de un Compilador

1. ¿Que hace la declaración `package nombredepaquete`?
2. ¿Cual es la función de la declaración `use 5.008004`?
3. ¿Cuál es la función de la declaración `use strict`?
4. ¿Cuál es la función de la declaración `use warnings`?
5. ¿Que diferencia hay entre `use warnings` y `perl -w`?
6. ¿Cuál es la función de la declaración `use Carp`? ¿Que diferencia hay entre `croak` y `die`?

7. ¿Qué hace la declaración `our`?
8. ¿Qué es una variable de paquete?
9. ¿Cuál es el nombre completo de una variable de paquete?
10. ¿En que variable especial se situán los argumentos pasados a una subrutina?
11. ¿Que hace la declaración `local`?
12. ¿Cómo se declara una variable léxica?
13. ¿Cuál es el prefijo para los hashes?
14. ¿Cómo se hace referencia a un elemento de un hash `%h` de clave `k`?
15. ¿Cómo se hace referencia a un elemento de un array `@a` de índice 7? ¿Que lugar ocupa ese elemento en el array?
16. ¿Cuál es el significado de `undef`?
17. ¿Cuál es el prefijo para las subrutinas?
18. Señale la diferencia entre

    ```
    my ($input) = @_;
    ```

    y

    ```
    my $input = @_;
    ```

    Repase [10].
19. Toda referencia es un escalar: ¿Cierto o falso?
20. Toda referencia es verdadera ¿Cierto o falso?
21. ¿Que diferencia hay entre `use` y `require`? ¿La línea `require Exporter` se ejecuta en tiempo de compilación o en tiempo de ejecución?
22. ¿Que hace la línea `our @ISA = qw(Exporter)`?. Repáse [10].
23. ¿Que hace la línea `our @EXPORT = qw( compile compile_from_file)`?
24. ¿Que diferencia hay entre `EXPORT` y `EXPORT_OK`?. Repase [10].
25. ¿Que hace la línea `our $VERSION = '0.01`?
26. ¿Que valor tiene una variable no incializada? ¿y si es un array?
27. ¿Que es un array anónimo? (Repase [10])
28. ¿Que es un hash anónimo? (Repase [10])
29. ¿Que hace el operador `=>`?. Repase [10].
30. ¿En que lugar se dejan los ejecutables asociados con una distribución? ¿Cómo se informa a Perl que se trata de ejecutables?
31. ¿Cuál es la función de `MANIFEST.SKIP`? ¿Que hace `make manifest`?
32. ¿Que hace la opción `-I`? ¿Porqué la primera línea de `tutu.pl` comienza: `#!/usr/bin/perl -w -I../lib/`?

33. ¿Cómo puedo saber lo que hace el módulo `lib`? ¿Qué hace la línea `use lib ('../lib')` en el programa `tutu`?

34. ¿Que contiene la variable `PERL5LIB`?

35. ¿Cómo se crea una distribución?

36. ¿Que devuelve `-r $input_name` en la línea 79? Repase [10].

37. ¿Cuál es la función de la variable mágica `$/`? ¿Que se leerá en la línea 86

```
85   local $/ = undef;
86   my $input = <$fhi>;
```

38. ¿Que hace el operador \? ¿Que relación hay entre `\$target` y `$target`?.

39. Si `$targetref` es una referencia a la cadena que va a contener el código objeto, ¿Cómo se denota a la cadena referenciada por `$targetref`? Explique la línea

```
92   $fh->print($$targetref);
```

33.2.2. Práctica: Fases de un Compilador

Reproduzca los pasos explicados en la sección 33.2 extendiendo el módulo `PL::Tutu` con las funciones de compilación y los correspondientes guiones de compilación.

Mejore el script `tutu` para que acepte opciones desde la línea de comandos. Debera soportar al menos las siguientes opciones:

- `--usage`

  Muestra de forma concisa el comando de uso

- `--help`

  Un resumen de cada opción disponible

- `--version`

  Muestra la versión del programa

- `--man`

  Muestra la documentación

Use para ello el módulo `Getopt::Long` . Este módulo provee la función `GetOptions` la cual se atiene a los estándares de especificación de opciones en la línea de comandos POSIX y GNU. Esta función soporta el uso del guión doble `--` y el simple así como admitir el prefijo mas corto que deshace la ambiguedad entre las diferentes opciones.

La llamada a `GetOptions` analiza la línea de comandos en `ARGV` inicializa la variable asociada de manera adecuada. Retorna un valor verdadero si la línea de comandos pudo ser procesada con En caso contrario emitirá un mensaje de error y devolverá falso. Recuerde hacer `perldoc Getopt::Long` para obtener información mas detallada

El siguiente ejemplo ilustra el uso de `Getopt::Long`. Se hace uso también del módulo (función `pod2usage` en la línea 63) `Pod::Usage` el cual permite la documentación empotrada.

```
nereida:~/LEyapp/examples> cat -n treereg
 1  #!/usr/bin/perl -w
 2  use strict;
 3  use Parse::Eyapp::YATW;
 4  use Parse::Eyapp::Node;
```

```
 5  use Parse::Eyapp::Treeregexp;
 6  use Carp;
 7  use Getopt::Long;
 8  use Pod::Usage;
 9
10  my $infile;
11  my $outfile;
12  my $packagename;
13  my $prefix = '';
14  my $syntax = 1;
15  my $numbers = 1;
16  my $severity = 0; # 0 = Don't  check arity. 1 = Check arity.
17                    # 2 = Check arity and give a warning 3 = ... and croak
18  GetOptions(
19    'in=s'       => \$infile,
20    'out=s'      => \$outfile,
21    'mod=s'      => \$packagename,
22    'prefix=s'   => \$prefix,
23    'severity=i' => \$severity,
24    'syntax!'    => \$syntax,
25    'numbers!'   => \$numbers,
26    'version'    => \&version,
27    'usage'      => \&usage,
28    'help'       => \&man,
29  ) or croak usage();
30
31  # If an argument remains is the inputfile
32  ($infile) = @ARGV unless defined($infile);
33  die usage() unless defined($infile);
34
35  my $treeparser = Parse::Eyapp::Treeregexp->new(
36                       INFILE   => $infile,
37                       OUTFILE  => $outfile,
38                       PACKAGE  => $packagename,
39                       PREFIX   => $prefix,
40                       SYNTAX   => $syntax,
41                       NUMBERS  => $numbers,
42                       SEVERITY => $severity
43                     );
44
45  $treeparser->generate();
46
47  sub version {
48    print "Version $Parse::Eyapp::Treeregexp::VERSION\n";
49    exit;
50  }
51
52  sub usage {
53    print <<"END_ERR";
54  Supply the name of a file containing a tree grammar (.trg)
55  Usage is:
56  treereg [-m packagename] [[no]syntax] [[no]numbers] [-severity 0|1|2|3] \
57          [-p treeprefix] [-o outputfile] -i filename[.trg]
```

```
58  END_ERR
59    exit;
60  }
61
62  sub man {
63    pod2usage(
64      -exitval => 1,
65      -verbose => 2
66    );
67  }
68  __END__
69
70  =head1 SYNOPSIS
71
72    treereg [-m packagename] [[no]syntax] [[no]numbers] [-severity 0|1|2|3] \
73            [-p treeprefix] [-o outputfile] -i filename[.trg]
74    treereg [-m packagename] [[no]syntax] [[no]numbers] [-severity 0|1|2|3] \
75            [-p treeprefix] [-o outputfile] filename[.trg]
76 ... # Follows the documentation bla, bla, bla
```

Ahora podemos ejecutar el guión de múltiples formas:

```
nereida:~/LEyapp/examples> treereg -nos -nonu -se 3 -m Tutu Foldonly1.trg
nereida:~/LEyapp/examples> treereg -nos -nonu -s 3 -m Tutu Foldonly1.trg
Option s is ambiguous (severity, syntax)
nereida:~/LEyapp/examples> treereg -nos -bla -nonu -m Tutu Foldonly1.trg
Unknown option: bla
nereida:~/LEyapp/examples>
```

La librería estandar de Perl incluye el módulo `Getopt::Long`. No es el caso de `Pod::Usage`. Descarge el módulo e instalelo en un directorio local en el que tenga permisos. Si es preciso repase las secciones [10] y [10] de los apuntes de introducción a Perl.

33.3. Análisis Léxico

Comenzaremos con la parte mas sencilla del compilador: el analizador léxico. Habitualmente el término "análisis léxico" se refiere al tratamiento de la entrada que produce como salida la lista de *tokens*. Un *token* hace alusión a las unidades mas simples que tiene significado. Habitualmente un *token* o lexema queda descrito por una expresión regular. Léxico viene del griego *lexis*, que significa "palabra". Perl es, sobra decirlo, una herramienta eficaz para encontrar en que lugar de la cadena se produce un emparejamiento. Sin embargo, en el análisis léxico, el problema es encontrar la subcadena a partir de la última posición en la que se produjo un emparejamiento y que es aceptada por una de las expresiones regulares que definen los lexemas del lenguaje dado.

La estructura general del analizador léxico consiste en un bucle en el que se va recorriendo la entrada, buscando por un emparejamiento con uno de los patrones/lexemas especificados y, cuando se encuentra, se retorna esa información al analziador sintáctico. Como no tenemos escrito el analaizador sintáctico simplemente iremos añadiéndo los terminales al final de una lista.

Una iteración del bucle tiene la forma de una secuencia de condicionales en las que se va comprobando si la entrada casa con cada una de las expresiones regulares que definen los terminales del lenguaje. Las condiciones tienen un aspecto similar a este:

```
...
if (m{\G\s*(\d+)}gc) {
  push @tokens, 'NUM', $1;
```

```
}
elsif (m{\G\s*([a-z_]\w*)\b}igc) {
  push @tokens, 'ID', $1;
}
...
```

Una expresión como `m{\G\s*(\d+)}gc` es una expresión regular. Es conveniente que en este punto repase la introducción a las expresiones regulares en [10].

El Operador de Binding

Nótese que, puesto que no se menciona sobre que variable se hace el *binding* (no se usa ninguno de los operadores de *binding* `=~` y `!~`): se entiende que es sobre la variable por defecto `$_` sobre la que se hace el matching.

Casando a partir del Ültimo Emparejamiento

El ancla `\G` casa con el punto en la cadena en el que terminó el último emparejamiento.

La expresión regular describiendo el patrón de interés se pone entre paréntesis para usar la estrategia de los paréntesis con memoria (véanse 31.1.4 y 31.1.1).

Las opciones `c` y `g` son especialmente útiles para la construcción del analizador.

La opción `g`

Como lo usamos en un contexto escalar, la opción `g` *itera sobre la cadena, devolviendo cierto cada vez que casa, y falso cuando deja de casar.* Se puede averiguar la posicion del emparejamiento utilizando la función `pos`. (véase la sección 31.1.6 para mas información sobre la opción `g`).

La opción `c`

La opción `/c` afecta a las operaciones de emparejamiento con `/g` en un contexto escalar. Normalmente, *cuando una búsqueda global escalar tiene lugar y no ocurre casamiento, la posición de comienzo de búsqueda es reestablecida* al comienzo de la cadena. La opción `/c` hace que la posición inicial de emparejamiento permanezca donde la dejó el último emparejamiento con éxito y no se vaya al comienzo. Al combinar esto con el ancla `\G`, la cuál casa con el final del último emparejamiento, obtenemos que la combinación

```
m{\G\s*(...)}gc
```

logra el efecto deseado: Si la primera expresión regular en la cadena `elsif` fracasa, la posición de búsqueda no es inicializada de nuevo gracias a la opción `c` y el ancla `\G` sigue recordando donde terminó el ultimo casamiento.

La opción `i`

Por último, la opción `i` permite ignorar el tipo de letra (mayúsculas o minúsculas).

Repase la sección 31.1.6 para ver algunas de las opciones mas usadas.

Código del Analizador Léxico

Este es el código completo de la subrutina `scanner` que se encarga del análisis léxico:

```
1 package Lexical::Analysis;
2 sub scanner {
3   local $_ = shift;
4   { # Con el redo del final hacemos un bucle "infinito"
5     if (m{\G\s*(\d+)}gc) {
6       push @tokens, 'NUM', $1;
7     }
```

```
 8     elsif (m{\G\s*int\b}igc) {
 9       push @tokens, 'INT', 'INT';
10     }
11     elsif (m{\G\s*string\b}igc) {
12       push @tokens, 'STRING', 'STRING';
13     }
14     elsif (m{\G\s*p\b}igc) {
15       push @tokens, 'P', 'P'; # P para imprimir
16     }
17     elsif (m{\G\s*([a-z_]\w*)\b}igc) {
18       push @tokens, 'ID', $1;
19     }
20     elsif (m{\G\s*\"([^"]*)\"}igc) {
21       push @tokens, 'STR', $1;
22     }
23     elsif (m{\G\s*([+*()=;,])}gc) {
24       push @tokens, 'PUN', $1;
25     }
26     elsif (m{\G\s*(\S)}gc) { # Hay un caracter "no blanco"
27       Error::fatal "Caracter invalido: $1\n";
28     }
29     else {
30       last;
31     }
32     redo;
33   }
34 }
```

Relación de corutina con el Analizador Sintáctico

Si decidieramos establecer una relación de corutina con el analizador léxico los condicionales se pueden programar siguiendo secuencias con esta estructura:

```
      return ('INT', 'INT') if (m{\G\s*int\b}igc);
```

Manejo de Errores

Para completar el analizador solo quedan declarar las variables usadas y las subrutinas de manejo de errores:

```
######## global scope variables
our @tokens = ();
our $errorflag = 0;

package Error;

sub error($) {
  my $msg = shift;
  if (!$errorflag) {
    warn "Error: $msg\n";
    $errorflag = 1;
  }
}

sub fatal($) {
  my $msg = shift;
```

```
  die "Error: $msg\n";
}
```

El uso de `our` es necesario porque hemos declarado al comienzo del módulo `use strict`. El pragma `use strict` le indica al compilador Perl que debe considerar como obligatorias un conjunto de reglas de buen estilo de programación. Entre otras restricciones, el uso del pragma implica que todas las variables (no-mágicas) deben ser declaradas explícitamente (uso de `my`, `our`, etc.) La declaración `our` se describe en [10].

33.3.1. Ejercicio: La opción g

Explique cada una de las líneas que siguen:

```
$ perl -wde 0
main::(-e:1):   0
  DB<1> $x = "abababab"
  DB<2> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 2
  DB<3> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 4
  DB<4> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 6
  DB<5> print "falso" unless $x =~ m{b}g
falso
```

33.3.2. Ejercicio: Opciones g y c en Expresiones Regulares

Explique cada una de las conductas que siguen

- ¿Porqué en la línea 18 se casa con la primera b?

```
  DB<5> $x = "bbabab"
  DB<6> $x =~ m{a}g; print "match= ".$&." pos = ".pos($x)
match= a pos = 3
  DB<7> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 4
  DB<8> $x =~ m{c}g; print "match= ".$&." pos = ".pos($x)
Use of uninitialized value in concatenation (.)
  DB<18> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 1
```

- ¿Porqué en la línea 27 se casa con la última b?

```
  DB<23> $x = "bbabab"
  DB<24> $x =~ m{a}g; print "match= ".$&." pos = ".pos($x)
match= a pos = 3
  DB<25> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 4
  DB<26> $x =~ m{c}gc
  DB<27> $x =~ m{b}g; print "match= ".$&." pos = ".pos($x)
match= b pos = 6
```

- ¿Porqué en la línea 5 se produce casamiento y en la línea 8 no?

```
  DB<3> $x = "bcbabab"
  DB<4> $x =~ m{b}gc; print "match= ".$&." pos = ".pos($x)
match= b pos = 1
  DB<5> $x =~ m{a}gc; print "match= ".$&." pos = ".pos($x)
match= a pos = 4

  DB<6> $x = "bcbabab"
  DB<7> $x =~ m{b}gc; print "match= ".$&." pos = ".pos($x)
match= b pos = 1
  DB<8> $x =~ m{\Ga}gc; print "match= ".$&." pos = ".pos($x)
Use of uninitialized value in concatenation
match=  pos = 1
```

33.3.3. Ejercicio: El orden de las expresiones regulares

¿Que ocurriría en la subrutina `scanner` si el código en las líneas 17-19 que reconoce los identificadores se adelanta a la línea 8? ¿Que ocurriría con el reconocimiento de las palabras reservadas como `INT`? ¿Seguiría funcionando correctamente el analizador?

33.3.4. Ejercicio: Regexp para cadenas

En la rutina `scanner`. ¿Es legal que una cadena correspondiente al terminal `STR` contenga retornos de carro entre las comillas dobles?

33.3.5. Ejercicio: El `or` es vago

Explique el resultado de la siguiente sesión con el depurador:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL/Lexical$ perl -de 0
  DB<1> 'bb' =~ m{b|bb}; print $&
  b
  DB<2> 'bb' =~ m{bb|b}; print $&
  bb
```

Perl convierte la expresión regular en un NFA. A diferencia de lo que ocurre en otras herramientas, el NFA no es convertido en un DFA. El NFA es entonces simulado. ¿Que está ocurriendo en la simulación?

33.3.6. Práctica: Números de Línea, Errores, Cadenas y Comentarios

Extienda el analizador léxico para que:

- Reescriba la expresión regular para las cadenas de manera que acepte comillas dobles escapadas `\"` en el interior de una cadena. Por ejemplo en `"esta \"palabra\" va entre comillas"`.

  Analice esta solución ¿Es correcta?:

  ```
  DB<1> $stringre = qr{"(\\.|[^\\"])*"}
  DB<2> print $& if '"esta \"palabra\" va entre comillas"' =~ $stringre
  "esta \"palabra\" va entre comillas"
  ```

- Consuma comentarios a la Perl: cualesquiera caracteres después de una almohadilla hasta el final de la línea (`# ...`).

- Consuma comentarios no anidados a la *C* (`/* ... */`.). Repase las secciones sobre expresiones regulares no "greedy" (p. ej. sección 31.1.1) y la sección 31.1.1. Recuerde que, en una expresión regular, la opción `/s` hace que el punto `'.'` empareje con un retorno de carro `\n`. Esto es, el punto "casa" con cualquier carácter.

  Observe el siguiente ejemplo:

```
pl@nereida:~/src/perl/testing$ cat -n ccomments.pl
 1  #!/usr/bin/perl -w
 2  use strict;
 3
 4  sub showmatches {
 5   my ($x, $re) = @_;
 6
 7    for (my $i = 0; $x =~ /$re/gsx; $i++) {
 8      print "matching $i: $1\n";
 9      $i++;
10    }
11  }
12
13  my $x = <<'EOC';
14  if (x) {
15    /* a comment */ return x + 1; /* another comment */
16  }
17  else {
18    return x + 2; /* a last comment */
19  }
20  EOC
21
22  print "\n************************\n";
23
24  my $greedy = q{ (
25           /\*  # Abrir comentario
26           .*   # Consumir caracteres (greedy)
27           \*/  # Cerrar comentario
28           )
29  };
30  print "Greedy:\n";
31  showmatches($x, $greedy);
32
33  print "\n************************\n";
34
35  my $lazy = q{ (
36           /\*  # Abrir comentario
37           .*?  # Consumir caracteres (lazy)
38           \*/  # Cerrar comentario
39           )
40  };
41  print "Lazy:\n";
42  showmatches($x, $lazy);
```

Cuando se ejecuta produce:

```
pl@nereida:~/src/perl/testing$ ccomments.pl

************************
Greedy:
matching 0: /* a comment */ return x + 1; /* another comment */
}
else {
  return x + 2; /* a last comment */
```

```
**********************
Lazy:
matching 0: /* a comment */
matching 2: /* another comment */
matching 4: /* a last comment */
```

Explique la conducta.

- Números en punto flotante (como `-1.32e-04` o `.91`). El siguiente ejemplo intenta ayudarle en la búsqueda de la solución:

```
lhp@nereida:~$ perl -de 0
  DB<1> print "Si" if  ('.5' =~ m{\d+})
Si
  DB<2> print "Si" if  ('.5' =~ m{^\d+})

  DB<3> print "Si" if  '0.7' =~ m{^\d+(\.\d+)?(e[+-]?\d+)?$}
Si
  DB<4> print "Si" if  '.7' =~ m{^\d+(\.\d+)?(e[+-]?\d+)?$}

  DB<5> print "Si" if  '1e2' =~ m{^\d+(\.\d+)?(e[+-]?\d+)?$}
Si
  DB<6> print "Si " while  'ababa' =~ m{a}g
Si Si Si
  DB<7> print "@a" if @a = 'ababa' =~ m{(a)}g
a a a
```

¿Sabría decir porque la respuesta al primer comando del depurador es `si`?.

- Tenga presente el posible conflicto entre los terminales `INT` y `FLOAT`. Si la entrada contiene `3.5` el terminal debería ser `(FLOAT, '3.5')` y no `(INT, 3), ('.', '.'), (INT, 5)`.

- En esta práctica si lo desea puede instalar y usar el módulo `Regexp::Common` mantenido por *Abigail* el cual provee expresiones regulares para las situaciones mas comunes: números, teléfonos, IP, códigos postales, listas, etc. Puede incluso usarlo para encontrar soluciones a las cuestiones planteadas en esta práctica:

```
nereida:~/doc/casiano/PLBOOK/PLBOOK> perl -MRegexp::Common -e 'print "$RE{num}{int}\n"'
(?:(?:[+-]?)(?:[0123456789]+))
```

Podemos hacer uso directo del hash `%RE` directamente en las expresiones regulares aprovechando que estas interpolan las variables en su interior:

```
nereida:/tmp> cat -n prueba.pl
     1  #!/usr/bin/perl -w
     2  use strict;
     3  use Regexp::Common;
     4
     5  my $input = <>;
     6
     7  print "$&\n" if $input =~ /^$RE{num}{real}$/;
nereida:/tmp> ./prueba.pl
23.45
```

```
23.45
nereida:/tmp> ./prueba.pl
jshdf
nereida:/tmp>
```

- Para mejorar la calidad de los mensajes de error extienda el par `(terminal, valor)` devuelto por el `scanner` a un par `(terminal, [valor, número de línea ])` cuya segunda componente es un array anónimo conteniendo el valor y el número de línea en el que aparece el terminal.

  El siguiente extracto de un analizador léxico muestra como hacerlo:

```
sub _Lexer {

  return('', undef) unless defined($input);

  #Skip blanks
      $input=~m{\G((?:
              \s+       # any white space char
          |   \#[^\n]*  # Perl like comments
          )+)}xsgc
    and do {
        my($blanks)=$1;

        #Maybe At EOF
            pos($input) >= length($input)
        and return('', undef);
        $tokenend += $blanks =~ tr/\n//;
    };

    $tokenbegin = $tokenend;

      $input=~/\G(and)/gc
    and return($1, [$1, $tokenbegin]);

      $input=~/\G(?:[A-Za-z_][A-Za-z0-9_]*::)*([A-Za-z_][A-Za-z0-9_]*)/gc
    and do {
      return('IDENT', [$1, $tokenbegin]);
    };

    ..................

        $input=~/\G{/gc
    and do {
        my($level,$from,$code);

        $from=pos($input);

        $level=1;
        while($input=~/([{}])/gc) {
                substr($input,pos($input)-1,1) eq '\\' #Quoted
            and next;
                $level += ($1 eq '{' ? 1 : -1)
            or last;
        }
```

```
            $level
        and  _SyntaxError("Not closed open curly bracket { at $tokenbegin");
        $code = substr($input,$from,pos($input)-$from-1);
        $tokenend+= $code=~tr/\n//;
        return('CODE', [$code, $tokenbegin]);
    };


    #Always return something
      $input=~/\G(.)/sg
    and do {
      $1 eq "\n" and ++$tokenend;
      return ($1, [$1, $tokenbegin]);
    };
    #At EOF
    return('', undef);
}
```

El operador `tr` ha sido utilizado para contar los retornos de carro (descrito en la sección 31.5). El operador, ademas de reemplazar devuelve el número de carácteres reeemplazados o suprimidos:

```
$cuenta = $cielo =~ tr/*/*/; # cuenta el numero de estrellas en cielo
```

Para aprender soluciones alternativas consulte `perldoc -q 'substring'`.

- Mejore sus mensajes de error ahora que lleva la cuenta de los números de línea. En vez de usar las rutinas `error` y `fatal` introducidas en la sección anterior escriba una sola rutina que recibe el nivel de severidad del error (parámetro `$level` en el siguiente ejemplo) y ejecuta la acción apropiada. El código de `_SyntaxError` ilustra como hacerlo:

```
sub _SyntaxError {
  my($level,$message,$lineno)=@_;

  $message= "*".
      [ 'Warning', 'Error', 'Fatal' ]->[$level].
      "* $message, at ".
      ($lineno < 0 ? "eof" : "line $lineno")." at file $filename\n";

      $level > 1
  and die $message;

  warn $message;

  $level > 0 and ++$nberr;

      $nberr == $max_errors
  and die "*Fatal* Too many errors detected.\n"
}
```

33.4. Pruebas para el Analizador Léxico

Queremos separar/aislar las diferentes fases del compilador en diferentes módulos.

Módulo `PL::Error`

Para ello comenzamos creando un módulo conteniendo las rutinas de tratamiento de errores:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL$ pwd
/home/lhp/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL$ cat -n Error.pm
   1  package Error;
   2  use strict;
   3  use warnings;
   4  use Carp;
   5
   6  require Exporter;
   7
   8  our @ISA = qw(Exporter);
   9  our @EXPORT = qw( error fatal);
  10  our $VERSION = '0.01';
  11
  12  sub error {
  13    my $msg = join " ", @_;
  14    if (!$PL::Tutu::errorflag) {
  15      carp("Error: $msg\n");
  16      $PL::Tutu::errorflag = 1;
  17    }
  18  }
  19
  20  sub fatal {
  21    my $msg = join " ", @_;
  22    croak("Error: $msg\n");
  23  }
```

Observa como accedemos a la variable `errorflag` del paquete `PL::Tutu`. Para usar este módulo desde `PL::Tutu`, tenemos que declarar su uso:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL$ cat -n Tutu.pm | head -8
     1  package PL::Tutu;
     2
     3  use 5.008004;
     4  use strict;
     5  use warnings;
     6  use IO::File;
     7  use Carp;
     8  use PL::Error;
```

En la línea 8 hacemos `use PL::Error` y no `use Error` ya que el módulo lo hemos puesto en el directorio `PL`. No olvides hacer `make manifest` para actualizar el fichero `MANIFEST`.

Módulo `PL::Lexical::Analysis`

Supongamos que además de modularizar el grupo de rutinas de tratamiento de errores queremos hacer lo mismo con la parte del análisis léxico. Parece lógico que el fichero lo pongamos en un subdirectorio de `PL/` por lo que cambiamos el nombre del módulo a `PL::Lexical::Analysis` quedando la jerarquía de ficheros asi:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL$ tree
.
|-- Error.pm
```

```
|-- Lexical
|   `-- Analysis.pm
`-- Tutu.pm
```

Por supuesto debemos modificar las correspondientes líneas en `Tutu.pm`:

```
 1 package PL::Tutu;
 2
 3 use 5.008004;
 4 use strict;
 5 use warnings;
 6 use IO::File;
 7 use Carp;
 8 use PL::Error;
 9 use PL::Lexical::Analysis;
10 ...
11
12 sub compile {
13   my ($input) = @_;
14   local %symbol_table = ();
15   local $data = ""; # Contiene todas las cadenas en el programa fuente
16   local $target = ""; # target code
17   my @tokens = ();
18   local $errorflag = 0;
19   local ($lookahead, $value) = ();
20   local $tree = undef; # abstract syntax tree
21   local $global_address = 0;
22
23
24   ########lexical analysis
25   @tokens = &PL::Lexical::Analysis::scanner($input);
26   print "@tokens\n";
27
28   ...
29
30   return \$target;
31 }
```

Observe que ahora `PL::Lexical::Analysis::scanner` devuelve ahora la lista con los terminales y que `@tokens` se ha ocultado en `compile` como una variable léxica (línea 17). En la línea 26 mostramos el contenido de la lista de terminales.

Sigue el listado del módulo conteniendo el analizador léxico. Obsérve las líneas 6, 16 y 44.

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/lib/PL/Lexical$ cat -n Analysis.pm
     1  # Lexical analyzer
     2  package PL::Lexical::Analysis;
     3  use strict;
     4  use warnings;
     5  use Carp;
     6  use PL::Error;
     7
     8  require Exporter;
     9
    10  our @ISA = qw(Exporter);
    11  our @EXPORT = qw( scanner );
```

```
12  our $VERSION = '0.01';
13
14  sub scanner {
15    local $_ = shift;
16    my @tokens;
17
18    { # Con el redo del final hacemos un bucle "infinito"
19      if (m{\G\s*(\d+)}gc) {
20        push @tokens, 'NUM', $1;
21      }
..      ...
37      elsif (m{\G\s*([+*()=;,])}gc) {
38        push @tokens, 'PUN', $1;
39      }
40      elsif (m{\G\s*(\S)}gc) {
41        Error::fatal "Caracter invalido: $1\n";
42      }
43      else {
44        return @tokens;
45      }
46      redo;
47    }
48  }
```

El Programa Cliente

Puesto que en el paquete `PL::Lexical::Analysis` exportamos `scanner` no es necesario llamar la rutina por el nombre completo desde `compile`. Podemos simplificar la línea en la que se llama a `scanner` que queda así:

```
########lexical analysis
@tokens = &scanner($input);
print "@tokens\n";
```

De la misma forma, dado que `PL::Tutu` exporta la función `compile_from_file`, no es necesario llamarla por su nombre completo desde el guión `tutu`. Reescribimos la línea de llamada:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/scripts$ cat tutu
#!/usr/bin/perl -w
use lib ('../lib');
use PL::Tutu;

&compile_from_file(@ARGV);
```

Actualización del MANIFEST

Como siempre que se añaden o suprimen archivos es necesario actualizar `MANIFEST`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu$ make manifest
/usr/bin/perl "-MExtUtils::Manifest=mkmanifest" -e mkmanifest
Added to MANIFEST: lib/PL/Lexical/Analysis.pm
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu$ cat -n MANIFEST
     1  Changes
     2  lib/PL/Error.pm
     3  lib/PL/Lexical/Analysis.pm
     4  lib/PL/Tutu.pm
     5  Makefile.PL
```

```
 6  MANIFEST
 7  MANIFEST.SKIP
 8  README
 9  scripts/test01.tutu
10  scripts/tutu
11  scripts/tutu.pl
12  t/PL-Tutu.t
```

33.4.1. Comprobando el Analizador Léxico

Queremos comprobar si nuestro código funciona. ¿Cómo hacerlo?. Lo adecuado es llevar una aproximación sistemática que permita validar el código.

Principios Básicos del Desarrollo de Pruebas

En general, la filosofía aconsejable para realizar un banco de pruebas de nuestro módulo es la que se articula en la metodología denominada *Extreme Programming*, descrita en múltiples textos, en concreto en el libro de Scott [**?**]:

- Todas las pruebas deben automatizarse
- Todos los fallos que se detecten deberían quedar traducidos en pruebas
- La aplicación debería pasar todas las pruebas después de cualquier modificación importante y también al final del día
- El desarrollo de las pruebas debe preceder el desarrollo del código
- Todos los requerimientos deben ser expresados en forma de pruebas

La Jerarquía de Una Aplicación

Pueden haber algunas diferencias entre el esquema que se describe aqui y su versión de Perl. Lea detenidamente el capítulo Test Now, test Forever del libro de Scott [**?**] y el libro [11] de Ian Langworth y chromatic.

Si usas una versión de Perl posterior la 5.8.0, entonces tu versión del programa `h2xs` creará un subdirectorio `/t` en el que guardar los ficheros de prueba Estos ficheros deberán ser programas Perl de prueba con el tipo `.t`. La utilidad `h2xs` incluso deja un programa de prueba `PL-Tutu.t` en ese directorio. La jerarquía de ficheros con la que trabajamos actualmente es:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu$ make veryclean
rm -f blib/script/tutu blib/script/tutu.pl
rm -rf ./blib Makefile.aperl ...
mv Makefile Makefile.old > /dev/null 2>&1
rm -rf blib/lib/auto/PL/Tutu blib/arch/auto/PL/Tutu
rm -rf PL-Tutu-0.01
rm -f  blib/lib/PL/.Tutu.pm.swp ...
rm -f *~ *.orig */*~ */*.orig
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu$ tree
.
|-- .svn                # use siempre un sistema de control de versiones
|-- Changes             # la historia de cambios
|-- MANIFEST            # lista de ficheros que componen la distribución
|-- MANIFEST.SKIP       # regexps para determinar que ficheros no pertenecen
|-- META.yml            # YML no es XML
|-- Makefile.PL         # generador del Makefle independiente de la plataforma
|-- PL-Tutu-0.01.tar.gz
|-- README              # instrucciones de instalacion
```

```
|-- lib
|   `-- PL
|       |-- Error.pm            # rutinas de manejo de errores
|       |-- Lexical
|       |   `-- Analysis.pm     # modulo con el analizador lexico
|       `-- Tutu.pm            # modulo principal
|-- scripts
|   |-- test01.sal   # salida del programa de prueba
|   |-- test01.tutu  # programa de prueba
|   |-- tutu         # compilador
|   `-- tutu.pl      # compilador
`-- t
    `-- 01Lexical.t  # prueba consolidada
```

Un Ejemplo de Programa de Prueba

Estos son los contenidos de nuestro primer test:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu$ cd t
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/t$ ls -l
total 4
-rw-r--r--  1 lhp lhp 767 2005-10-10 11:27 01Lexical.t
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/t$ cat -n 01Lexical.t
 1  # Before `make install' is performed this script should be runnable with
 2  # `make test'. After `make install' it should work as `perl PL-Tutu.t'
 3
 4  #########################
 5
 6  # change 'tests => 1' to 'tests => last_test_to_print';
 7
 8  use Test::More tests => 5;
 9  use Test::Exception;
10
11  BEGIN { use_ok('PL::Lexical::Analysis') };
12  BEGIN { use_ok('PL::Tutu') };
13
14  #########################
15
16  # Insert your test code below, the Test::More module is use()ed here so read
17  # its man page ( perldoc Test::More ) for help writing this test script.
18
19  can_ok('PL::Lexical::Analysis', 'scanner');
20
21  # Test result of call
22  my $a = 'int a,b; string c; c = "hello"; a = 4; b = a +1; p b';
23  my @tokens = scanner($a);
24  my @expected_tokens = qw{
25  INT INT
26  ID a
27  PUN ,
28  ID b
29  PUN ;
30  STRING STRING
31  ID c
32  PUN ;
```

```
33  ID c
34  PUN =
35  STR "hello"
36  PUN ;
37  ID a
38  PUN =
39  NUM 4
40  PUN ;
41  ID b
42  PUN =
43  ID a
44  PUN +
45  NUM 1
46  PUN ;
47  P P
48  ID b
49  };
50  is(@tokens, @expected_tokens, "lexical analysis");
51
52  # test a lexically  erroneous program
53  $a = 'int a,b; string c[2]; c = "hello"; a = 4; b = a +1; p b';
54  throws_ok { scanner($a) } qr{Error: Caracter invalido:}, 'erroneous program';
```

El nombre del fichero de prueba debe cumplir que:

- Sea significativo del tipo de prueba
- Que los prefijos de los nombres `01`, `02`, ... nos garanticen el orden de ejecución

Ejecución de Las Pruebas

Ahora ejecutamos las pruebas:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu$ make test
PERL_DL_NONLAZY=1 /usr/bin/perl "-MExtUtils::Command::MM" "-e" "test_harness(0, 'blib/lib', 'b
t/01Lexical....ok 1/5Possible attempt to separate words with commas at t/01Lexical.t line 49.
t/01Lexical....ok
All tests successful.
Files=1, Tests=5,  0 wallclock secs ( 0.08 cusr +  0.00 csys =  0.08 CPU)
```

O bien usamos `prove`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/t$ prove -I../lib 01Lexical.t
01Lexical....ok
All tests successful.
Files=1, Tests=2,  0 wallclock secs ( 0.03 cusr +  0.01 csys =  0.04 CPU)
```

También podemos añadir la opción `verbose` a `prove`:

```
lhp@nereida:~/Lperl/src/topdown/PL0506/03lexico/PL-Tutu/t$ prove -v 01Lexical.t
01Lexical....1..5
ok 1 - use PL::Lexical::Analysis;
ok 2 - use PL::Tutu;
ok 3 - PL::Lexical::Analysis->can('scanner')
ok 4 - lexical analysis
ok 5 - erroneous program
ok
All tests successful.
Files=1, Tests=5,  0 wallclock secs ( 0.07 cusr +  0.01 csys =  0.08 CPU)
```

Repáse [10] para un mejor conocimiento de la metodología de pruebas en Perl.

Versiones anteriores a la 5.8

En esta sección he usado la versión 5.6.1 de Perl. Creeemos un subdirectorio `tutu_src/` y en él un programa de prueba `pruebalex.pl`:

```
$ pwd
/home/lhp/projects/perl/src/tmp/PL/Tutu/tutu_src
$ cat pruebalex.pl
#!/usr/bin/perl -w -I..
#use PL::Tutu;
use Tutu;

my $a = 'int a,b; string c; c = "hello"; a = 4; b = a +1; p b';
Lexical::Analysis::scanner($a);
print "prog = $a\ntokens = @PL::Tutu::tokens\n";
```

Observa como la opción `-I..` hace que se busque por las librerías en el directorio padre del actual. Cuando ejecutamos `pruebalex.pl` obtenemos la lista de terminales:

```
$ ./pruebalex.pl
prog = int a,b; string c; c = "hello"; a = 4; b = a +1; p b
tokens = INT INT ID a PUN , ID b PUN ; STRING STRING ID c PUN ;
ID c PUN = STR hello PUN ; ID a PUN = NUM 4 PUN ; ID b PUN =
ID a PUN + NUM 1 PUN ; P P ID b
```

La última línea ha sido partida por razones de legibilidad, pero consituye una sóla línea. Editemos el fichero `test.pl` en el directorio del módulo. Sus contenidos son como sigue:

```
$ cat -n test.pl
 1  # Before ‘make install’ is performed this script should be runnable with
 2  # ‘make test’. After ‘make install’ it should work as ‘perl test.pl’
 3
 4  ##########################
 5
 6  # change 'tests => 1' to 'tests => last_test_to_print';
 7
 8  use Test;
 9  BEGIN { plan tests => 1 };
10  use PL::Tutu;
11  ok(1); # If we made it this far, we're ok.
12
13  ##########################
14
15  # Insert your test code below, the Test module is use()ed here so read
16  # its man page ( perldoc Test ) for help writing this test script.
17
```

En la línea 9 se establece el número de pruebas a realizar. La primera prueba aparece en la línea 11. Puede parecer que no es una prueba, ¡pero lo es!. Si se ha alcanzado la línea 11 es que se pudo cargar el módulo `PL::Tutu` y eso ¡tiene algún mérito!.

Seguiremos el consejo de la línea 15 y escribiremos nuestra segunda prueba al final del fichero `test.pl`:

```
$ cat -n test.pl | tail -7
```

```
16  # its man page ( perldoc Test ) for help writing this test script.
17
18  my $a = 'int a,b; string c; c = "hello"; a = 4; b = a +1; p b';
19  local @PL::Tutu::tokens = ();
20  Lexical::Analysis::scanner($a);
21  ok("@PL::Tutu::tokens" eq
22  'INT INT ID a PUN , ID b PUN ; STRING STRING ID c PUN ; ID c
     PUN = STR hello PUN ; ID a PUN = NUM 4 PUN ; ID b PUN =
     ID a PUN + NUM 1 PUN ; P P ID b');
```

La línea 22 ha sido partida por razones de legibilidad, pero constituye una sóla línea. Ahora podemos ejecutar `make test` y comprobar que las dos pruebas funcionan:

```
$ make test
PERL_DL_NONLAZY=1 /usr/bin/perl -Iblib/arch -Iblib/lib -I/usr/lib/perl/5.6.1 \
                                -I/usr/share/perl/5.6.1 test.pl
1..2
ok 1
ok 2
```

¿Recordaste cambiar la línea 9 de `test.pl`? ¿Añadiste los nuevos ficheros a la lista en `MANIFEST`?

33.4.2. Práctica: Pruebas en el Análisis Léxico

Extienda su compilador para modularizar el analizador léxico tal y como se explicó en la sección 33.4.

1. Lea los siguientes documentos
   - `Test::Tutorial` (Test::Tutorial - A tutorial about writing really basic tests) por Michael Schwern.
   - Test Now, test Forever " del libro de Scott [**?**].
   - Perl Testing Reference Card por Ian Langworth.
   - Chapter 4: Distributing Your Tests (and Code) del libro Perl Testing: A Developer's Notebook
2. Incluya la estrategia de pruebas de no regresión explicada en las secciones previas. Dado que ahora la estructura del terminal es una estructura de datos mas compleja `(token, [value, line_number])` no podrá usar `is`, ya que este último sólo comprueba la igualdad entre escalares. Use `is_deeply` para comprobar que la estructura de datos devuelta por el analizador léxico es igual a la esperada. Sigue un ejemplo:

   ```
   nereida:~/src/perl/YappWithDefaultAction/t> cat -n  15treeregswith2arrays.t
    1  #!/usr/bin/perl -w
    2  use strict;
    3  #use Test::More qw(no_plan);
    4  use Test::More tests => 3;
    5  use_ok qw(Parse::Eyapp) or exit;

   ..  ..... etc., etc.

   84  my $expected_tree = bless( {
   85    'children' => [
   86      bless( { 'children' => [
   87          bless( { 'children' => [], 'attr' => 'a', 'token' => 'a' }, 'TERMINAL' )
   ```

```
88        ]
89      }, 'A' ),
90      bless( { 'children' => [
91          bless( { 'children' => [], 'attr' => 'c', 'token' => 'c' }, 'TERMINAL' )
92        ]
93      }, 'C' )
94    ]
95  }, 'ABC' );
96  is_deeply($t, $expected_tree, "deleting node between arrays");
```

3. Extienda los tests con una prueba en la que la entrada contenga un carácter ilegal. Obsérve que, tal y como esta escrito la rutina `scanner`, si la entrada tiene un carácter ilegal se ejecutarán las líneas

   ```
   26     elsif (/\G\s*(.)/gc) {
   27       Error::fatal "Caracter invalido: $1\n";
   28     }
   ```

   lo que causa la parada del programa de prueba, al ejecutarse `fatal` el cuál llama a `croak`.

   ```
   sub fatal {
     my $msg = join " ", @_;
     croak("Error: $msg\n");
   }
   ```

   El objetivo es lograr que el programa de pruebas continúe ejecutando las subsiguientes pruebas.

   Para ello puede usar `Test::Exception` o bien `eval` y la variable especial `$@` para controlar que el programa `.t` no termine prematuramente. Repase las secciones sobre pruebas en [10].

4. Pruebe a dar como entrada un fichero vacío

5. Pruebe a dar como entrada un fichero que no existe

6. Pruebe a dar como entrada un fichero binario

7. Si tiene sentido en su caso, llame a las subrutinas con mas argumentos (y también con menos) de los que esperan.

8. Si tiene sentido en su caso, llame a las subrutinas con argumentos cuyo tipo no es el que se espera.

   No use prototipos para lograrlo. No es una buena idea. Los prototipos en Perl a menudo producen un preprocesado del parámetro. Escriba código que controle que la naturaleza del parámetro es la que se espera. Por ejemplo:

   ```
   sub tutu {
     my $refhash = shift;
     croak "Error" unless UNIVERSAL::isa($refhash, 'HASH');
     ...
   }
   ```

9. Cambie los argumentos de orden (si es que se aplica a su código)

10. Comentarios: pruebe con `/* * */ a = 4; /* * / */`. Tambien con comentarios anidados (debería producirse un error)

11. Flotantes: compruebe su expresión regular con 0.0 0e0 .0 0 1.e-5 1.0e2 -2.0 . (un punto sólo)

12. Cadenas. Pruebe con las cadenas

```
""
"h\"a\"h"
"\""
```

    Pruebe también con una cadena con varias líneas y otra que contenga un carácter de control en su interior.

13. Convierta los fallos (bugs) que encontró durante el desarrollo en pruebas

14. Compruebe la documentación usando el módulo Test::Pod de Andy Lester. Instálelo si es necesario.

15. Utilice el módulo Test::Warn para comprobar que los mensajes de warning (uso de `warn` and `carp`) se muestran correctamente.

16. Una prueba `SKIP` declara un bloque de pruebas que - bajo ciertas circustancias - puede saltarse. Puede ser que sepamos que ciertas pruebas sólo funcionan en ciertos sistemas operativos o que la prueba requiera que ciertos paquetes están instalados o que la máquina disponga de ciertos recursos (por ejemplo, acceso a internet). En tal caso queremos que los tests se consideren si se dan las circustancias favorables pero que en otro caso se descarten sin protestas. Consulte la documentación de los módulos `Test::More` y `Test::Harness` sobre pruebas tipo `SKIP`. El ejemplo que sigue declara un bloque de pruebas que pueden saltarse. La llamada a `skip` indica cuantos tests hay, bajo que condición saltarselos.

```
 1  SKIP: {
 2      eval { require HTML::Lint };
 3
 4      skip "HTML::Lint not installed", 2 if $@;
 5
 6      my $lint = new HTML::Lint;
 7      isa_ok( $lint, "HTML::Lint" );
 8
 9      $lint->parse( $html );
10      is( $lint->errors, 0, "No errors found in HTML" );
11  }
```

    Si el usuario no dispone del módulo `HTML::Lint` el bloque no será ejecutado. El módulo `Test::More` producirá `oks` que serán interpretados por `Test::Harness` como tests *skipped* pero `ok`.

    Otra razón para usar una prueba `SKIP` es disponer de la posibilidad de saltarse ciertos grupos de pruebas. Por ejemplo, aquellas que llevan demasiado tiempo de ejecución y no son tan significativas que no se pueda prescindir de ellas cuando se introducen pequeños cambios en el código. El siguiente código muestra como usando una variable de entorno `TEST_FAST` podemos controlar que pruebas se ejecutan.

```
nereida:~/src/perl/YappWithDefaultAction/t> cat 02Cparser.t | head -n 56 -
#!/usr/bin/perl -w
use strict;
#use Test::More qw(no_plan);
use Test::More tests => 6;

use_ok qw(Parse::Eyapp) or exit;
```

```
SKIP: {
  skip "You decided to skip C grammar test (env var TEST_FAST)", 5 if $ENV{TEST_FAST} ;
  my ($grammar, $parser);
  $grammar=join('',<DATA>);
  $parser=new Parse::Eyapp(input => $grammar, inputfile => 'DATA', firstline => 52);

  #is($@, undef, "Grammar module created");

  # Does not work. May I have done s.t. wrong?
  #is(keys(%{$parser->{GRAMMAR}{NULLABLE}}), 43, "43 nullable productions");

  is(keys(%{$parser->{GRAMMAR}{NTERM}}), 233, "233 syntactic variables");

  is(scalar(@{$parser->{GRAMMAR}{UUTERM}}), 3, "3 UUTERM");

  is(scalar(keys(%{$parser->{GRAMMAR}{TERM}})), 108, "108 terminals");

  is(scalar(@{$parser->{GRAMMAR}{RULES}}), 825, "825 rules");

  is(scalar(@{$parser->{STATES}}), 1611, "1611 states");
}

__DATA__
/*
   This grammar is a stripped form of the original C++ grammar
   from the GNU CC compiler :

   YACC parser for C++ syntax.
   Copyright (C) 1988, 89, 93-98, 1999 Free Software Foundation, Inc.
   Hacked by Michael Tiemann (tiemann@cygnus.com)

   The full gcc compiler an the original grammar file are freely
   available under the GPL license at :

   ftp://ftp.gnu.org/gnu/gcc/
   ..................... etc. etc.
*/
nereida:~/src/perl/YappWithDefaultAction> echo $TEST_FAST
1
nereida:~/src/perl/YappWithDefaultAction> make test
PERL_DL_NONLAZY=1 /usr/bin/perl "-MExtUtils::Command::MM" "-e" "test_harness(0, 'blib/lib
t/01calc.....................................ok
t/02Cparser..................................ok
        5/6 skipped: various reasons
t/03newgrammar...............................ok
t/04foldandzero..............................ok
t/05treewithvars.............................ok
t/06meta.....................................ok
t/07translationschemetype....................ok
t/08tschemetypestar..........................ok
t/09ts_with_defaultaction....................ok
t/10ts_with_treereg..........................ok
```

```
etc., etc.................................ok

t/28unshifttwoitems.......................ok
t/29foldinglistsofexpressions.............ok
t/30complextreereg........................ok
t/32deletenodewithwarn....................ok
t/33moveinvariantoutofloop................ok
t/34moveinvariantoutofloopcomplexformula....ok
All tests successful, 5 subtests skipped.
Files=33, Tests=113,  5 wallclock secs ( 4.52 cusr +  0.30 csys =  4.82 CPU)
```

Introduzca una prueba `SKIP` similar a la anterior y otra que si el módulo `Test::Pod` esta instalado comprueba que la documentación esta bien escrita. Estudie la documentación del módulo `Test::Pod`.

17. Introduzca pruebas `TODO` (que, por tanto, deben fallar) para las funciones que están por escribir (`parser`, `Optimize`, `code_generator`, `transform`). Repáse [10]. Sigue un ejemplo:

```
42 TODO: {
43   local $TODO = "Randomly generated problem";
44   can_ok('Algorithm::Knap01DP', 'GenKnap'); # sub GenKnap no ha sido escrita aún
45 }
```

18. Cuando compruebe el funcionamiento de su módulo *nunca descarte que el error pueda estar en el código de la prueba*. En palabras de Schwern

> Code has bugs. Tests are code. Ergo, tests have bugs.
>
> Michael Schwern

19. Instale el módulo Devel::Cover. El módulo `Devel::Cover` ha sido escrito por Paul Johnson y proporciona estadísticas del cubrimiento alcanzado por una ejecución. Para usarlo siga estos pasos:

```
pl@nereida:~/src/perl/YappWithDefaultAction$ cover -delete
Deleting database /home/pl/src/perl/YappWithDefaultAction/cover_db
pl@nereida:~/src/perl/YappWithDefaultAction$ HARNESS_PERL_SWITCHES=-MDevel::Cover make te
PERL_DL_NONLAZY=1 /usr/bin/perl "-MExtUtils::Command::MM" "-e" "test_harness(0, 'blib/lib
t/01calc...................................ok
t/01calc...................................ok
t/02Cparser................................ok
        5/6 skipped: various reasons
t/03newgrammar.............................ok
t/03newgrammar.............................ok
t/04foldandzero............................ok
etc., etc. ................................ok
t/34moveinvariantoutofloopcomplexformula....ok
All tests successful, 5 subtests skipped.
Files=33, Tests=113, 181 wallclock secs (177.95 cusr +  2.94 csys = 180.89 CPU)
```

La ejecución toma ahora mucho mas tiempo: ¡181 segundos frente a los 5 que toma la ejecución sin `cover` !. Al ejecutar `cover` de nuevo obtenemos una tabla con las estadísticas de cubrimiento:

```
pl@nereida:~/src/perl/YappWithDefaultAction$ cover
Reading database from /home/pl/src/perl/YappWithDefaultAction/cover_db

---------------------------- ------ ------ ------ ------ ------ ------ ------
File                          stmt   bran   cond    sub    pod   time  total
---------------------------- ------ ------ ------ ------ ------ ------ ------
blib/lib/Parse/Eyapp.pm      100.0    n/a    n/a  100.0    n/a    0.2  100.0
...lib/Parse/Eyapp/Driver.pm  72.4   63.2   50.0   64.3    0.0   21.3   64.4
...ib/Parse/Eyapp/Grammar.pm  90.9   77.8   66.7  100.0    0.0   16.6   84.3
blib/lib/Parse/Eyapp/Lalr.pm  91.4   72.6   78.6  100.0    0.0   48.3   85.6
blib/lib/Parse/Eyapp/Node.pm  74.4   58.3   29.2   88.2    0.0    1.6   64.7
...ib/Parse/Eyapp/Options.pm  86.4   50.0    n/a  100.0    0.0    2.7   72.8
...lib/Parse/Eyapp/Output.pm  82.3   47.4   60.0   70.6    0.0    3.7   70.0
.../lib/Parse/Eyapp/Parse.pm 100.0    n/a    n/a  100.0    n/a    0.2  100.0
...Parse/Eyapp/Treeregexp.pm 100.0    n/a    n/a  100.0    n/a    0.1  100.0
blib/lib/Parse/Eyapp/YATW.pm  89.4   63.9   66.7   85.7    0.0    4.8   77.6
...app/_TreeregexpSupport.pm  73.1   33.3   50.0  100.0    0.0    0.4   60.8
main.pm                       52.2    0.0    n/a   80.0    0.0    0.0   45.7
Total                         83.8   64.7   60.0   84.5    0.0  100.0   75.5
---------------------------- ------ ------ ------ ------ ------ ------ ------

Writing HTML output to /home/pl/src/perl/YappWithDefaultAction/cover_db/coverage.html ...
pl@nereida:~/src/perl/YappWithDefaultAction$
```

El HTML generado nos permite tener una visión mas detallada de los niveles de cubrimiento.

Para mejorar el cubrimiento de tu código comienza por el informe de cubrimiento de subrutinas. Cualquier subrutina marcada como no probada es un candidato a contener errores o incluso a ser *código muerto*.

Para poder hacer el cubrimiento del código usando Devel::Cover, si se usa una `csh` o `tcsh` se debe escribir:

```
nereida:~/src/perl/YappWithDefaultAction> setenv HARNESS_PERL_SWITCHES -MDevel::Cover
nereida:~/src/perl/YappWithDefaultAction> make test
PERL_DL_NONLAZY=1 /usr/bin/perl "-MExtUtils::Command::MM" "-e" "test_harness(0, 'blib/lib
t/01calc.....................................ok
t/01calc.....................................ok
t/02Cparser..................................ok
        5/6 skipped: various reasons
t/03newgrammar...............................ok
t/03newgrammar...............................ok
t/04foldandzero..............................ok
t/05treewithvars.............................ok
t/06meta.....................................ok
t/06meta.....................................ok
t/07translationschemetype....................ok
.............................................ok
t/38tspostfix_resultisarray..................ok
t/39tspostfix................................ok
All tests successful, 5 subtests skipped.
Files=38, Tests=135, 210 wallclock secs (206.28 cusr +  3.27 csys = 209.55 CPU)
nereida:~/src/perl/YappWithDefaultAction>
```

Aún mas robusto - más independiente de la shell que usemos - es pasar las opciones en `HARNESS_PERL_SWITCHE` como parámetro a `make`:

```
make HARNESS_PERL_SWITCHES=-MDevel::Cover test
```

Añade el informe de cubrimiento al `MANIFEST` para que se incluya en la distribución que subas. Si lo consideras conveniente añade un directorio informes en los que vayan los informes asociados a esta práctica. Incluye en el `README` o en la documentación una breve descripción de donde están los informes.

20. Se conoce con el nombre de *perfilado* o *profiling* de un programa al estudio de su rendimiento mediante un programa (conocido como *profiler*) que monitoriza la ejecución del mismo mediante una técnica que interrumpe cada cierto tiempo el programa para comprobar en que punto de la ejecución se encuentra. Las estadísticas acumuladas se vuelcan al final de la ejecución en un fichero que puede ser visualizado mediante la aplicación apropiada.

    En Perl hay dos módulos que permiten realizar profiling. El mas antiguo es `Devel::DProf` . La aplicación para visualizar los resultados se llama `dprofpp` . Sigue un ejemplo de uso:

```
nereida:~/src/perl/YappWithDefaultAction/t> perl -d:DProf 02Cparser.t
1..6
ok 1 - use Parse::Eyapp;
ok 2 - 233 syntactic variables
ok 3 - 3 UUTERM
ok 4 - 108 terminals
ok 5 - 825 rules
ok 6 - 1611 states
nereida:~/src/perl/YappWithDefaultAction/t> dprofpp tmon.out
Total Elapsed Time = 3.028396 Seconds
  User+System Time = 3.008396 Seconds
Exclusive Times
%Time ExclSec CumulS #Calls sec/call Csec/c  Name
 31.4   0.945  1.473   1611   0.0006 0.0009  Parse::Eyapp::Lalr::_Transitions
 17.5   0.528  0.528   1611   0.0003 0.0003  Parse::Eyapp::Lalr::_Closures
 16.1   0.486  0.892      1   0.4861 0.8918  Parse::Eyapp::Lalr::_ComputeFollows
 8.04   0.242  0.391      1   0.2419 0.3906  Parse::Yapp::Driver::_Parse
 8.04   0.242  0.242  11111   0.0000 0.0000  Parse::Eyapp::Lalr::__ANON__
 4.59   0.138  0.138   8104   0.0000 0.0000  Parse::Eyapp::Lalr::_Preds
 2.66   0.080  0.080      1   0.0800 0.0800  Parse::Eyapp::Lalr::_SetDefaults
 2.66   0.080  0.972      1   0.0800 0.9718  Parse::Eyapp::Lalr::_ComputeLA
 2.46   0.074  0.074   3741   0.0000 0.0000  Parse::Eyapp::Parse::_Lexer
 1.89   0.057  0.074   8310   0.0000 0.0000  Parse::Eyapp::Parse::__ANON__
 0.96   0.029  0.028      1   0.0288 0.0276  Parse::Eyapp::Lalr::_SolveConflict
                                             s
 0.66   0.020  0.050      6   0.0033 0.0083  Parse::Eyapp::Output::BEGIN
 0.60   0.018  1.500      1   0.0176 1.4997  Parse::Eyapp::Lalr::_LR0
 0.53   0.016  0.259      3   0.0054 0.0863  Parse::Eyapp::Lalr::_Digraph
 0.33   0.010  0.010      1   0.0100 0.0100  Parse::Eyapp::Grammar::_SetNullable
```

    Tambien es posible usar el módulo `-MDevel::Profiler` :

```
nereida:~/src/perl/YappWithDefaultAction/examples> perl -MDevel::Profiler eyapp 02Cparser
Unused terminals:

        END_OF_LINE, declared line 128
        ALL, declared line 119
        PRE_PARSED_CLASS_DECL, declared line 120
```

```
27 shift/reduce conflicts and 22 reduce/reduce conflicts
nereida:~/src/perl/YappWithDefaultAction/examples> dprofpp tmon.out
Total Elapsed Time = 3.914144 Seconds
  User+System Time = 3.917144 Seconds
Exclusive Times
%Time ExclSec CumulS #Calls sec/call Csec/c  Name
 22.3   0.877  1.577   1611   0.0005 0.0010  Parse::Eyapp::Lalr::_Transitions
 17.8   0.700  0.700   1611   0.0004 0.0004  Parse::Eyapp::Lalr::_Closures
 15.6   0.614  1.185      1   0.6142 1.1854  Parse::Eyapp::Lalr::_ComputeFollow
                                             s
 9.60   0.376  0.545      1   0.3758 0.5453  Parse::Yapp::Driver::_Parse
 7.99   0.313  0.313   8104   0.0000 0.0000  Parse::Eyapp::Lalr::_Preds
 5.85   0.229  0.229      3   0.0763 0.0763  Parse::Eyapp::Lalr::_Digraph
 4.06   0.159  0.159   3741   0.0000 0.0000  Parse::Eyapp::Parse::_Lexer
 3.32   0.130  0.130      1   0.1300 0.1300  Parse::Eyapp::Lalr::DfaTable
 2.27   0.089  0.089      1   0.0890 0.0890  Parse::Eyapp::Lalr::_SetDefaults
 2.04   0.080  1.265      1   0.0800 1.2654  Parse::Eyapp::Lalr::_ComputeLA
 1.17   0.046  0.057      1   0.0464 0.0567  Parse::Eyapp::Grammar::Rules
 1.02   0.040  1.617      1   0.0397 1.6169  Parse::Eyapp::Lalr::_LR0
 0.77   0.030  0.030   1185   0.0000 0.0000  Parse::Eyapp::Lalr::_FirstSfx
 0.71   0.028  0.039      1   0.0284 0.0387  Parse::Eyapp::Grammar::RulesTable
 0.54   0.021  0.021   1650   0.0000 0.0000  Parse::Eyapp::Grammar::classname
```

Presente un informe del perfil de su compilador. Añade el informe del perfil al `MANIFEST` para que se incluya en la distribución que subas.

21. El módulo `Devel::Size` proporciona la posibilidad de conocer cuanto ocupa una estructura de datos. Considere el siguiente ejemplo:

```
 71 ................................... codigo omitido
 72
 73 use Devel::Size qw(size total_size);
 74 use Perl6::Form;
 75
 76 sub sizes {
 77   my $d = shift;
 78   my ($ps, $ts) = (size($d), total_size($d));
 79   my $ds = $ts-$ps;
 80   return ($ps, $ds, $ts);
 81 }
 82
 83 print form(
 84 ' ============================================================',
 85 '| VARIABLE | SOLO ESTRUCTURA |     SOLO DATOS |          TOTAL |',
 86 '|----------+-----------------+----------------+----------------|',
 87 '| $parser  | {>>>>>>} bytes  | {>>>>>>} bytes | {>>>>>>} bytes |', sizes($parser),
 88 '| $t       | {>>>>>>} bytes  | {>>>>>>} bytes | {>>>>>>} bytes |', sizes($t),
 89 ' ============================================================',
 90 );
```

Al ejecutarlo se obtiene esta salida:

```
 ....... salida previa omitida
```

```
 ============================================================
| VARIABLE | SOLO ESTRUCTURA |     SOLO DATOS |          TOTAL |
|----------+-----------------+----------------+----------------|
| $parser  |      748 bytes  |      991 bytes |     1739 bytes |
| $t       |       60 bytes  |     1237 bytes |     1297 bytes |
 ============================================================
```

Elabore un informe con el consumo de memoria de las variables mas importantes de su programa. Añadelo el informe al `MANIFEST` para que se incluya en la distribución que subas. Explica en el `README` o en la documentación el significado de los ficheros de informe.

33.4.3. Repaso: Pruebas en el Análisis Léxico

1. ¿Cuál es la diferencia entre los operadores `==` y `eq`?
2. ¿Cuáles son los parámetros de la función `ok`?
3. ¿Cuáles son los parámetros de la función `is`?
4. ¿Porqué es conveniente nombrar las pruebas con un nombre que empiece por un número?
5. ¿Como puedo ejecutar los tests en modo *verbose*?
6. ¿Como puedo probar un código que produce la detención del programa?
7. ¿Que contiene la variable `$@`?
8. ¿Que hace la función `like`?
9. ¿Que contiene la variable `$#-`? ¿Y `$+`? (Consulte 31.1.4)
10. ¿Porqué la función `use_ok` es llamada dentro de un `BEGIN`?
11. ¿Que es una prueba `SKIP`?
12. ¿Que es una prueba `TODO`?
13. ¿Que hace la función `pod_file_ok`? ¿A que módulo pertenece?
14. ¿Que hace el operador `tr`?
15. ¿Qué devuelve el operador `tr`?
16. ¿Que hace la opción `d` de `tr`? (consulte 31.5)
17. Explique la conducta de la siguiente sesión con el depurador:

```
  DB<1> $a = '"Focho \"mucha\" chufa"'
  DB<2> print $a
"Focho \"mucha\" chufa"
  DB<3> print $& if $a =~ m{^"([^"]|\\")*"}
"Focho \"
  DB<4> print $& if $a =~ m{^"(\\"||[^"])*"}
"Focho \"mucha\" chufa"
```

18. Supuesto que tuvieramos el operador menos en nuestro lenguaje y dada la entrada `a = 2-3`, ¿Que devolverá su analizador léxico? ¿Devuelve `ID PUN NUM NUM` o bien `ID PUN NUM PUN NUM`? (hemos simplificado el flujo eliminando los atributos).

19. ¿Que hace la llamada `use Test::More qw(no_plan);`?

20. ¿Que hace la función `can_ok`? ¿Qué argumentos tiene?

21. Explique las causas de la siguiente conducta del depurador:

```
  DB<1> $a='4+5'
  DB<2> print "($&) " while ($a =~ m/(\G\d+)/gc) or ($a =~ m/(\G\+)/gc);
(4) (+) (5)
  DB<3> $a='4+5' # inicializamos la posición de búsqueda
  DB<4> print "($&) " while ($a =~ m/(\G\d+)/g) or ($a =~ m/(\G\+)/g);
(4)
  DB<5> $a='4+5'
  DB<6> print "($&) " while ($a =~ m/\G\d+|\G\+/g)
(4) (+) (5)
```

22. ¿Que diferencia hay entre `is_deeply` e `is`?

23. ¿Que argumentos recibe la función `throws_ok`? ¿En que módulo se encuentra?

24. ¿Que hace el comando `HARNESS_PERL_SWITCHES=-MDevel::Cover make test`?

25. ¿Cómo se interpreta el cubrimiento de las sentencias? ¿y de las subrutinas? ¿y de las ramas? ¿y las condiciones lógicas? ¿En cual de estos factores es realista y deseable lograr un cubrimiento del %100 con nuestras pruebas?

26. ¿Que pasa si después de haber desarrollado un número de pruebas cambio la interfaz de mi API?

27. ¿Que hace el comando `perl -d:DProf programa`? ¿Para que sirve?

33.5. Conceptos Básicos para el Análisis Sintáctico

Suponemos que el lector de esta sección ha realizado con éxito un curso en teoría de autómatas y lenguajes formales. Las siguientes definiciones repasan los conceptos mas importantes.

Definición 33.5.1. *Dado un conjunto A, se define A^* el cierre de Kleene de A como: $A^* = \cup_{n=0}^{\infty} A^n$*

Se admite que $A^0 = \{\epsilon\}$, donde ϵ denota la palabra vacía, esto es la palabra que tiene longitud cero, formada por cero símbolos del conjunto base A.

Definición 33.5.2. *Una gramática G es una cuaterna $G = (\Sigma, V, P, S)$. Σ es el conjunto de terminales. V es un conjunto (disjunto de Σ) que se denomina conjunto de* variables sintácticas *o* categorías gramáticales*, P es un conjunto de pares de $V \times (V \cup \Sigma)^*$. En vez de escribir un par usando la notación $(A, \alpha) \in P$ se escribe $A \to \alpha$. Un elemento de P se denomina producción. Por último, S es un símbolo del conjunto V que se denomina símbolo de arranque.*

Definición 33.5.3. *Dada una gramática $G = (\Sigma, V, P, S)$ y $\mu = \alpha A \beta \in (V \cup \Sigma)^*$ una frase formada por variables y terminales y $A \to \gamma$ una producción de P, decimos que μ deriva en un paso en $\alpha\gamma\beta$. Esto es, derivar una cadena $\alpha A \beta$ es sustituir una variable sintáctica A de V por la parte derecha γ de una de sus reglas de producción. Se dice que μ deriva en n pasos en δ si deriva en $n-1$ pasos en una cadena $\alpha A \beta$ la cual deriva en un paso en δ. Se escribe entonces que $\mu \stackrel{*}{\Longrightarrow} \delta$. Una cadena deriva en 0 pasos en si misma.*

Definición 33.5.4. *Dada una gramática $G = (\Sigma, V, P, S)$ se denota por $L(G)$ o lenguaje generado por G al lenguaje:*

$$L(G) = \{x \in \Sigma^* : S \stackrel{*}{\Longrightarrow} x\}$$

Esto es, el lenguaje generado por la gramática G esta formado por las cadenas de terminales que pueden ser derivados desde el símbolo de arranque.

Definición 33.5.5. *Una derivación que comienza en el símbolo de arranque y termina en una secuencia formada por sólo terminales de* Σ *se dice* completa.

Una derivación $\mu \stackrel{*}{\Longrightarrow} \delta$ *en la cual en cada paso* αAx *la regla de producción aplicada* $A \to \gamma$ *se aplica en la variable sintáctica mas a la derecha se dice* una derivación a derechas

Una derivación $\mu \stackrel{*}{\Longrightarrow} \delta$ *en la cual en cada paso* $xA\alpha$ *la regla de producción aplicada* $A \to \gamma$ *se aplica en la variable sintáctica mas a la izquierda se dice* una derivación a izquierdas

Definición 33.5.6. *Observe que una derivación puede ser representada como un árbol cuyos nodos están etiquetados en* $V \cup \Sigma$*. La aplicación de la regla de producción* $A \to \gamma$ *se traduce en asignar como hijos del nodo etiquetado con A a los nodos etiquetados con los símbolos* $X_1 \dots X_n$ *que constituyen la frase* $\gamma = X_1 \dots X_n$*. Este árbol se llama* árbol sintáctico concreto *asociado con la derivación.*

Definición 33.5.7. *Observe que, dada una frase* $x \in L(G)$ *una derivación desde el símbolo de arranque da lugar a un árbol. Ese árbol tiene como raíz el símbolo de arranque y como hojas los terminales* $x_1 \dots x_n$ *que forman x. Dicho árbol se denomina* árbol de análisis sintáctico concreto *de x. Una derivación determina una forma de recorrido del árbol de análisis sintáctico concreto.*

Definición 33.5.8. *Una gramática G se dice ambigua si existe alguna frase* $x \in L(G)$ *con al menos dos árboles sintácticos. Es claro que esta definición es equivalente a afirmar que existe alguna frase* $x \in L(G)$ *para la cual existen dos derivaciones a izquierda (derecha) distintas.*

33.5.1. Ejercicio

Dada la gramática con producciones:

program → declarations statements | statements
declarations → declaration ';' declarations | declaration ';'
declaration → INT idlist | STRING idlist
statements → statement ';' statements | statement
statement → ID '=' expression | P expression
expression → term '+' expression | term
term → factor '*' term | factor
factor → '(' expression ')' | ID | NUM | STR
idlist → ID ',' idlist | ID

En esta gramática, Σ esta formado por los caracteres entre comillas simples y los símbolos cuyos identificadores están en mayúsculas. Los restantes identificadores corresponden a elementos de V. El símbolo de arranque es $S =$ program.

Conteste a las siguientes cuestiones:

1. Describa con palabras el lenguaje generado.
2. Construya el árbol de análisis sintáctico concreto para cuatro frases del lenguaje.
3. Señale a que recorridos del árbol corresponden las respectivas derivaciones a izquierda y a derecha en el apartado 2.
4. ¿Es ambigua esta gramática?. Justifique su respuesta.

33.6. Análisis Sintáctico Predictivo Recursivo

La siguiente fase en la construcción del analizador es la fase de análisis sintáctico. Esta toma como entrada el flujo de terminales y construye como salida el árbol de análisis sintáctico abstracto.

El árbol de análisis sintáctico abstracto es una representación compactada del árbol de análisis sintáctico concreto que contiene la misma información que éste.

Existen diferentes métodos de análisis sintáctico. La mayoría caen en una de dos categorías: ascendentes y descendentes. Los ascendentes construyen el árbol desde las hojas hacia la raíz. Los descendentes lo hacen en modo inverso. El que describiremos aqui es uno de los mas sencillos: se denomina método de análisis predictivo descendente recursivo.

33.6.1. Introducción

En este método se asocia una subrutina con cada variable sintáctica $A \in V$. Dicha subrutina (que llamaremos `A`) reconocerá el lenguaje generado desde la variable A:

$$L_A(G) = \{x \in \Sigma^* : A \stackrel{*}{\Longrightarrow} x\}$$

En este método se escribe una rutina `A` por variable sintáctica $A \in V$. Se le da a la rutina asociada el mismo nombre que a la variable sintáctica asociada. La función de la rutina `A` asociada con la variable $A \in V$ es reconocer el lenguaje $L(A)$ generado por A. La estrategia general que sigue la rutina `A` para reconocer $L(A)$ es decidir en términos del terminal a en la entrada que regla de producción concreta $A \to \alpha$ se aplica para a continuación comprobar que la entrada que sigue pertenece al lenguaje generado por α. En un analizador predictivo descendente recursivo (APDR) se asume que el símbolo que actualmente esta siendo observado (denotado `lookahead`) permite determinar unívocamente que producción de A hay que aplicar. Una vez que se ha determinado que la regla por la que continuar la derivación es $A \to \alpha$ se procede a reconocer $L_\alpha(G)$, el lenguaje generado por α. Si $\alpha = X_1 \dots X_n$, las apariciones de terminales X_i en α son emparejadas con los terminales en la entrada mientras que las apariciones de variables $X_i = B$ en α se traducen en llamadas a la correspondiente subrutina asociada con `B`.

Para ilustrar el método, simplificaremos la gramática presentada en el ejercicio 34.1.1 eliminando las declaraciones:

statements → statement ';' statements | statement
statement → ID '=' expression | P expression
expression → term '+' expression | term
term → factor '*' term | factor
factor → '(' expression ')' | ID | NUM

La secuencia de llamadas cuando se procesa la entrada mediante el siguiente programa construye "implícitamente" el árbol de análisis sintáctico concreto.

Dado que estamos usando `strict` se requiere prototipar las funciones al comienzo del fichero:

```
sub parse();
sub statements();
sub statement();
sub expression();
sub term();
sub factor();
sub idlist();
sub declaration();
sub declarations();
```

Para saber mas sobre prototipos consulte [10].

Programa 33.6.1.
```
 1 sub match {
 2   my $t = shift;
 3
 4   if ($lookahead eq $t) {
 5     ($lookahead, $value) = splice @tokens,0,2;
```

```
 6     if (defined($lookahead)) {
 7       $lookahead = $value if ($lookahead eq 'PUN');
 8     } else { $lookahead = 'EOI'; }
 9   }
10   else { error("Se esperaba $t y se encontro $lookahead\n"); }
11 }
12
13 sub statement {
14   if ($lookahead eq 'ID') { match('ID'); match('='); expression; }
15   elsif ($lookahead eq 'P') { match('P'); expression; }
16   else { error('Se esperaba un identificador'); }
17 }
18
19 sub term() {
20   factor;
21   if ($lookahead eq '*') { match('*'); term; }
22 }
23
24 sub expression() {
25   term;
26   if ($lookahead eq '+') { match('+'); expression; }
27 }
28
29 sub factor() {
30   if ($lookahead eq 'NUM') { match('NUM'); }
31   elsif ($lookahead eq 'ID') { match('ID'); }
32   elsif ($lookahead eq '(') { match('('); expression; match(')'); }
33   else { error("Se esperaba (, NUM o ID"); }
34 }
35
36 sub statements {
37   statement;
38   if ($lookahead eq ';') { match(';'); statements; }
39 }
40
41 sub parser {
42   ($lookahead, $value) = splice @tokens,0,2;
43   statements; match('EOI');
44 }
```

Como vemos en el ejemplo, el análisis predictivo confía en que, si estamos ejecutando la entrada del procedimiento `A`, el cuál está asociado con la variable $A \in V$, el símbolo terminal que esta en la entrada a determine de manera unívoca la regla de producción $A \to a\alpha$ que debe ser procesada.

Si se piensa, esta condición requiere que todas las partes derechas α de las reglas $A \to \alpha$ de A "comiencen" por diferentes símbolos. Para formalizar esta idea, introduciremos el concepto de conjunto $FIRST(\alpha)$:

Definición 33.6.1. *Dada una gramática $G = (\Sigma, V, P, S)$ y un símbolo $\alpha \in (V \cup \Sigma)^*$ se define el conjunto $FIRST(\alpha)$ como:*

$$FIRST(\alpha) = \left\{ b \in \Sigma : \alpha \stackrel{*}{\Longrightarrow} b\beta \right\} \cup N(\alpha)$$

donde:

$$N(\alpha) = \begin{cases} \{\epsilon\} & si\ \alpha \stackrel{*}{\Longrightarrow} \epsilon \\ \emptyset & en\ otro\ caso \end{cases}$$

Podemos reformular ahora nuestra afirmación anterior en estos términos: Si $A \to \gamma_1 \mid \ldots \mid \gamma_n$ y los conjuntos $FIRST(\gamma_i)$ son disjuntos podemos construir el procedimiento para la variable A siguiendo este seudocódigo:

```
sub A {
  if ($lookahead in FIRST(gamma_1)) { imitar gamma_1 }
  elsif ($lookahead in FIRST(gamma_2)) { imitar gamma_2 }
  ...
  else ($lookahead in FIRST(gamma_n)) { imitar gamma_n }
}
```

Donde si γ_j es $X_1 \ldots X_k$ el código `gamma_j` consiste en una secuencia $i = 1 \ldots k$ de llamadas de uno de estos dos tipos:

- Llamar a la subrutina `X_i` si X_i es una variable sintáctica
- Hacer una llamada a `match(X_i)` si X_i es un terminal

33.6.2. Ejercicio: Recorrido del árbol en un ADPR

¿En que forma es recorrido el árbol de análisis sintáctico concreto en un analizador descendente predictivo recursivo? ¿En que orden son visitados los nodos?

33.6.3. Ejercicio: Factores Comunes

En el programa 33.6.1 el reconocimiento de las categorías gramáticales statements, expression y term (líneas 19-27) difiere del resto. Observe las reglas:

statements → statement ';' statements | statement
expression → term '+' expression | term
term → factor '*' term | factor

¿Son disjuntos los conjuntos $FIRST(\gamma_i)$ para las partes derechas de las reglas de statements? ¿Son disjuntos los conjuntos $FIRST(\gamma_i)$ para las partes derechas de las reglas de expression? ¿Son disjuntos los conjuntos $FIRST(\gamma_i)$ para las partes derechas de las reglas de term?

Si se tiene una variable con producciones:

$A \to \alpha\beta \mid \alpha\gamma$

Las dos producciones tienen un *máximo factor común* en la izquierda de su parte derecha α. Asumimos que $FIRST(\beta) \cap FIRST(\gamma) = \emptyset$.

1. ¿Cómo puede modificarse la gramática para obtener una nueva gramática que cumpla la condición de que las partes derechas tienen conjuntos $FIRST(\gamma_i)$ disjuntos?
2. ¿Puede modificarse la técnica APDR para que funcione sobre gramáticas con este tipo de producciones?. Observe el código asociado con statements, expression y term. ¿Cómo sería el esquema general?

33.6.4. Derivaciones a vacío

Surge un problema cuando $A \to \gamma_1 \mid \ldots \mid \gamma_n$ y la palabra vacía está en alguno de los conjuntos $FIRST(\gamma_i)$. ¿Que hacer entonces?

Nótese que si $A \to \gamma$ y $\epsilon \in FIRST(\gamma)$ es porque existe una derivación $\gamma \stackrel{*}{\Longrightarrow} \epsilon$. ¿Que terminales podemos legalmente encontrarnos cuando estamos en la subrutina `A`? Consideremos una derivación desde el símbolo de arranque en la que se use la producción $A \to \gamma$. Dicha derivación forzosamente tendrá la forma:

$$S \stackrel{*}{\Longrightarrow} \beta A\ a\mu \Longrightarrow \beta\gamma\ a\mu \stackrel{*}{\Longrightarrow} \beta\ a\mu.$$

Cualquier terminal $a \in \Sigma$ que pueda aparecer en una derivación desde el símbolo de arranque inmediatamente a continuación de la variable A es susceptible de ser visto cuando se esta analizando A y se aplicó $A \to \gamma$ con $\gamma \stackrel{*}{\Longrightarrow} \epsilon$. Esto nos lleva a la definición del conjunto $FOLLOW(A)$ como conjunto de terminales que pueden aparecer a continuación de A en una derivación desde el símbolo de arranque:

Definición 33.6.2. *Dada una gramática $G = (\Sigma, V, P, S)$ y una variable $A \in V$ se define el conjunto $FOLLOW(A)$ como:*

$$FOLLOW(A) = \left\{ b \in \Sigma : \exists\ S \stackrel{*}{\Longrightarrow} \alpha A b \beta \right\} \cup E(A)$$

donde

$$E(A) = \begin{cases} \{\$\} & si\ S \stackrel{*}{\Longrightarrow} \alpha A \\ \emptyset & en\ otro\ caso \end{cases}$$

Aqui \$ denota el final de la entrada (que se corresponde en el código Perl anterior con el terminal `EOI`).

Si $A \to \gamma_1 \mid \ldots \mid \gamma_n$ dado que los conjuntos $FIRST(\gamma_i)$ han de ser disjuntos para que un analizador predictivo APDR funcione, sólo una parte derecha puede contener la palabra vacía en su $FIRST$. Supongamos que es γ_n. Podemos reformular la construcción del procedimiento para la variable A siguiendo este seudocódigo:

```
sub A {
  if ($lookahead in FIRST(gamma_1)) { imitar gamma_1 }
  elsif ($lookahead in FIRST(gamma_2)) { imitar gamma_2 }
  ...
  else ($lookahead in FIRST(gamma_n) or $lookahead in FOLLOW(A)) { imitar gamma_n }
}
```

Un caso particular de $\gamma_n \stackrel{*}{\Longrightarrow} \epsilon$ es que $\gamma_n = \epsilon$. En tal caso, y como es obvio, el significado de `imitar gamma_n` es equivalente a ejecutar una sentencia vacía.

33.6.5. Construcción de los conjuntos de Primeros y Siguientes

Algoritmo 33.6.1. *Construcción de los conjuntos $FIRST(X)$*

Repita el siguiente conjunto de reglas hasta que no se puedan añadir mas símbolos terminales o a ningún conjunto $FIRST(X)$:

1. *Si $X \in \Sigma$ entonces $FIRST(X) = X$*
2. *Si $X \to \epsilon$ entonces $FIRST(X) = FIRST(X) \cup \{\epsilon\}$*
3. *Si$X \in V$ y $X \to Y_1 Y_2 \cdots Y_k \in P$ entonces*

$$\begin{array}{l} i = 1; \\ hacer \\ \quad FIRST(X) = FIRST(X) \cup FIRST^*(Y_i); \\ \quad i++; \\ mientras\ (i \leq k\ y\ \epsilon \in FIRST(Y_i)) \end{array}$$

4. *Añadir ϵ a $FIRST(X)$ si $i \geq k$ y $\epsilon \in FIRST(Y_k)$*

Aqui $FIRST^*(Y)$ denota al conjunto $FIRST(Y) - \{\epsilon\}$.

Este algoritmo puede ser extendido para calcular $FIRST(\alpha)$ para $\alpha = X_1 X_2 \cdots X_n \in (V \cup \Sigma)^*$. El esquema es anólogo al de un símbolo individual.

Algoritmo 33.6.2. *Construcción del conjunto $FIRST(\alpha)$*

Repita siguiente conjunto de reglas hasta que no se puedan añadir mas símbolos terminales o a ningún conjunto $FIRST(\alpha)$:

$$
\begin{aligned}
&i = 1;\\
&FIRST(\alpha) = \emptyset;\\
&hacer\\
&\quad FIRST(\alpha) = FIRST(\alpha) \cup FIRST^{*}(X_i);\\
&\quad i++;\\
&mientras\ (i \leq n\ y\ \epsilon \in FIRST(X_i))
\end{aligned}
$$

Algoritmo 33.6.3. *Construcción de los conjuntos $FOLLOW(A)$ $\forall A \in V$:*

Repetir los siguientes pasos hasta que ninguno de los conjuntos $FOLLOW$ cambie:

1. *$FOLLOW(S) = \{\$\}$ ($ representa el final de la entrada)*
2. *Si $A \rightarrow \alpha B \beta$ entonces*

$$FOLLOW(B) = FOLLOW(B) \cup (FIRST(\beta) - \{\epsilon\})$$

3. *Si $A \rightarrow \alpha B$ o $A \rightarrow \alpha B \beta$ y $\epsilon \in FIRST(\beta)$ entonces*

$$FOLLOW(B) = FOLLOW(B) \cup FOLLOW(A)$$

33.6.6. Ejercicio: Construir los $FIRST$

Construya los conjuntos $FIRST$ de las partes derechas de las reglas de producción de la gramática presentada en el ejercicio 34.1.1.

33.6.7. Ejercicio: Calcular los $FOLLOW$

Modificamos la gramática de la sección 33.6.1 para que admita la sentencia vacía:

statements → statement ';' statements | statement
statement → ID '=' expression | P expression | ϵ
expression → term '+' expression | term
term → factor '*' term | factor
factor → '(' expression ')' | ID | NUM

Calcule los conjuntos $FOLLOW$. ¿Es la nueva gramática susceptible de ser analizada por un analizador predictivo descendente recursivo? ¿Cómo sería el código para la subrutina `statements`?. Escríbalo.

33.6.8. Práctica: Construcción de los FIRST y los FOLLOW

He escrito un módulo llamado Grammar que provee la función `Grammar::Parse` la cual recibe una cadena conteniendo la gramática en formato `yacc` o `eyapp` y devuelve una referencia a un hash conteniendo la información pertinente para el tratamiento de la gramática. Para instalar el módulo tenga en cuenta que depende del módulo `Parse::Yapp`.

Para ilustrar el uso vea los ejemplos en el directorio `scripts`. En concreto veamos el programa `grammar.pl`.

```
Grammar/scripts$ cat -n grammar.pl
 1  #!/usr/bin/perl -w -I../lib
 2  use strict;
 3  use Grammar;
```

```
 4  use Data::Dumper;
 5
 6  sub usage {
 7    print <<"EOI";
 8  usage:
 9  $0 input_grammar
10  EOI
11    die "\n";
12  }
13
14  usage() unless @ARGV;
15  my $filename = shift;
16
17  local $/ = undef;
18  open my $FILE, "$filename";
19  my $grammar = <$FILE>;
20  my $x = Grammar::Parse($grammar);
21
22  print Dumper($x);
```

Vamos a darle como entrada la gramática en el fichero `aSb.yp` conteniendo una gramática:

```
Grammar/scripts$ cat -n aSb.yp
 1  %%
 2  S:
 3      |   'a' S 'b'
 4  ;
 5  %%
```

Las gramáticas aceptadas por `Grammar::Parse` se adaptan a la sintáxis de las gramáticas reconocidas por `Parse::Yapp`. Una gramática (normalmente con tipo `.yp`) consta de tres partes: la cabeza, el cuerpo y la cola. Cada una de las partes va separada de las otras por el símbolo `%%` en una línea aparte. Así, el `%%` de la línea 1 separa la cabeza del cuerpo. En la cabecera se colocan las declaraciones de terminales (directiva `%token`), cual es el símbolo de arranque (directiva `%start`), etc. El cuerpo contiene las reglas de la gramática y las acciones asociadas. Por último, la cola en nuestro caso no es usada y es vacía. En general, la cola contiene las rutinas de soporte al código que aparece en las acciones asi como, posiblemente, rutinas para el análisis léxico y el tratamiento de errores.

La salida de `Grammar::Parse` es una referencia a un hash cuyas entradas vienen explicadas por los comentarios.

```
Grammar/scripts$ grammar.pl aSb.yp
$VAR1 = {
    'SYMS' => { 'S' => 2, '"b"' => 3, '"a"' => 3 }, # Símbolo => línea
    'NULL' => { 'S' => 1 }, # símbolos que se anulan
    'RULES' => [
                [ 'S', [] ], # S produce vacío
                [ 'S', [ '"a"', 'S', '"b"' ] ] # S -> aSb
              ],
    'START' => 'S', # Símbolo de arranque
    'TERM' => [ '"b"', '"a"' ], # terminales /tokens
    'NTERM' => { 'S' => [ 0, 1 ] }  # índices de las reglas de las variables sintácticas
  };
```

Usando la estructura devuelta por la función `Grammar::Parse` escriba un módulo que provea funciones para computar los `FIRST` y los `FOLLOW` de las variables sintácticas de la gramática. No olvide escribir

la documentación. Incluya una prueba por cada una de las gramáticas que figuran en el directorio `scripts` del módulo `Grammar`.

Puede encontrar la práctica *casi hecha* en PL::FirstFollow. Asegúrese de entender el algoritmo usado. Aumente el número de pruebas y haga un análisis de cubrimiento.

33.6.9. Gramáticas LL(1)

Una gramática $G = (\Sigma, V, P, S)$ cuyo lenguaje generado $L(G)$ puede ser analizado por un analizador sintáctico descendente recursivo predictivo se denomina *LL(1)*. Una gramática es LL(1) si y sólo si para cualesquiera dos producciones $A \to \alpha$ y $A \to \beta$ de G se cumple:

1. $FIRST(\alpha) \cap FIRST(\beta) = \emptyset$
2. Si $\epsilon \in FIRST(\alpha)$, entonces $FIRST(\alpha) \cap FOLLOW(A) = \emptyset$

¿De donde viene el nombre LL(1)? La primera L hace alusión al hecho de que el flujo de terminales se lee de izquierda a derecha, accediendo a la entrada por su izquierda (*Left*). La segunda L se refiere a que el método de análisis predictivo construye una derivación a izquierdas. El número entre paréntesis indica el número de terminales que debemos consultar para decidir que regla de producción se aplica. Asi, en una gramática LL(2) la decisión final de que producción elegir se hace consultando los dos terminales a la entrada.

33.6.10. Ejercicio: Caracterización de una gramática LL(1)

Cuando se dice que una gramática es LL(1) si, y sólo si:

1. $FIRST(\alpha) \cap FIRST(\beta) = \emptyset$
2. Si $\epsilon \in FIRST(\alpha)$, entonces $FIRST(\alpha) \cap FOLLOW(A) = \emptyset$

se asume que los conjuntos $FIRST(\alpha)$ no son vacíos.

- ¿Que se puede decir de la regla $A \to \alpha$ si $FIRST(\alpha) = \emptyset$?
- ¿Que se puede decir de la variable A si $FOLLOW(A) = \emptyset$?

33.6.11. Ejercicio: Ambiguedad y LL(1)

¿Puede una gramática LL(1) ser ambigua?. Razone su respuesta.

33.6.12. Práctica: Un analizador APDR

Siguiendo con la construcción del compilador para el lenguaje Tutu, escriba un analizador APDR para la siguiente gramática. Reutilice el código de las prácticas de las secciones anteriores (33.3 y 33.4).

program → declarations statements | statements
declarations → declaration ';' declarations | declaration ';'
declaration → INT idlist | STRING idlist
statements → statement ';' statements | statement
statement → ID '=' expression | P expression | ϵ
expression → term '+' expression | term
term → factor '*' term | factor
factor → '(' expression ')' | ID | NUM | STR
idlist → ID ',' idlist | ID

33.6.13. Práctica: Generación Automática de Analizadores Predictivos

Objetivo Escriba un módulo `GAP.pm` que provea una subrutina `gap` para la generación automática de un APDR supuesto que la gramática de entrada es LL(1).

La subrutina `gap` recibe como entrada la gramática según la estructura de datos generada por la función `Grammar::Parse` de la versión 0.3 del módulo Grammar.

El Módulo `Grammar` La estructura de datos generada por la función `Grammar::Parse` se explicó en la práctica 33.6.8. La estructura ha sido extendida en esta versión para incluir el código que se sitúe en la zona de cola. Por ejemplo, dada la gramática de entrada:

```
Grammar/03/scripts$ cat -n aSb.yp
     1  %%
     2  S:
     3      |   'a' S 'b'
     4  ;
     5  %%
     6
     7  sub Lex {
     8    local $_ = shift; # input
     9    my @tokens;
    10
    11
    12    while ($_) {
    13      s/^\s*//; # fuera blancos
    14      push @tokens, $1, $1 if s/^(.)//s
    15    }
    16    @tokens;
    17  }
    18
    19  sub main {
    20    my $filename = shift;
    21    my $input;
    22
    23    if (defined($filename)) {
    24      local $/ = undef;
    25      open my $FILE, $filename or die "No se pudo abrir $filename\n";
    26      $input = <$FILE>;
    27      close($FILE);
    28    }
    29    else { $input = <STDIN> }
    30
    31    my @tokens = Lex($input);
    32    Parse(@tokens); # Llamada al analizador generado
    33    print "Sintácticamente correcto\n";
    34  }
```

se genera la siguiente estructura de datos:

```
{
  'SYMS' => { 'S' => 2, 'b' => 3, 'a' => 3 },  # Símbolo => línea de aparición
  'NULL' => { 'S' => 1 }, # Símbolos que se anulan
  'RULES' => [ # Reglas
               [ 'S', [] ], # S produce vacío
               [ 'S', [ 'a', 'S', 'b' ] ] # S-> a S b
```

```
            ],
  'START' => 'S', # Símbolo de arranque
  'TERM' => [ 'b', 'a' ], # Terminales
  'NTERM' => { 'S' => [ 0, 1 ] } # Variables sintácticas e índices de las reglas de esa variab
  'TAIL' => [ # [ 'Código de cola', línea en la que está el segundo %% ]
  '

  sub Lex {
    local $_ = shift; # input
    my @tokens;


    while ($_) {
      s/^\\s*//; # fuera blancos
      push @tokens, $1, $1 if s/^(.)//s
    }
    @tokens;
  }

  sub main {
    my $filename = shift;
    my $input;

    if (defined($filename)) {
      local $/ = undef;
      open my $FILE, $filename or die "No se pudo abrir $filename\\n";
      $input = <$FILE>;
      close($FILE);
    }
    else { $input = <STDIN> }

    my @tokens = Lex($input);
    my $ok = Parse(@tokens); # Llamada al analizador generado
    print "Sintácticamente correcto\\n" if $ok;
  }

  ', 5 ], # línea en la que está el segundo %%
 };
```

Asi pues la entrada con clave `TAIL` contiene el código auxiliar de cola. Este código debe ser incluido por su programa dentro del texto del paquete generado por `gap`.

Descripción del objetivo: La función `gap` La función `gap` también recibe como entrada el nombre del package:

```
$package_text = &gap($grammar, 'Package_name');
```

La función `gap` retorna una cadena conteniendo el `package` en el que estan las subrutinas del analizador sintáctico.

La idea es que dicha cadena se salvará en un fichero con nombre `Package_name.pm` que podrá posteriormente ser usado (`use Package_name`) por un programa que necesite analizar entradas que se conforman de acuerdo a la especificación de la gramática.

Descripción del objetivo: La función `parser` La rutina principal del paquete generado se ha de llamar `parser` (esto es, su nombre completo es: `Package_name::parser`. Evidentemente `Package_name`

debe ser un nombre Perl válido). Ninguna subrutina deberá ser exportada sino que deberán ser llamadas por su nombre completo.

La subrutina `parser` recibe como argumento el array de terminales, obtiene el primer terminal y llama a la subrutina asociada con el símbolo de arranque. Los terminales están representados como parejas $(terminal, atributo)$.

Observe que, una vez que la cadena `$package_text` conteniendo el paquete ha sido generada y salvada en un fichero con nombre `Package_name.pm`, podemos escribir un programa cliente:

```
use strict;
use Package_name;

&Package_name::main;
```

Este programa espera una entrada desde fichero o `STDIN` e informa si dicha entrada es sintácticamente correcta o no para la gramática en cuestión.

Cálculo de los First y los Follow con `PL::FirstFollow` Para facilitar la escritura de `GAP.pm` pueden hacer uso del módulo PL::FirstFollow el cual calcula los $FIRST$ y los $FOLLOW$. El módulo `PL::FirstFollow` depende de Set::Scalar escrito por Jarkko Hietaniemi: instálelo primero.

Deberá familiarizarse con `PL::FirstFollow`, rellenar la documentación de todas las subrutinas (apariciones de `????` en el texto) y escribir la documentación siguiendo el template que se provee. *Rellene los fragmentos de código que se han sustituido por signos de interrogación.* Haga un estudio de cubrimiento y añada pruebas para mejorar el actual. El actual cubrimiento es:

```
---------------------------- ------ ------ ------ ------ ------ ------ ------
File                          stmt   bran   cond    sub    pod   time  total
---------------------------- ------ ------ ------ ------ ------ ------ ------
...ammar-0.03/lib/Grammar.pm  100.0    n/a    n/a  100.0    0.0   75.3   97.2
blib/lib/PL/FirstFollow.pm    100.0   92.9   50.0  100.0    0.0   24.7   95.1
Total                         100.0   92.9   50.0  100.0    0.0  100.0   95.5
---------------------------- ------ ------ ------ ------ ------ ------ ------
```

Si observa un fallo en `PL::FirstFollow` háganoslo saber y además de resolverlo escriba una prueba para detectar el fallo.

Haga un estudio de profiling de su aplicación.

Uso de Templates Un módulo que puede facilitar la escritura de esta práctica es Text::Template debido a Mark Jason Dominus. El siguiente ejemplo de uso es un fragmento de un traductor - que nunca acabo de terminar - que toma con fuente un fichero en el formato que usa Moodle para los cuestionarios (conocido como formato GIFT) y lo convierte en un cuestionario LaTeX:

```
lhp@nereida:~/projects/Gift2LaTeX/Gift2LaTeX/lib$ cat -n Gift2LaTeX.pm
   1  package Gift2LaTeX;
   2
   3  use strict;
   4  use warnings;
   5  use Gift;
   6  use Text::Template;
   7  use HTML::Latex;
  ..    ..................................................
  49  package Gift::TRUEFALSE; # True-false questions belong to this class
  50
  51  { # closure
  52
  53    die "Can't find $TEMPLATE_DIR/TRUEFALSE_question.tep\n"
```

```
 54          unless -e "$TEMPLATE_DIR/TRUEFALSE_question.tep";
 55    my $tfq_tmpl = Text::Template->new( #tfq = true-false question
 56      DELIMITERS => ['%<', '%>'];
 57      SOURCE => "$TEMPLATE_DIR/TRUEFALSE_question.tep",
 58    );
 ..    ..................................................
 67    sub gen_latex {
 68      my $self = shift;
 69
 70      ########## Generate latex for question
 71      my $prefix  = $self->PREFIX;
 72
 73      my $sufix = $self->POSTSTATE;
 74
 75      $self->Error("Only HTML and PLAIN formats are supported\n")
 76          unless (!$self->FORMAT or ($self->FORMAT =~ m{html|plain}i));
 77
 78      my ($prefix_tex, $sufix_tex);
 79      if (defined($self->FORMAT) and $self->FORMAT =~ m{plain}i) {
 80        $prefix_tex = $prefix;
 81        $sufix_tex  = $sufix;
 82      }
 83      else { # HTML by default
 ..    ..................................................
 86      }
 87      my $params = {
 88        prefix  => $prefix_tex,
 89        sufix   => $sufix_tex,
 90        separator => $separator,
 91        label   => $label_prefix.$question_number,
 92        question_number => $question_number
 93      };
 94      my $question_tex = $tfq_tmpl->fill_in(HASH => $params);
 96      ########## Generate latex for answer
...      ................................
101    }
102  }
```

En la línea 55 se crea el template. El template se lee desde el fichero `"$TEMPLATE_DIR/TRUEFALSE_question.tep"` cuyo contenido es una mezcla de texto (en este caso texto LaTeX y HTML) con código Perl: El código Perl aparece acotado entre los delimitadores `'%<'` y `'%>'`.

```
lhp@nereida:~/projects/Gift2LaTeX/Gift2LaTeX/etc/en$ cat -n TRUEFALSE_question.tep
 1  \ begin{latexonly}
 2    %<$separator%>
 3    \ label{question:%<$label%>}
 4    %<$prefix%>
 5
 6    \ begin{center}
 7      \ begin{tabular}{llll}
 8          $\ bigcirc$ & TRUE & $\ bigcirc$ & FALSE
 9      \ end{tabular}
10
11      \noindent %<$sufix%>
```

```
12    \ end{center}
13  \ end{latexonly}
14
15  \ begin{htmlonly}
16  %<$separator%>
17  \ label{question:%<$label%>}
18  %<$prefix%>
19
20  \ begin{center}
21    \ begin{tabular}{llll}
22        \ htmlref{$\bigcirc$}{answer:%<$label%>} & TRUE &
23        \ htmlref{$\bigcirc$}{answer:%<$label%>} & FALSE
24    \ end{tabular}
25
26    \ noindent %<$sufix%>
27  \ end{center}
28  \ end{htmlonly}
```

El template se rellena en las líneas 87-94. En esa llamada se ejecuta el código Perl incrustado en el esqueleto y su resultado se inserta en la posición que ocupa en el texto.

Concatenación y Documentos `HERE` Cuando concatene sangre adecuadamente las concatenaciones:

```
my $usage = "Usage: $0 <file> [-full] [-o] [-beans]\n"
            . "Options:\n"
            . "    -full  : produce a full dump\n"
            . "    -o     : dump in octal\n"
            . "    -beans : source is Java\n"
            ;
```

ponga el punto al principio de la siguiente línea, no al final.

Pero cuando el número de líneas es grande es mejor usar un *here document* o *documento aqui*. Veamos un ejemplo:

```
print <<"EOI";
El programa se deberá ejecutar con:

$0 numfiles $opciones initialvalue
EOI
```

Para definir un "documento aqui" se escribe la etiqueta entrecomillada y precedida de `<<` y sigue el texto que consituye el *here document* que se delimita por una línea en blanco que empieza por la etiqueta. Al documento aquí se le trata como una cadena de doble comilla si la etiqueta aparece en doble comilla y como de comilla simple si la etiqueta esta entre comillas simples. Observe que el punto y coma se escribe despues de la primera aparición de la etiqueta.

Un problema con el uso de los heredoc es que rompen la estructura normal del sangrado:

```
if ($usage_error) {
    warn <<'END_USAGE';
Usage: qdump <file> [-full] [-o] [-beans]
Options:
    -full  : produce a full dump
    -o     : dump in octal
    -beans : source is Java
END_USAGE
}
```

Es mejor que cada heredoc se aisle en una subrutina y se parametrice con las variables que van a ser interpoladas:

```
sub build_usage {
    my ($prog_name, $file_name) = @_;

    return <<"END_USAGE";
Usage: $prog_name $file_name [-full] [-o] [-beans]
Options:
    -full  : produce a full dump
    -o     : dump in octal
    -beans : source is Java
END_USAGE
}
```

que mas tarde puede ser llamado con los valores de interpolación adecuados:

```
if ($usage_error) {
    warn build_usage($PROGRAM_NAME, $requested_file);
}
```

Véase el libro de Conway Perl Best Practices [**?**] para mas detalles sobre buenas prácticas de programación con heredocs.

Descarga de los Módulos Necesarios

- El módulo `Grammar` : http://nereida.deioc.ull.es/~pl/perlexamples/Grammar-0.03.tar.gz
- El módulo `PL::FirstFollow` : http://nereida.deioc.ull.es/~pl/perlexamples/PL-FirstFollow-0.02.tar.gz

33.7. Esquemas de Traducción

Definición 33.7.1. *Un* esquema de traducción *es una gramática independiente del contexto en la cual se han insertado fragmentos de código en las partes derechas de sus reglas de producción. Los fragmentos de código asi insertados se denominan* acciones semánticas. *Dichos fragmentos actúan, calculan y modifican los atributos asociados con los nodos del árbol sintáctico. El orden en que se evalúan los fragmentos es el de un recorrido primero-profundo del árbol de análisis sintáctico.*

Obsérvese que, en general, para poder aplicar un esquema de traducción hay que construir el árbol sintáctico y después aplicar las acciones empotradas en las reglas en el orden de recorrido primero-profundo. Por supuesto, si la gramática es ambigua una frase podría tener dos árboles y la ejecución de las acciones para ellos podría dar lugar a diferentes resultados. Si se quiere evitar la multiplicidad de resultados (interpretaciones semánticas) es necesario precisar de que árbol sintáctico concreto se esta hablando.

Por ejemplo, si en la regla $A \to \alpha\beta$ insertamos un fragmento de código:

$$A \to \alpha\{action\}\beta$$

La acción $\{action\}$ se ejecutará después de todas las acciones asociadas con el recorrido del subárbol de α y antes que todas las acciones asociadas con el recorrido del subárbol β.

El siguiente esquema de traducción recibe como entrada una expresión en infijo y produce como salida su traducción a postfijo para expresiones aritmeticas con sólo restas de números:

$expr \to expr_1 - NUM$ `{ $expr{TRA} = $expr[1]{TRA}." ".$NUM{VAL}." - "}`
$expr \to NUM$ `{ $expr{TRA} = $NUM{VAL} }`

Las apariciones de variables sintácticas en una regla de producción se indexan como se ve en el ejemplo, para distinguir de que nodo del árbol de análisis estamos hablando. Cuando hablemos del atributo de un nodo utilizaremos una indexación tipo *hash*. Aquí `VAL` es un atributo de los nodos de tipo NUM denotando su valor numérico y para accederlo escribiremos `$NUM{VAL}`. Análogamente `$expr{TRA}` denota el atributo "traducción" de los nodos de tipo *expr*.

Ejercicio 33.7.1. *Muestre la secuencia de acciones a la que da lugar el esquema de traducción anterior para la frase 7 -5 -4.*

En este ejemplo, el cómputo del atributo `$expr{TRA}` depende de los atributos en los nodos hijos, o lo que es lo mismo, depende de los atributos de los símbolos en la parte derecha de la regla de producción. Esto ocurre a menudo y motiva la siguiente definición:

Definición 33.7.2. *Un atributo tal que su valor en un nodo puede ser computado en términos de los atributos de los hijos del nodo se dice que es un* atributo sintetizado.

Ejemplo 33.7.1. *Un ejemplo de atributo heredado es el tipo de las variables en las declaraciones:*

decl → *type* `{ $list{T} = $type{T} }` *list*
type → INT `{ $type{T} = $int }`
type → $STRING$ `{ $type{T} = $string }`
list → ID , `{ $ID{T} = $list{T}; $list_1{T} = $list{T} }` $list_1$
list → ID `{ $ID{T} = $list{T} }`

Definición 33.7.3. *Un* atributo heredado *es aquel cuyo valor se computa a partir de los valores de sus hermanos y de su padre.*

Ejercicio 33.7.2. *Escriba un esquema de traducción que convierta expresiones en infijo con los operadores* `+-*/()` *y números en expresiones en postfijo. Explique el significado de los atributos elegidos.*

33.8. Recursión por la Izquierda

Definición 33.8.1. *Una* gramática es recursiva por la izquierda *cuando existe una derivación* $A \stackrel{*}{\Longrightarrow} A\alpha$.

En particular, es recursiva por la izquierda si contiene una regla de producción de la forma $A \to A\alpha$. *En este caso se dice que la recursión por la izquierda es directa.*

Cuando la gramática es *recursiva por la izquierda*, el método de análisis recursivo descendente predictivo no funciona. En ese caso, el procedimiento `A` asociado con A ciclaría para siempre sin llegar a consumir ningún terminal.

33.8.1. Eliminación de la Recursión por la Izquierda en la Gramática

Es posible modificar la gramática para eliminar la recursión por la izquierda. En este apartado nos limitaremos al caso de recursión por la izquierda directa. La generalización al caso de recursión por la izquierda no-directa se reduce a la iteración de la solución propuesta para el caso directo.

Consideremos una variable A con dos producciones:

$$A \to A\alpha \quad | \beta$$

donde $\alpha, \beta \in (V \cup \Sigma)^*$ no comienzan por A. Estas dos producciones pueden ser sustituidas por:

$$\begin{array}{l} A \to \beta R \\ R \to \alpha R \quad | \epsilon \end{array}$$

eliminando así la recursión por la izquierda.

Definición 33.8.2. *La producción $R \to \alpha R$ se dice* recursiva por la derecha.

Las producciones recursivas por la derecha dan lugar a árboles que se hunden hacia la derecha. Es mas difícil traducir desde esta clase de árboles operadores como el menos, que son asociativos a izquierdas.

Ejercicio 33.8.1. *Elimine la recursión por la izquierda de la gramática*

$expr \to expr - NUM$
$expr \to NUM$

Ejercicio 33.8.2. *¿Que hay de erróneo en este esquema de traducción?*

```
expr → NUM − expr1   { $expr{T} = $NUM{VAL}." ".$expr[1]{T}." - "}
expr → NUM           { $expr{T} = $NUM{VAL} }
```

Ejercicio 33.8.3. *Dado el esquema de traducción:*

```
e →  NUM r   { $e{TRA} = $NUM{VAL}." ".$r{TRA} }
r → −e       { $r{TRA} = $e{TRA}." - " }
r → ϵ        { $r{TRA} = "" }
```

¿Cuál es el lenguaje generado por la gramática? ¿Puede el lenguaje ser analizado por un APDR? ¿Cual es la traducción de 4-5-6? ¿Es un esquema de traducción adecuado para traducir de infijo a postfijo? ¿Cuál es la traducción si cambiamos el anterior esquema por este otro?:

```
e → NUM r   { $e{TRA} = $NUM{VAL}." ".$r{TRA} }
r → −e      { $r{TRA} = " - ".$e{TRA} }
r → ϵ       { $r{TRA} = "" }
```

33.8.2. Eliminación de la Recursión por la Izquierda en un Esquema de Traducción

La eliminación de la recursión por la izquierda es sólo un paso: debe ser extendida a esquemas de traducción, de manera que no sólo se preserve el lenguaje sino la secuencia de acciones. Supongamos que tenemos un esquema de traducción de la forma:

```
A → Aα   { alpha_action }
A → Aβ   { beta_action }
A → γ    { gamma_action }
```

para una sentencia como $\gamma\beta\alpha$ la secuencia de acciones será:

```
gamma_action  beta_action alpha_action
```

¿Cómo construir un esquema de traducción para la gramática resultante de eliminar la recursión por la izquierda que ejecute las acciones asociadas en el mismo orden?. Supongamos para simplificar, que las acciones no dependen de atributos ni computan atributos, sino que actúan sobre variables globales. En tal caso, la siguiente ubicación de las acciones da lugar a que se ejecuten en el mismo orden:

```
A → γ { gamma_action } R
R → β  { beta_action } R
R → α  { alpha_action } R
R → ϵ
```

Si hay atributos en juego, la estrategia para construir un esquema de traducción equivalente para la gramática resultante de eliminar la recursividad por la izquierda se complica. Consideremos de nuevo el esquema de traducción de infijo a postfijo de expresiones aritméticas de restas:

$expr \rightarrow expr_1 - NUM$ `{ $expr{T} = $expr[1]{T}." ".$NUM{VAL}." - "}`
$expr \rightarrow NUM$ `{ $expr{T} = $NUM{VAL} }`

En este caso introducimos un atributo `H` para los nodos de la clase r el cuál acumula la traducción a postfijo hasta el momento. Observe como este atributo se computa en un nodo r a partir del correspondiente atributo del el padre y/o de los hermanos del nodo:

$expr \rightarrow NUM$ `{ $r{H} = $NUM{VAL} }` r `{ $expr{T} = $r{T} }`
$r \rightarrow -NUM$ `{ $r_1{H} = $r{H}." ".$NUM{VAL}." - " }` r_1 `{ $r{T} = $r_1{T} }`
$r \rightarrow \epsilon$ `{ $r{T} = $r{H} }`

El atributo `H` es un ejemplo de atributo heredado.

33.8.3. Ejercicio

Calcule los valores de los atributos cuando se aplica el esquema de traducción anterior a la frase `4 - 5 - 7`.

33.8.4. Convirtiendo el Esquema en un Analizador Predictivo

A partir del esquema propuesto, que se basa en una fase de descenso con un atributo heredado y una de ascenso con un atributo sintetizado:

$expr \rightarrow NUM$ `{ $r{H} = $NUM{VAL} }` r `{ $expr{T} = $r{T} }`
$r \rightarrow -NUM$ `{ $r_1{H} = $r{H}." ".$NUM{VAL}." - " }` r_1 `{ $r{T} = $r_1{T} }`
$r \rightarrow \epsilon$ `{ $r{T} = $r{H} }`

es posible construir un APDR que ejecuta las acciones semánticas en los puntos indicados por el esquema de traducción. El atributo heredado se convierte en un parámetro de entrada a la subrutina asociada con la variable sintáctica:

```
sub expression() {
  my $r = $value." "; #accion intermedia
  match('NUM');
  return rest($r); # accion final $expr{T} = $r{T}
}

sub rest($) {
  my $v = shift;

  if ($lookahead eq '-') {
    match('-');
    my $r = "$v $value -"; # accion intermedia
    match('NUM');
    return rest($r); # accion final $r{t} = $r_1{t}
  }
  elsif ($lookahead ne 'EOI') {
    error("Se esperaba un operador");
  }
  else { return $v; } # r -> epsilon { $r{t} = $r{h} }
```

```
}
```

33.8.5. Ejercicio

Generalize la estrategia anterior para eliminar la recursividad por la izquierda al siguiente esquema de traducción genérico recursivo por la izquierda y con un atributo sintetizado A^s:

$$
\begin{array}{ll}
A \to A_1 X_1 X_2 X_3 & \{A^s = f_X(A_1^s, X_1^s, X_2^s, X_3^s)\} \\
A \to A_1 Y_1 Y_2 Y_3 & \{A^s = f_Y(A_1^s, Y_1^s, Y_2^s, Y_3^s)\} \\
A \to Z_1 Z_2 Z_3 & \{A^s = f_Z(Z_1^s, Z_2^s, Z_3^s)\}
\end{array}
$$

donde f_X, f_Y y f_Z son funciones cualesquiera.

33.8.6. Práctica: Eliminación de la Recursividad por la Izquierda

En esta práctica vamos a extender las fases de análisis léxico y sintáctico del compilador del lenguaje Tutu cuya gramática se definió en el ejercicio 34.1.1 con expresiones que incluyen diferencias y divisiones. Además construiremos una representación del árbol sintáctico concreto. Para ello consideremos el siguiente esquema de traducción recursivo por la izquierda (en concreto las reglas recursivas por la izquierda son las 10, 11, 13 y 14):

0	p → ds ss	`{ $p{t} = { n => 0, ds => $ds{t}, ss => $ss{t} } }`
1	p → ss	`{ $p{t} = { n => 1, ss => $ss{t} } }`
2	ds → d ';' ds	`{ $ds{t} = { n => 2, d => $d{t}, ; => ';', ds => $ds{t} } }`
3	ds → d ';'	`{ $ds{t} = { n => 3, d => $d{t} ; => ';' } }`
4	d → INT il	`{ $d{t} = { n => 4, INT => 'INT', il >$il{t} } }`
5	d → STRING il	`{ $d{t} = { n => 5, STRING => 'STRING', il >$il{t} } }`
6	ss → s ';' ss	`{ $ss{t} = { n => 6, s => $s{t}, ; => ';' ss => $ss{t} } }`
7	ss → s	`{ $ss{t} = { n => 7, s => $s{t} } }`
8	s → ID '=' e	`{ $s{t} = { n => 8, ID => $ID{v}, = => '=', e => $e{t} } }`
9	s → P e	`{ $s{t} = { n => 9, P => 'P', e => $e{t} } }`
10	e → e1 '+' t	`{ $e{t} = { n => 10, e => $e1{t}, + => '+', t => $t{t} } }`
11	e → e1 '-' t	`{ $e{t} = { n => 11, e => $e1{t}, - => '-', t => $t{t} } }`
12	e → t	`{ $e{t} = { n => 12, t => $t{t} } }`
13	t → t1 '*' f	`{ $t{t} = { n => 13, t => $t1{t}, * => '*', f => $f{t} } }`
14	t → t '/' f	`{ $t{t} = { n => 14, t => $t1{t}, / => '/', f => $f{t} } }`
15	t → f	`{ $t{t} = { n => 15, f => $f{t} } }`
16	f → '(' e ')'	`{ $f{t} = { n => 16, ( => '(', e => $e{t}, ) => ')' } }`
17	f → ID	`{ $f{t} = { n => 17, ID => $ID{v} } }`
18	f → NUM	`{ $f{t} = { n => 18, NUM => $NUM{v} } }`
19	f → STR	`{ $f{t} = { n => 19, STR => $STR{v} } }`
20	il → ID ',' il	`{ $il{t} = { n => 20, ID => $ID{v}, ',' => ',', il => $il{t} } }`
21	il → ID	`{ $il{t} = { n => 21, ID => $ID{v} } }`
22	s → ϵ	`{ $s{t} = { n => 22, s => '' }}`

Por razones de espacio hemos abreviado los nombres de las variables. El atributo `t` (por *tree*) es una referencia a un hash. La entrada `n` contiene el número de la regla en juego. Hay una entrada por símbolo en la parte derecha. El atributo `v` de `ID` es la cadena asociada con el identificador. El atributo `v` de `NUM` es el valor numérico asociado con el terminal. Se trata de, siguiendo la metodología explicada en la sección anterior, construir un analizador descendente predictivo recursivo que sea equivalente al esquema anterior. Elimine la recursión por la izquierda. Traslade las acciones a los lugares convenientes en el nuevo esquema e introduzca los atributos heredados que sean necesarios. Genere pruebas siguiendo la metodología explicada en la sección 33.4.1. ¡Note que el árbol que debe producir es el de la gramática inicial, ¡No el de la gramática transformada!

33.9. Árbol de Análisis Abstracto

33.9.1. Lenguajes Árbol y Gramáticas Árbol

Un *árbol de análisis abstracto* (denotado AAA, en inglés *abstract syntax tree* o AST) porta la misma información que el árbol de análisis sintáctico pero de forma mas condensada, eliminándose terminales y producciones que no aportan información.

Definición 33.9.1. *Un* alfabeto con función de aridad *es un par* (Σ, ρ) *donde* Σ *es un conjunto finito y* ρ *es una función* $\rho : \Sigma \to \mathbb{N}_0$*, denominada* función de aridad*. Denotamos por* $\Sigma_k = \{a \in \Sigma : \ \rho(a) = k\}$*.*

Definimos el lenguaje árbol homogéneo $B(\Sigma)$ *sobre* Σ *inductivamente:*

- *Todos los elementos de aridad 0 están en* $B(\Sigma)$*:* $a \in \Sigma_0$ *implica* $a \in B(\Sigma)$
- *Si* $b_1, \ldots, b_k \in B(\Sigma)$ *y* $f \in \Sigma_k$ *es un elemento k-ario, entonces* $f(b_1, \ldots, b_k) \in B(\Sigma)$

Los elementos de $B(\Sigma)$ *se llaman* árboles *o* términos*.*

Ejemplo 33.9.1. *Sea* $\Sigma = \{A, CONS, NIL\}$ *con* $\rho(A) = \rho(NIL) = 0, \rho(CONS) = 2$*. Entonces*
$B(\Sigma) = \{A, NIL, CONS(A, NIL), CONS(NIL, A), CONS(A, A), CONS(NIL, NIL), \ldots\}$

Ejemplo 33.9.2. *Una versión simplificada del alfabeto con aridad en el que estan basados los árboles construidos por el compilador de Tutu es:*
$\Sigma = \{ID, NUM, LEFTVALUE, STR, PLUS, TIMES, ASSIGN, PRINT\}$
$\rho(ID) = \rho(NUM) = \rho(LEFTVALUE) = \rho(STR) = 0$
$\rho(PRINT) = 1$
$\rho(PLUS) = \rho(TIMES) = \rho(ASSIGN) = 2.$
Observe que los elementos en $B(\Sigma)$ *no necesariamente son árboles "correctos". Por ejemplo, el árbol* $ASSIGN(NUM, PRINT(ID))$ *es un elemento de* $B(\Sigma)$*.*

Definición 33.9.2. *Una* gramática árbol regular *es una cuadrupla* $((\Sigma, \rho), N, P, S)$*, donde:*

- (Σ, ρ) *es un alfabeto con aricidad* $\rho : \Sigma \to \mathbb{N}$
- N *es un conjunto finito de variables sintácticas o no terminales*
- P *es un conjunto finito de reglas de producción de la forma* $X \to s$ *con* $X \in N$ *y* $s \in B(\Sigma \cup N)$
- $S \in N$ *es la variable o símbolo de arranque*

Definición 33.9.3. *Dada una gramática* (Σ, N, P, S)*, se dice que un árbol* $t \in B(\Sigma \cup N)$ *es del tipo* $(X_1, \ldots X_k)$ *si el* j*-ésimo noterminal, contando desde la izquierda, que aparece en* t *es* $X_j \in N$*.*

Si $p = X \to s$ *es una producción y* s *es de tipo* $(X_1, \ldots X_n)$*, diremos que la producción* p *es de tipo* $(X_1, \ldots X_n) \to X$*.*

Definición 33.9.4. *Consideremos un árbol* $t \in B(\Sigma \cup N)$ *que sea del tipo* $(X_1, \ldots X_n)$*, esto es las variables sintácticas en el árbol leídas de izquierda a derecha son* $(X_1, \ldots X_n)$*.*

- *Si* $X_i \to s_i \in P$ *para algún* i*, entonces decimos que el árbol* t deriva en un paso en el árbol t' *resultante de sustituir el nodo* X_i *por el árbol* s_i *y escribiremos* $t \Longrightarrow t'$*. Esto es,* $t' = t\{X_i/s_i\}$
- *Todo árbol deriva en cero pasos en si mismo* $t \overset{0}{\Longrightarrow} t$*.*
- *Decimos que un árbol* t *deriva en* n *pasos en el árbol* t' *y escribimos* $t \overset{n}{\Longrightarrow} t'$ *si* t *deriva en un paso en un árbol* t'' *el cuál deriva en* $n-1$ *pasos en* t'*. En general, si* t *deriva en un cierto número de pasos en* t' *escribiremos* $t \overset{*}{\Longrightarrow} t'$*.*

Definición 33.9.5. *Se define el* lenguaje árbol generado por una gramática $G = (\Sigma, N, P, S)$ *como el lenguaje* $L(G) = \{t \in B(\Sigma) : \exists S \overset{*}{\Longrightarrow} t\}$*.*

Ejemplo 33.9.3. *Sea* $G = (\Sigma, V, P, S)$ *con* $\Sigma = \{A, CONS, NIL\}$ *y* $\rho(A) = \rho(NIL) = 0, \rho(CONS) = 2$ *y sea* $V = \{E, L\}$. *El conjunto de producciones* P *es:*

$$P_1 = \{L \to NIL, L \to CONS(E, L), E \to a\}$$

La producción $L \to CONS(E, L)$ *es del tipo* $(E, L) \to L$.

Informalmente, el lenguaje generado por G *se obtiene realizando sustituciones sucesivas (derivando) desde el símbolo de arranque hasta producir un árbol cuyos nodos estén etiquetados con elementos de* Σ. *Debería ser claro que, en este ejemplo,* $L(G)$ *es el conjunto de las listas en* A, *incluyendo la lista vacía:*

$$L(G) = \{NIL, CONS(A, NIL), CONS(A, CONS(A, NIL)), \ldots\}$$

Ejercicio 33.9.1. *Construya una derivación para el árbol* $CONS(A, CONS(A, NIL))$. *¿De que tipo es el árbol* $CONS(E, CONS(A, CONS(E, L)))$*?.*

Cuando hablamos del AAA producido por un analizador sintáctico, estamos en realidad hablando de un lenguaje árbol cuya definición precisa debe hacerse a través de una gramática árbol regular. Mediante las gramáticas árbol regulares disponemos de un mecanismo para describir formalmente el lenguaje de los AAA que producirá el analizador sintáctico para las sentencias Tutu.

Ejemplo 33.9.4. *Sea* $G = (\Sigma, V, P, S)$ *con*

$$\begin{aligned}
&\Sigma = \{ID, NUM, LEFTVALUE, STR, PLUS, TIMES, ASSIGN, PRINT\}\\
&\rho(ID) = \rho(NUM) = \rho(LEFTVALUE) = \rho(STR) = 0\\
&\rho(PRINT) = 1\\
&\rho(PLUS) = \rho(TIMES) = \rho(ASSIGN) = 2\\
&V = \{st, expr\}
\end{aligned}$$

y las producciones:

$$\begin{aligned}
P = \quad & \{\\
& st \quad \to ASSIGN(LEFTVALUE, expr)\\
& st \quad \to PRINT(expr)\\
& expr \to PLUS(expr, expr)\\
& expr \to TIMES(expr, expr)\\
& expr \to NUM\\
& expr \to ID\\
& expr \to STR\\
& \}
\end{aligned}$$

Entonces el lenguaje $L(G)$ *contiene árboles como el siguiente:*

$$\begin{array}{lllll}
ASSIGN & (& & & \\
 & LEFTVALUE, & & & \\
 & PLUS & (& & \\
 & & ID, & & \\
 & & TIMES & (& \\
 & & & NUM, & \\
 & & & ID & \\
 & & &) & \\
 & &) & & \\
 &) & & &
\end{array}$$

El cual podría corresponderse con una sentencia como `a = b + 4 * c`.

El lenguaje de árboles descrito por esta gramática árbol es el lenguaje de los AAA de las sentencias de Tutu.

Ejercicio 33.9.2. *Redefina el concepto de árbol de análisis concreto dado en la definición 34.1.7 utilizando el concepto de gramática árbol. Con mas precisión, dada una gramática $G = (\Sigma, V, P, S)$ defina una gramática árbol $T = (\Omega, N, R, U)$ tal que $L(T)$ sea el lenguaje de los árboles concretos de G. Puesto que las partes derechas de las reglas de producción de P pueden ser de distinta longitud, existe un problema con la aricidad de los elementos de Ω. Discuta posibles soluciones.*

Ejercicio 33.9.3. *¿Cómo son los árboles sintácticos en las derivaciones árbol? Dibuje varios árboles sintácticos para las gramáticas introducidas en los ejemplos 33.9.3 y 33.9.4.*

Intente dar una definición formal del concepto de árbol de análisis sintáctico asociado con una derivación en una gramática árbol

Definición 33.9.6. *La notación de Dewey es una forma de especificar los subárboles de un árbol $t \in B(\Sigma)$. La notación sigue el mismo esquema que la numeración de secciones en un texto: es una palabra formada por números separados por puntos. Así $t/2.1.3$ denota al tercer hijo del primer hijo del segundo hijo del árbol t. La definición formal sería:*

- $t/\epsilon = t$
- *Si $t = a(t_1, \ldots t_k)$ y $j \in \{1 \ldots k\}$ y n es una cadena de números y puntos, se define inductivamente el subárbol $t/j.n$ como el subárbol n-ésimo del j-ésimo subárbol de t. Esto es: $t/j.n = t_j/n$*

Ejercicio 33.9.4. *Sea el árbol:*

$$
\begin{array}{lllll}
t = ASSIGN & (& & & \\
 & LEFTVALUE, & & & \\
 & PLUS & (& & \\
 & & ID, & & \\
 & & TIMES & (& \\
 & & & NUM, & \\
 & & & ID & \\
 & & &) & \\
 & &) & & \\
 &) & & &
\end{array}
$$

Calcule los subárboles t/ϵ, $t/2.2.1$, $t/2.1$ y $t/2.1.2$.

33.9.2. Realización del AAA para Tutu en Perl

En la sección 33.6.1 nos limitamos a realizar un recorrido del árbol de análisis sintáctico concreto. En esta sección construimos un *árbol de análisis sintáctico abstracto*. Este proceso puede verse como *la traducción desde el lenguaje de árboles concretos hasta el lenguaje de árboles abstractos.*

Definición 33.9.7. *La gramática árbol extendida que especifica los árboles AAA para el compilador de Tutu es esta:*

1	$prog$	$\rightarrow PROGRAM(decls, sts)$
2	$decls$	$\rightarrow$ *list decl*
3	sts	$\rightarrow$ *list st*
4	$decl$	$\rightarrow INT(idlist)$
5	$decl$	$\rightarrow STRING(idlist)$
6	$idlist$	$\rightarrow$ *list* $SIMPLEID$
7	st	$\rightarrow ASSIGN(LEFTVALUE, expr)$
8	st	$\rightarrow PRINT(expr)$
9	$expr$	$\rightarrow PLUS(expr, expr)$
10	$expr$	$\rightarrow TIMES(expr, expr)$
11	$expr$	$\rightarrow NUM$
12	$expr$	$\rightarrow ID$
13	$expr$	$\rightarrow STR$

Hemos extendido el concepto de gramática árbol con el concepto de *lista de no terminales*. A la hora de construir las estructuras de datos las listas de variables se van a traducir por listas de árboles.

Por ejemplo, un árbol abstracto para el programa

```
int a,b;
string c, d;
a = 4;
p a
```

Sería de la forma:

```
PROGRAM(
         DECLS[INT[ID, ID], STRING[ID, ID]],
         STS[ASSIGN(LEFTVALUE, NUM), PRINT(ID)]
       )
```

Donde los corchetes indican listas y los paréntesis tuplas.

Para llevar a cabo la traducción deberemos tomar decisiones sobre que forma de representación nos conviene. Cada nodo del AAA va a ser un objeto y la clase indicará si es un nodo suma, producto, una declaración, una asignación, etc.

Cada nodo del árbol AAA va a ser un objeto. De este modo el acceso a los atributos del nodo se hará a través de los métodos asociados. Además, el procedimiento de traducción al lenguaje objetivo depende del tipo de nodo. Así por ejemplo, el método *traducción* es diferente para un nodo de tipo $PLUS$ que para otro de tipo $ASSIGN$.

Resumamos antes de entrar en detalle, la forma de manejar los objetos en Perl:

- Para crear una *clase* se construye un "package":

  ```
  package NUM;
  ```

- Para crea un *método* se escribe una subrutina:

  ```
  package NUM;
  sub incr { my $self = shift; $self->{VAL}++ }
  ```

  el primer argumento de un método suele ser la referencia al objeto en cuestión.

- Para crear un *objeto*, se bendice ("bless") una referencia. Los objetos Perl son datos normales como hashes y arrays que han sido "bendecidos" en un paquete. Por ejemplo:

  ```
  my $a = bless {VAL => 4}, 'NUM';
  ```

  crea un objeto referenciado por `$a` que pertenece a la clase `NUM`. Los métodos del objeto son las subrutinas que aparecen en el `package NUM`.

- Para referirse a un método de un objeto se usa la sintáxis "flecha":

  ```
  $a->incr;
  ```

  Cuando se usa la sintáxis flecha, el primer argumento de la rutina es la referencia al objeto, esto es, la llamada anterior es equivalente a `NUM::incr($a)`

- Constructores: En Perl son rutinas que retornan una referencia a un objeto recién creado e inicializado

  ```
  sub new { my ($class, $value) = @_; return bless {VAL => $value}, $class; }
  ```

Normalmente se llaman usando la sintáxis flecha, pero a la izquierda de la flecha va el nombre de la clase. Por ejemplo:

```
my $a = NUM->new(4)
```

En este caso, el primer argumento es el nombre de la clase. La llamada anterior es equivalente a

```
NUM::new('NUM', 4)
```

Volviendo a nuestro problema de crear el AAA, para crear los objetos de las diferentes clases de nodos usaremos el módulo `Class::MakeMethods::Emulator::MethodMaker` (véase la línea 9):

```
 1 package PL::Syntax::Analysis;
 2
 3 use 5.006;
 4 use strict;
 5 use warnings;
 6 use Data::Dumper;
 7 use IO::File;
 8 use Class::MakeMethods::Emulator::MethodMaker '-sugar';
 9
10 require Exporter;
11 our @ISA = qw(Exporter);
12 our @EXPORT = qw( );
13 our $VERSION = '0.02';
14
15 ##########################################################
16
17 # Grammar:
18 # P : DD L      | L
19 # DD: D ';' DD  | D ';'
20 # D : int I     | string I
21 # L : S         | S ; L
22 # S : ID '=' E  | p E  | epsilon
23 # E : T '+' E   | T
24 # T : F '*' T   | F
25 # F : '(' E ')' | id | num | str
26 # I : id ',' I  | id
```

Hemos aislado la fase de análisis sintáctica en un módulo aparte denominado `PL::Syntax::Analysis`. La dependencia se actualiza en `Makefile.PL`:

```
PL0506/04sintactico/PL-Tutu$ cat -n Makefile.PL
  1  use 5.008004;
  2  use ExtUtils::MakeMaker;
  3  WriteMakefile(
  4      NAME              => 'PL::Tutu',
  5      VERSION_FROM      => 'lib/PL/Tutu.pm', # finds $VERSION
  6      PREREQ_PM         => {Class::MakeMethods::Emulator::MethodMaker => 0},1
  7      EXE_FILES         => [ 'scripts/tutu.pl', 'scripts/tutu' ],
  8      ($] >= 5.005 ?     ## Add these new keywords supported since 5.005
  9        (ABSTRACT_FROM  => 'lib/PL/Tutu.pm', # retrieve abstract from module
 10         AUTHOR         => 'Casiano Rodriguez Leon <casiano@ull.es>') : ()),
 11  );
```

Se actualiza también MANIFEST:

```
$ cat -n MANIFEST
 1  Changes
 2  lib/PL/Error.pm
 3  lib/PL/Lexical/Analysis.pm
 4  lib/PL/Syntax/Analysis.pm
 5  lib/PL/Tutu.pm
 6  Makefile.PL
 7  MANIFEST
 8  MANIFEST.SKIP
 9  README
10  scripts/test01.tutu
11  scripts/tutu
12  scripts/tutu.pl
13  t/01Lexical.t
```

Ahora `compile` llama a `Syntax::Analysis::parser` pasándole como argumento la lista de terminales `@tokens`. La función `parser` devuelve el AAA:

```
04sintactico/PL-Tutu/lib/PL$ sed -ne '/sub compile\>/,/^}/p' Tutu.pm | cat -n
  1  sub compile {
  2    my ($input) = @_;
  3    #my %symbol_table = ();
  4    #my $data = ""; # Contiene todas las cadenas en el programa fuente
  5    my $target = ""; # target code
  6    my @tokens = ();
  7    my $tree = undef; # abstract syntax tree
  8    #my $global_address = 0;
  9
 10
 11    ########lexical analysis
 12    @tokens = scanner($input);
 13    #print "@tokens\n";
 14
 15    ########syntax (and semantic) analysis
 16    $tree = &Syntax::Analysis::parser(@tokens);
 17    print Dumper($tree);
 18
 19    ########machine independent optimizations
 20    &Machine::Independent::Optimization::Optimize;
 21
 22    ########code generation
 23    &Code::Generation::code_generator;
 24
 25    ########peephole optimization
 26    &Peephole::Optimization::transform($target);
 27
 28    return \$target;
 29  }
```

El módulo `Class::MakeMethods::Emulator::MethodMaker` permite crear constructores y métodos de acceso. El módulo no viene con la distribución de Perl, así que, en general, deberá descargarlo desde CPAN e instalarlo. Así definimos que existe una clase de nodos `TYPE` que nuestro AAA va a tener:

```
package TYPE;
```

```
make methods
  get_set       => [ 'NAME', 'LENGTH' ],
  new_hash_init => 'new';
```

El uso de los argumentos `get_set => [ 'NAME', 'LENGTH' ]` hace que se cree un objeto de tipo hash con claves `'NAME'` y `'LENGTH'` así como métodos `NAME` y `LENGTH` que cuando se llaman con un argumento devuelven el valor y cuando se llaman con dos argumentos modifican el valor correspondiente. La clave `get_set` produce métodos de acceso y modificación de los atributos del objeto que tienen la forma:

```
sub TYPE::NAME {
  my ($self, $new) = @_;
  defined($new) and $self->{NAME} = $new;
  return $self->{NAME};
}
```

Asi mismo el uso de `new_hash_init => 'new'` genera un constructor cuyo nombre será `'new'` y que cuando es llamado inicializará el objeto con los argumentos con nombre especificados en la llamada. El constructor construído (vaya retruécano) cuando se usa la clave `new_hash_init` tiene el siguiente aspecto:

```
sub TYPE::new {
  my ($class, %args) = @_;
  my $self = {};

  bless $self, $class;
  foreach my $attribute (keys %args) {
    $self->$attribute($args{$attribute});
  }
  return $self;
}
```

Ahora podemos crear objetos de la clase `TYPE` haciendo:

```
my $int_type = TYPE->new(NAME => 'INTEGER', LENGTH => 1);
my $string_type = TYPE->new(NAME => 'STRING', LENGTH => 2);
my $err_type = TYPE->new(NAME => 'ERROR', LENGTH => 0);
```

Cada uno de estos objetos es un hash con las correspondientes claves para el nombre y el tipo.

Otros tipos de nodos del AAA son:

```
package PROGRAM; # raíz del AAA
make methods
  get_set       => [ 'DECLS', 'STS' ],
  new_hash_init => 'new';

package STRING; # tipo
make methods
  get_set       => [ 'TYPE', 'IDLIST' ],
  new_hash_init => 'new';

package INT; # tipo
make methods
  get_set       => [ 'TYPE', 'IDLIST' ],
  new_hash_init => 'new';
```

```
package ASSIGN; #sentencia
make methods
  get_set       => [ 'LEFT', 'RIGHT' ],
  new_hash_init => 'new';

package PRINT; #sentencia
make methods
  get_set       => [ 'EXPRESSION' ],
  new_hash_init => 'new';

package NUM; # para los números
make methods
  get_set       => [ 'VAL', 'TYPE' ],
  new_hash_init => 'new';

package ID; # Nodos identificador. Parte derecha
make methods
  get_set       => [ 'VAL', 'TYPE' ],
  new_hash_init => 'new';

package STR; # Clase para las constantes cadena
make methods
  get_set       => [ 'OFFSET', 'LENGTH', 'TYPE' ],
  new_hash_init => 'new';

package PLUS; # Nodo suma
make methods
  get_set       => [ 'LEFT', 'RIGHT', 'TYPE' ],
  new_hash_init => 'new';

package TIMES;
make methods
  get_set       => [ 'LEFT', 'RIGHT', 'TYPE' ],
  new_hash_init => 'new';

package LEFTVALUE; # Identificador en la parte izquierda
make methods       # de una asignación
  get_set       => [ 'VAL', 'TYPE' ],
  new_hash_init => 'new';
```

Hemos extendido el concepto de gramática árbol con el concepto de *lista de no terminales*. A la hora de construir las estructuras de datos las listas de variables se van a traducir por listas de árboles. Los tipos de nodos ($ASSIGN$, $PRINT$, ...) se traducen en nombres de clases. Hemos hecho una excepción con $SIMPLEID$ el cual es simplemente una variable cadena conteniendo el identificador correspondiente.

El siguiente esquema de traducción resume la idea para una gramática simplificada: cada vez que encontremos un nodo en el árbol sintáctico concreto con una operación crearemos un nodo en el AAA cuya clase viene definida por el tipo de operación. Para los terminales creamos igualmente nodos indicando de que clase de terminal se trata. El atributo nodo lo denotaremos por `n`:

$e \to e_1 + f$ `{ $e{n} = PLUS->new(LEFT=>$e_1{n}, RIGHT=>$f{n}) }`
$f \to NUM$ `{ $f{n} = NUM->new(VAL => $NUM{VAL}) }`
$f \to ID$ `{ $f{n} = ID->new(VAL => $ID{VAL}) }`

La estructura de cada rutina sigue siendo la misma, sólo que ampliada con las acciones para la construcción de los correspondientes nodos. Veamos por ejemplo, como modificamos la subrutina factor:

```
sub factor() {
  my ($e, $id, $str, $num);

  if ($lookahead eq 'NUM') {
    $num = $value;
    match('NUM');
    return NUM->new(VAL => $num, TYPE => $int_type);
  }
  elsif ($lookahead eq 'ID') {
    $id = $value;
    match('ID');
    return ID->new( VAL => $id, TYPE => undef);
  }
  elsif ($lookahead eq 'STR') {
    $str = $value;
    match('STR');
    return STR->new(OFFSET => undef, LENGTH => undef, TYPE => $string_type);
  }
  elsif ($lookahead eq '(') {
    match('(');
    $e = expression;
    match(')');
    return $e;
  }
  else {
    Error::fatal("Se esperaba (, NUM o ID");
  }
}
```

33.9.3. AAA: Otros tipos de nodos

Hemos optado por que las rutinas asociadas a variables sintácticas que describen listas de subcategorías devuelvan las correspondientes listas de nodos. Teníamos tres variables tipo lista. Las reglas para las listas eran:

Gramática de los Árboles de Tutu	Gramática del lenguaje Tutu
decls → *list decl*	declarations → declaration ';' declarations | declaration ';'
sts → *list st*	statements → statement ';' statements | statement
idlist → *list SIMPLEID*	idlist → ID ',' idlist | ID

En este caso las subrutinas asociadas no devuelven objetos sino listas de objetos. Esto da lugar a una compactación del AAA. Veánse los códigos de `statements` y `idlist`:

```
sub statements() {
  my @s;

  @s = (statement());
  if ($lookahead eq ';') {
    match(';');
    push @s, statements();
```

```
  }
  return @s;
}

sub idlist() {
  my @id;

  if ($lookahead eq 'ID') {
    @id = ($value); # no es un objeto
    match('ID');
    if ($lookahead eq ',') {
      match(',');
      push @id, idlist();
    }
  }
  else {
    Error::fatal('Se esperaba un identificador');
    @id = ('ERROR');
  }
  return @id;
}
```

33.9.4. Declaraciones

Los nodos del tipo declaration no existen propiamente, son nodos de la clase INT o de la clase $STRING$. La parte de la gramática árbol de la que hablamos es:

2	*decls*	→ list *decl*
4	*decl*	→ $INT(idlist)$
5	*decl*	→ $STRING(idlist)$
6	*idlist*	→ list $SIMPLEID$

Los nodos declaration son un hash con una clave `TYPE` la cual apunta a la estructura de datos/objeto describiendo el tipo. La otra clave del hash `IDLIST` apunta a una lista de identificadores. Los elementos de esta lista son simples identificadores (identificados en la gramática árbol anterior como $SIMPLEID$ y no como objetos ID). La parte de la gramática implicada en las declaraciones es:

declaration → INT idlist | STRING idlist
idlist → ID ',' idlist | ID

Así pues, el código construye un nodo de la clase `INT` o `STRING` según sea el caso.

```
sub declaration() {
  my ($t, $class, @il);

  if (($lookahead eq 'INT') or ($lookahead eq 'STRING')) {
    $class = $lookahead;
    $t = &type();
    @il = &idlist();
    return $class->new(TYPE => $t, IDLIST => \@il);
  }
  else {
    Error::fatal('Se esperaba un tipo');
  }
}
```

Observe la llamada `$class->new(TYPE => $t, IDLIST => \@il)` en la cual la clase se usa a través de una referencia simbólica.

33.9.5. Práctica: Arbol de Análisis Abstracto

Complete la fase de análisis sintáctico para la gramática de Tutu extendida con sentencias de bloque (vea las reglas 1,2,3 y 11) construyendo el AAA según el lenguaje árbol especificado por una gramática árbol que extienda la dada en la definición 33.9.7. Genere pruebas, usando `make test` para comprobar el correcto funcionamiento de su analizador sobre las mismas. Utilize el módulo `Data::Dumper` para volcar las estructuras de datos resultantes.

1	p → b
2	b → ds ss
3	b → ss
4	ds → d ';' ds
5	ds → d ';'
6	d → INT il
7	d → STRING il
8	ss → s ';' ss
9	ss → s
10	s → ID = e
11	s → '{' b '}'
12	s → P e
13	s → ϵ
14	e → e1 '+' t
15	e → e1 '-' t
16	e → t
17	t → t1 '*' f
18	t → t '/' f
19	t → f
20	f → '(' e ')'
21	f → ID
22	f → NUM
23	f → STR
24	il → ID ',' il
25	il → ID

33.10. Análisis Semántico

Hay quien dice que el análisis semántico es la determinación de aquellas propiedades que, siendo dependientes del contexto, pueden ser computadas estáticamente en tiempo de compilación para cualquier programa correcto. Entre estas propiedades están: la comprobación de que las variables son declaradas, la compatibilidad de tipos en las expresiones, el correcto uso de las llamadas a función asi como el ámbito y visibilidad de las variables. La fase de análisis semántico puede verse como una fase de "adornado" o "etiquetado" del AAA, en la cual los atributos de los nodos del AAA son computados.

Aunque la veamos como una fase separada del análisis sintáctico, puede en numerosas ocasiones llevarse a cabo al mismo tiempo que se construye el árbol. Así lo hacemos en este ejemplo: incrustamos la acción semántica en la correspondiente rutina de análisis sintáctico. Así, en la rutina `term`, una vez que hemos obtenido los dos operandos, comprobamos que son de tipo numérico llamando (línea 8) a

`Semantic::Analysis::check_type_numeric_operator`:

Observe como aparece un nuevo atributo `TYPE` decorando el nodo creado (línea 9):

```
 1 sub term() {
 2   my ($t, $t2);
 3
 4   $t = factor;
 5   if ($lookahead eq '*') {
 6     match('*');
 7     $t2 = term;
 8     my $type = Semantic::Analysis::check_type_numeric_operator($t, $t2, '*');
 9     $t = TIMES->new( LEFT => $t, RIGHT => $t2, TYPE => $type);
10   }
11   return $t;
12 }
```

En el manejo de errores de tipo, un tipo especial `$err_type` es usado para indicar un error de tipo:

```
sub check_type_numeric_operator {
  my ($op1, $op2, $operator) = @_;

  my $type = numeric_compatibility($op1, $op2, $operator);
  if ($type == $err_type) {
    Error::fatal("Operandos incompatibles para el operador $operator")
  }
  else {
    return $type;
  }
}
```

La subrutina `numeric_compatibility` comprueba que los dos operandos son de tipo numérico y devuelve el correspondiente tipo. Si ha ocurrido un error de tipo, intenta encontrar un tipo conveniente para el operando:

```
sub numeric_compatibility {
  my ($op1, $op2, $operator) = @_;

  if (($op1->TYPE == $op2->TYPE) and is_numeric($op1->TYPE)) {
    return $op1->TYPE; # correct
  }
  ... # código de recuperación de errores de tipo
}

sub is_numeric {
  my $type = shift;

  return ($type == $int_type); # añadir long, float, double, etc.
}
```

Es parte del análisis semántico la declaración de tipos:

```
sub declaration() {
  my ($t, $class, @il);

  if (($lookahead eq 'INT') or ($lookahead eq 'STRING')) {
    $class = $lookahead;
    $t = &type();
    @il = &idlist();
```

```
    &Semantic::Analysis::set_types($t, @il);
    &Address::Assignment::compute_address($t, @il);
    return $class->new(TYPE => $t, IDLIST => \@il);
  }
  else {
    Error::fatal('Se esperaba un tipo');
  }
}
```

Para ello se utiliza una tabla de símbolos que es un hash `%symbol_table` indexado en los identificadores del programa:

```
sub set_types {
  my $type = shift;
  my @vars = @_;

  foreach my $var (@vars) {
    if (!exists($symbol_table{$id})) { $symbol_table{$var}->{TYPE} = $type; }
    else { Error::error("$id declarado dos veces en el mismo ámbito"); }
  }
}
```

Cada vez que aparece una variable en el código, bien en un factor o en una asignación, comprobamos que ha sido declarada:

```
sub factor() {
  my ($e, $id, $str, $num);

  if ($lookahead eq 'NUM') { ... }
  elsif ($lookahead eq 'ID') {
    $id = $value;
    match('ID');
    my $type = Semantic::Analysis::check_declared($id);
    return ID->new( VAL => $id, TYPE => $type);
  }
  elsif ($lookahead eq 'STR') { ... }
  elsif ($lookahead eq '(') { ... }
  else { Error::fatal("Se esperaba (, NUM o ID"); }
}
```

La función `check_declared` devuelve el atributo `TYPE` de la correspondiente entrada en la tabla de símbolos.

```
sub check_declared {
  my $id = shift;

  if (!exists($symbol_table{$id})) {
    Error::error("$id no ha sido declarado!");
    # auto-declaración de la variable a err_type
    Semantic::Analysis::set_types($err_type, ($id));
  }
  return $symbol_table{$id}->{TYPE};
}
```

33.10.1. Práctica: Declaraciones Automáticas

Modifique la subrutina `check_declared` para que cuando una variable no haya sido declarada se declare "sobre la marcha". ¿Puede utilizar información dependiente del contexto para decidir cual es la mejor forma de declararla?

33.10.2. Práctica: Análisis Semántico

Extienda el código de la práctica 33.9.5 para comprobar la compatibilidad de tipos.

1	p → b
2	b → ds ss
3	b → ss
4	ds → d ';' ds
5	ds → d ';'
6	d → INT il
7	d → STRING il
8	ss → s ';' ss
9	ss → s
10	s → ID '=' e
11	s → '{' b '}'
12	s → P e
13	s → ϵ
14	e → e1 '+' t
15	e → e1 '-' t
16	e → t
17	t → t1 '*' f
18	t → t '/' f
19	t → f
20	f → '(' e ')'
21	f → ID
22	f → NUM
23	f → STR
24	il → ID ',' il
25	il → ID

En cuanto a las sentencias de bloque, se pretende que el ámbito y visibilidad de las variables sea como en el lenguaje C, esto es, las declaraciones mas internas con el mismo identificador ocultan las mas externas. Así:

```
int a;
a = 4;
{
  int a;
  a = 5;
  p a
}; /* el ; es necesario */
p a
```

Imprimiría 5 y 4. Para traducir esta sentencia es necesario *usar una lista/pila de referencias a tablas de símbolos. Cada sentencia compuesta o bloque tendrá su propia tabla de símbolos.* Los identificadores se búscan en la lista de referencias a tablas de símbolos, primero en la última tabla de símbolos insertada y sino se encuentra se busca en la penúltima insertada, etc.

Guarde como un atributo del identificador ($SYMTABLE$) la referencia a la tabla de símbolos a la que pertenece. Guarde como un atributo del nodo bloque ($BLOCK$) la referencia a la tabla de símbolos asociada.

33.11. Optimización Independiente de la Máquina

En esta fase se hace un análisis del árbol, sometiéndolo a transformaciones que aumenten la eficiencia del código final producido.

Ejemplos de tareas que se pueden llevar a cabo en esta fase son:

- Extracción del interior de un bucle de cálculos que son invariantes del bucle
- Plegado de constantes: computar las expresiones constantes en tiempo de compilación, no de ejecución
- Propagación de las constantes: si se sabe que una variable en un punto del programa tiene un valor constante `a = 4`, se puede sustituir su uso por el de la constante
- Eliminación de computaciones redundantes, cuando la misma expresión aparece repetidas veces con los mismos valores de las variables
- La eliminación de código "muerto": código que se sabe que nunca podrá ser ejecutado

En nuestro primer ejemplo, reduciremos esta fase a realizar una tarea de plegado de las constantes. Primero lo haremos mediante la rutina

`&Machine::Independent::Optimization::fold`

En esta fase transformamos los AAA: si tenemos un árbol de la forma OPERATION(left, right), esto es, su raíz es una operación, primero plegamos los subárboles left y right, y si se han transformado en constantes numéricas, entonces plegamos el nodo que pasa a ser numérico:

```
 1 sub operator_fold { # Obsérvese el uso del aliasing
 2
 3   if ($_[0]->LEFT->is_operation) {
 4     $_[0]->{LEFT}->fold;
 5   }
 6   if ($_[0]->RIGHT->is_operation) {
 7     $_[0]->{RIGHT}->fold;
 8   }
 9   if (ref($_[0]->LEFT) eq "NUM" and ref($_[0]->RIGHT) eq "NUM") {
10     $_[0] = reduce_children($_[0]);
11   }
12 }
13
14 sub PLUS::fold {
15  operator_fold(@_);
16 }
17
18 sub TIMES::fold {
19  operator_fold(@_);
20 }
```

El plegado de las operaciones binarias se ha delegado en la subrutina `operator_fold`. En las líneas 3 y 6 se comprueba que se trata de un nodo de tipo operación. Si es así se procede a su plegado. Una vez plegados los dos subárboles hijo comprobamos en la línea 9 que los hijos actuales son de la clase `NUM`. Si es el caso, en la línea 10 cambiamos el nodo por el resultado de operar los dos hijos. Los nodos han sido extendidos con un método `is_operation` que determina si se trata de un nodo operación binaria o no. Para ello se han introducido nuevas clases de nodos: la clase `Node` está en la raíz de la jerarquía de herencia, las clases `Leaf` y `Binary` se usan para representar los nodos hoja y binarios y

heredan de la anterior. Una clase informa a Perl que desea heredar de otra clase añadiendo el nombre de esa clase a la variable `@ISA` de su paquete. La herencia en Perl determina la forma de búsqueda de un método. Si el objeto no se puede encontrar en la clase, recursivamente y en orden primero-profundo se busca en las clases de las cuales esta hereda, esto es en las clases especificadas en el vector `@ISA`.

```
package Node;

sub is_operation {
  my $node = shift;

  return ref($node) =~ /^(TIMES)|(PLUS)$/;
}

package Leaf; # hoja del AAA
our @ISA = ("Node");
sub children {
  return ();
}

package Binary;
our @ISA = ("Node");
sub children {
  my $self = shift;

  return (LEFT => $self->{LEFT}, RIGHT => $self->{RIGHT});
}
```

Así pues, los objetos de la clase `Leaf` tienen acceso al método `is_operation`.

Ahora hacemos que las clases `PLUS` y `TIMES` hereden de la clase `BINARY`:

```
package PLUS;
our @ISA = ("Binary");

sub operator {
  my $self = shift;

  $_[0]+$_[1];
}

....

package TIMES;
our @ISA = ("Binary");

sub operator {
  my $self = shift;

  $_[0]*$_[1];
}

....
```

Obsérvese que en las líneas 4 y 7 del código del plegado de nodos de operación se ha accedido directamente al dato en vez de usar el método para modificar el atributo, saltándonos lo que la buena

programación orientada a objetos indica. La forma en la que esta escrito hace que, por ejemplo, `$_[0]->{LEFT}` sea modificado. Recuérdese que en Perl *los argumentos son alias de los parámetros*.

La subrutina `reduce_children` es la encargada de crear el nuevo nodo con el resultado de operar los hijos izquierdo y derecho:

```
1 sub reduce_children {
2   my ($node) = @_;
3
4   my $value = $node->operator($node->LEFT->VAL, $node->RIGHT->VAL);
5   NUM->new(VAL => $value, TYPE => $PL::Tutu::int_type);
6 }
```

En la línea 4 se usa el método `operator` asociado con un nodo operación.

Plegar una sentencia de impresión es plegar la expresión a imprimir:

```
sub PRINT::fold {
  $_[0]->{EXPRESSION}->fold;
}
```

Plegar una sentencia de asignación es plegar la parte derecha de la asignación:

```
sub ASSIGN::fold {
  $_[0]->{RIGHT}->fold;
}
```

de nuevo, hemos accedido a los campos en vez de usar los métodos.

Las restantes operaciones de plegado son triviales:

```
sub ID::fold { }

sub NUM::fold { }

sub STR::fold { }
```

Por último, para plegar todas las expresiones recorremos la lista de sentencias del programa y las plegamos una a una.

```
sub fold {
  my @statements = @{$tree->{STS}};
  for my $s (@statements) {
    $s->fold;
  }
}
```

33.11.1. Práctica: Plegado de las Constantes

Complete su proyecto de compilador de Tutu con la fase de plegado de las constantes siguiendo la metodología explicada en los párrafos previos. Mejore la jerarquía de clases con una clase abstracta `Operation` que represente a los nodos que se corresponden con operaciones binarias. Defina el método abstracto `operation` en dicha clase. Un *método abstracto* es uno que, mas que proveer un servicio representa un servicio o categoría. La idea es que al definir un clase base abstracta se indica un conjunto de métodos que deberían estar definidos en todas las clases que heredan de la clase base abstracta. Es como una declaración de interfaz que indica la necesidad de definir su funcionalidad en las clases descendientes, pero que no se define en la clase base. Un método abstracto debe producir una excepción con el mensaje de error adecuado si no se ha redefinido en la clase desendiente.

Para ello use la clave `abstract` del módulo `Class::MethodMaker`. Consulte la documentación del módulo `Class::MethodMaker`. Consulte [10] para saber más sobre clases abstractas.

33.12. Patrones Árbol y Transformaciones Árbol

En la fase de optimización presentada en la sección 33.11 transformabamos el programa en su representación intermedia, como un AAA decorado, para obtener otro AAA decorado.

Una transformación de un programa puede ser descrita como un conjunto de *reglas de transformación* o *esquema de traducción árbol* sobre el árbol abstracto que representa el programa.

Antes de seguir, es conveniente que repase los conceptos en la sección 33.9.1 sobre lenguajes y gramáticas árbol.

En su forma mas sencilla, estas reglas de transformación vienen definidas por ternas $(p, e, action)$, donde la primera componente de la terna p es un *patrón árbol* que dice que árboles deben ser seleccionados. La segunda componente e dice cómo debe transformarse el árbol que casa con el patrón p. La acción *action* indica como deben computarse los atributos del árbol transformado a partir de los atributos del árbol que casa con el patrón p. Una forma de representar este esquema sería:

$$p \Longrightarrow e \texttt{ \{ action \}}$$

Por ejemplo:

$PLUS(NUM_1, NUM_2) \Longrightarrow NUM_3$ `{ $NUM_3{VAL} = $NUM_1{VAL} + $NUM_2{VAL} }`

cuyo significado es que dondequiera que haya un nódo del AAA que case con el *patrón de entrada* $PLUS(NUM, NUM)$ deberá sustituirse el subárbol $PLUS(NUM, NUM)$ por el subárbol NUM. Al igual que en los esquemas de traducción, enumeramos las apariciones de los símbolos, para distinguirlos en la parte semántica. La acción indica como deben recomputarse los atributos para el nuevo árbol: El atributo `VAL` del árbol resultante es la suma de los atributos `VAL` de los operandos en el árbol que ha casado. La transformación se repite hasta que se produce la *normalización del árbol.*

Las reglas de "casamiento" de árboles pueden ser mas complejas, haciendo alusión a propiedades de los atributos, por ejemplo

$ASSIGN(LEFTVALUE, x)$ *and* `{ notlive($LEFTVALUE{VAL}) }` $\Longrightarrow NIL$

indica que se pueden eliminar aquellos árboles de tipo asignación en los cuáles la variable asociada con el nodo $LEFTVALUE$ no se usa posteriormente.

Otros ejemplos con variables S_1 y S_2:

$IFELSE(NUM, S_1, S_2)$ and `{ $NUM{VAL} != 0 }` $\Longrightarrow S_1$
$IFELSE(NUM, S_1, S_2)$ and `{ $NUM{VAL} == 0 }` $\Longrightarrow S_2$

Observe que en el patrón de entrada $ASSIGN(LEFTVALUE, x)$ aparece un "comodín": la variable-árbol x, que hace que el árbol patrón $ASSIGN(LEFTVALUE, x)$ case con cualquier árbol de asignación, independientemente de la forma que tenga su subárbol derecho.

Las siguientes definiciones formalizan una aproximación simplificada al significado de los conceptos *patrones árbol* y *casamiento de árboles.*

Definición 33.12.1. *Sea (Σ, ρ) un alfabeto con función de aridad y un conjunto (puede ser infinito) de variables $V = \{x_1, x_2, \ldots\}$. Las variables tienen aridad cero:*

$$\rho(x) = 0 \ \forall x \in V.$$

Un elemento de $B(V \cup \Sigma)$ se denomina patrón *sobre Σ.*

Definición 33.12.2. *Se dice que un patrón es un* patrón lineal *si ninguna variable se repite.*

Definición 33.12.3. *Se dice que un patrón es de tipo $(x_1, \ldots x_k)$ si las variables que aparecen en el patrón leidas de izquierda a derecha en el árbol son $x_1, \ldots x_k$.*

Ejemplo 33.12.1. *Sea $\Sigma = \{A, CONS, NIL\}$ con $\rho(A) = \rho(NIL) = 0, \rho(CONS) = 2$ y sea $V = \{x\}$. Los siguientes árboles son ejemplos de patrones sobre Σ:*

$$\{\ x, CONS(A,x), CONS(A, CONS(x, NIL)), \ldots\}$$

El patrón $CONS(x, CONS(x, NIL))$ es un ejemplo de patrón no lineal. La idea es que un patrón lineal como éste "fuerza" a que los árboles t que casen con el patrón deben tener iguales los dos correspondientes subárboles $t/1$ y $t/2.1$ situados en las posiciones de las variables [1]

Ejercicio 33.12.1. *Dado la gramática árbol:*

$$S \to S_1(a, S, b)$$
$$S \to S_2(NIL)$$

la cuál genera los árboles concretos para la gramática

$$S \to aSb \mid \epsilon$$

¿Es $S_1(a, X(NIL), b)$ un patrón árbol sobre el conjunto de variables $\{X, Y\}$? ¿Lo es $S_1(X, Y, a)$? ¿Es $S_1(X, Y, Y)$ un patrón árbol?

Ejemplo 33.12.2. *Ejemplos de patrones para el AAA definido en el ejemplo 33.9.2 para el lenguaje Tutu son:*

$$x, y, PLUS(x,y), ASSIGN(x, TIMES(y, ID)), PRINT(y) \ldots$$

considerando el conjunto de variables $V = \{x, y\}$. El patrón $ASSIGN(x, TIMES(y, ID))$ es del tipo (x, y).

Definición 33.12.4. *Una* sustitución *es una aplicación θ que asigna variables a patrones $\theta : V \to B(V \cup \Sigma)$.*

Tal función puede ser naturalmente extendida de las variables a los árboles: los nodos (hoja) etiquetados con dichas variables son sustituidos por los correspondientes subárboles.

$$\theta : B(V \cup \Sigma) \to B(V \cup \Sigma)$$
$$t\theta = \begin{cases} x\theta & si\ t = x \in V \\ a(t_1\theta, \ldots, t_k\theta) & si\ t = a(t_1, \ldots, t_k) \end{cases}$$

Obsérvese que, al revés de lo que es costumbre, la aplicación de la sustitución θ al patrón se escribe por detrás: $t\theta$.

También se escribe $t\theta = t\{x_1/x_1\theta, \ldots x_k/x_k\theta\}$ si las variables que aparecen en t de izquierda a derecha son $x_1, \ldots x_k$.

Ejemplo 33.12.3. *Si aplicamos la sustitución $\theta = \{x/A, y/CONS(A, NIL)\}$ al patrón $CONS(x, y)$ obtenemos el árbol $CONS(A, CONS(A, NIL))$. En efecto:*

$$CONS(x,y)\theta = CONS(x\theta, y\theta) = CONS(A, CONS(A, NIL))$$

Ejemplo 33.12.4. *Si aplicamos la sustitución $\theta = \{x/PLUS(NUM, x), y/TIMES(ID, NUM)\}$ al patrón $PLUS(x, y)$ obtenemos el árbol $PLUS(PLUS(NUM, x), TIMES(ID, NUM))$:*

$$PLUS(x,y)\theta = PLUS(x\theta, y\theta) = PLUS(PLUS(NUM, x), TIMES(ID, NUM))$$

Definición 33.12.5. *Se dice que un patrón $\tau \in B(V \cup \Sigma)$ con variables $x_1, \ldots x_k$* casa con un árbol *$t \in B(\Sigma)$ si existe una sustitución de τ que produce t, esto es, si existen $t_1, \ldots t_k \in B(\Sigma)$ tales que $t = \tau\{x_1/t_1, \ldots x_k/t_k\}$. También se dice que τ* casa con la sustitución *$\{x_1/t_1, \ldots x_k/t_k\}$.*

Ejemplo 33.12.5. *El patrón $\tau = CONS(x, NIL)$ casa con el árbol $t = CONS(CONS(A, NIL), NIL)$ y con el subárbol $t.1$. Las respectivas sustituciones son $t\{x/CONS(A, NIL)\}$ y $t.1\{x/A\}$.*

[1] Repase la notación de Dewey introducida en la definición 33.9.6

$$t = \tau\{x/CONS(A, NIL)\}$$
$$t.1 = \tau\{x/A\}$$

Ejercicio 33.12.2. *Sea* $\tau = PLUS(x, y)$ *y* $t = TIMES(PLUS(NUM, NUM), TIMES(ID, ID))$. *Calcule los subárboles* t' *de* t *y las sustituciones* $\{x/t_1, y/t_2\}$ *que hacen que* τ *case con* t'.

Por ejemplo es obvio que para el árbol raíz t/ϵ *no existe sustitución posible:*
$t = TIMES(PLUS(NUM, NUM), TIMES(ID, ID)) = \tau\{x/t_1, y/t_2\} = PLUS(x, y)\{x/t_1, y/t_2\}$
ya que un término con raíz $TIMES$ *nunca podrá ser igual a un término con raíz* $PLUS$.

El problema aquí es equivalente al de las expresiones regulares en el caso de los lenguajes lineales. En aquellos, los autómatas finitos nos proveen con un mecanismo para reconocer si una determinada cadena "casa"' o no con la expresión regular. Existe un concepto análogo, el de *autómata árbol* que resuelve el problema del "casamiento" de patrones árbol. Al igual que el concepto de autómata permite la construcción de software para la búsqueda de cadenas y su posterior modificación, el concepto de autómata árbol permite la construcción de software para la búsqueda de los subárboles que casan con un patrón árbol dado.

Estamos ahora en condiciones de plantear una segunda aproximación al problema de la optimización independiente de la máquina utilizando una subrutina que busque por aquellos árboles que queremos optimizar (en el caso del plegado los árboles de tipo operación) y los transforme adecuadamente.

La función `match_and_transform_list` recibe una lista de árboles los cuales recorre sometiéndolos a las transformaciones especificadas. La llamada para producir el plegado sería:

```
Tree::Transform::match_and_transform_list(
  NODES => $tree->{STS}, # lista de sentencias
  PATTERN => sub {
     $_[0]->is_operation and $_[0]->LEFT->isa("NUM")
     and $_[0]->RIGHT->isa("NUM")
  },
  ACTION => sub {
    $_[0] = Machine::Independent::Optimization::reduce_children($_[0])
  }
);
```

Además de la lista de nodos le pasamos como argumentos una referencia a la subrutina encargada de reconocer los patrónes árbol (clave `PATTERN`) y una referencia a la subrutina que describe la acción que se ejecutará (clave `ACTION`) sobre el árbol que ha casado. Ambas subrutinas asumen que el primer argumento que se les pasa es la referencia a la raíz del árbol que está siendo explorado.

Los métodos *isa*, *can* y *VERSION* son proporcionados por una clase especial denominada clase `UNIVERSAL`, de la cual implícitamente hereda toda clase. El método `isa` nos permite saber si una clase hereda de otra.

La subrutina `match_and_transform_list` recibe los argumentos y da valores por defecto a los mismos en el caso de que no hayan sido establecidos. Finalmente, llama a `match_and_transform` sobre cada uno de los nodos "sentencia" del programa.

```
sub match_and_transform_list {
  my %arg = @_;
  my @statements = @{$arg{NODES}} or
   Error::fatal("Internal error. match_and_transform_list ".
   "espera una lista anónima de nodos");
  local $pattern = ($arg{PATTERN} or sub { 1 });
  local @pattern_args = @{$arg{PATTERN_ARGS}} if defined $arg{PATTERN_ARGS};
  local $action = ($arg{ACTION} or sub { print ref($_[0]),"\n" });
  local @action_args = @{$arg{ACTION_ARGS}} if defined $arg{ACTION_ARGS};
```

```
  for (@statements) {
    match_and_transform($_);
  }
}
```

La subrutina `match_and_transform` utiliza el método `can` para comprobar que el nodo actual dispone de un método para calcular la lista con los hijos del nodo. Una vez transformados los subárboles del nodo actual procede a comprobar que el nodo casa con el patrón y si es el caso le aplica la acción definida:

```
package Tree::Transform;

our $pattern;
our @pattern_args;
our $action;
our @action_args;
our @statements;

sub match_and_transform {
  my $node = $_[0] or Error::fatal("Error interno. match_and_transform necesita un nodo");
  Error::fatal("Error interno. El nodo de la clase",ref($node),
               " no dispone de método 'children'") unless $node->can("children");

  my %children = $node->children;

  for my $k (keys %children) {
    $node->{$k} = match_and_transform($children{$k});
  }

  if ($pattern->($node, @pattern_args)) {
    $action->( $node, @action_args);
  }
  return $node;
}
```

Recordemos el esquema de herencia que presentamos en la sección anterior. Las clases `Leaf` y `Binary` proveen versiones del método `children`. Teníamos:

```
package Node;

sub is_operation {
  my $node = shift;

  return ref($node) =~ /^(TIMES)|(PLUS)$/;
}

package Leaf; # hoja del AAA
our @ISA = ("Node");
sub children {
  return ();
}

package Binary;
our @ISA = ("Node");
```

```
sub children {
  my $self = shift;

  return (LEFT => $self->{LEFT}, RIGHT => $self->{RIGHT});
}
```

Los objetos de la clase `Leaf` tienen acceso al método `is_operation`.

Las clases `PLUS` y `TIMES` heredan de la clase `BINARY`:

```
package PLUS;
our @ISA = ("Binary");

sub operator {
  my $self = shift;

  $_[0]+$_[1];
}

....

package TIMES;
our @ISA = ("Binary");

sub operator {
  my $self = shift;

  $_[0]*$_[1];
}

....
```

La subrutina `reduce_children` introducida en la sección 33.11 es la encargada de crear el nuevo nodo con el resultado de operar los hijos izquierdo y derecho:

```
1 sub reduce_children {
2   my ($node) = @_;
3
4   my $value = $node->operator($node->LEFT->VAL, $node->RIGHT->VAL);
5   NUM->new(VAL => $value, TYPE => $PL::Tutu::int_type);
6 }
```

En la línea 4 se usa el método `operator` asociado con un nodo operación.

33.12.1. Práctica: Casando y Transformando Árboles

Complete su proyecto para el compilador de Tutu completando las subrutinas `match_and_transform` tal y como se explicó en la sección 33.12.

Ademas del plegado de constantes use las nuevas subrutinas para aplicar simultáneamente las siguientes transformaciones algebraicas:

$PLUS(NUM, x)$ $\wedge$ `{ $NUM{VAL} == 0 }` $\Longrightarrow x$
$PLUS(x, NUM)$ $\wedge$ `{ $NUM{VAL} == 0 }` $\Longrightarrow x$
$TIMES(NUM, x)$ $\wedge$ `{ $NUM{VAL} == 1 }` $\Longrightarrow x$
$TIMES(x, NUM)$ $\wedge$ `{ $NUM{VAL} == 1 }` $\Longrightarrow x$

1. Dado un programa como

```
int a; a = a * 4 * 5;
```

¿Será plegado el `4 * 5`? Sin embargo si que se pliega si el programa es de la forma:

```
int a; a = a * (4 * 5);
```

No intente en esta práctica que programas como el primero o como `4*a*5*b` sean plegados. Para lograrlo sería necesario introducir transformaciones adicionales y esto no se requiere en esta práctica.

2. ¿Existe un orden óptimo en el que ejecutar las transformaciones?
3. Ponga un ejemplo en el que sea beneficioso ejecutar el plegado primero.
4. Ponga otro ejemplo en el que sea beneficioso ejecutar el plegado después.
5. ¿Es necesario aplicar las transformaciones reiteradamente?
6. ¿Cuál es la condición de parada?
7. Como es habitual la pregunta 6 tiene una respuesta TIMTOWTDI: una posibilidad la da el módulo Data::Compare el cual puede obtenerse desde CPAN y que permite comparar estructuras de datos, pero existe una solución mas sencilla. ¿Cuál?

33.13. Asignación de Direcciones

Esta suele ser considerada la primera de las fases de síntesis. Las anteriores lo eran de análisis. La fase de análisis es una transformación *texto fuente* $\rightarrow$ *arbol*, mientras que la fase síntesis es una transformación inversa *arbol* $\rightarrow$ *texto objeto* que produce una "linealización" del árbol. En general, se ubican en la fase de síntesis todas las tareas que dependan del lenguaje objeto.

La asignación de direcciones depende de la máquina objetivo en cuanto conlleva consideraciones sobre la longitud de palabra, las unidades de memoria direccionables, la compactación de objetos pequeños (por ejemplo, valores lógicos), el alineamiento a fronteras de palabra, etc.

En nuestro caso debemos distinguir entre las cadenas y las variables enteras.

Las constantes literales (como `"hola"`) se almacenan concatenadas en orden de aparición textual. Una variable de tipo cadena ocupa dos palabras, una dando su dirección y otra dando su longitud.

```
sub factor() {
  my ($e, $id, $str, $num);

  if ($lookahead eq 'NUM') { ... }
  elsif ($lookahead eq 'ID') { ... }
  elsif ($lookahead eq 'STR') {
    $str = $value;
    my ($offset, $length) = Address::Assignment::str($str);
    match('STR');
    return STR->new(OFFSET => $offset, LENGTH => $length, TYPE => $string_type);
  }
  elsif ($lookahead eq '(') { ... }
  else { Error::fatal("Se esperaba (, NUM o ID"); }
}
```

El código de la subrutina `Address::Assignment::str` es:

```
  sub str {
    my $str = shift;
    my $len = length($str);
    my $offset = length($data);
    $data .= $str;
    return ($offset, $len);
  }
```

Una posible mejora es la que se describe en el punto 2 de la práctica 33.13.1.

Hemos supuesto que las variables enteras y las referencias ocupan una palabra. Cada vez que una variable es declarada, se le computa su dirección relativa:

```
# declaration -> type idlist
# type        -> INT | STRING
sub declaration() {
  my ($t, $decl, @il);

  if (($lookahead eq 'INT') or ($lookahead eq 'STRING')) {
    $t = &type(); @il = &idlist();
    &Semantic::Analysis::set_types($t, @il);
    &Address::Assignment::compute_address($t, @il);
    $decl = [$t, \@il];
    return bless $decl, 'DECL';
  }
  else { Error::fatal('Se esperaba un tipo'); }
}
```

Se usa una variable `$global_address` para llevar la cuenta de la última dirección utilizada y se introduce un atributo `ADDRESS` en la tabla de símbolos:

```
  sub compute_address {
    my $type = shift;
    my @id_list = @_;

    for my $i (@id_list) {
      $symbol_table{$i}->{ADDRESS} = $global_address;
      $global_address += $type->LENGTH;
    }
  }
```

Por último situamos todas las cadenas después de las variables del programa fuente:

```
  ...
  ##### En compile, despues de haber calculado las direcciones
  Tree::Transform::match_and_transform_list(
    NODES => $tree->{STS},
    PATTERN => sub {
      $_[0]->isa('STR')
    },
    ACTION => sub { $_[0]->{OFFSET} += $global_address; }
  );
```

Esta aproximación es bastante simplista en cuanto que usa una palabra por carácter. El punto 3 de la práctica 33.13.1 propone una mejora.

33.13.1. Práctica: Cálculo de las Direcciones

Modifique el cálculo de las direcciones para la gramática de Tutu extendida con sentencias compuestas que fué presentada en la sección 33.10.2.

1. Resuelva el problema de calcular las direcciones de las variables declaradas en los bloques. La memoria ocupada por dichas variables se libera al terminar el bloque. Por tanto la variable `$global_address` disminuye al final de cada bloque.

2. En el código de la subrutina `Address::Assignment::str`

   ```
   sub str {
     my $str = shift;
     my $len = length($str);
     my $offset = length($data);
     $data .= $str;
     return ($offset, $len);
   }
   ```

   no se comprueba si la cadena `$str` ya está en `$data`. Es natural que si ya está no la insertemos de nuevo. Introduzca esa mejora. Para ello use la función `pos`, la cual devuelve el desplazamiento en `$data` donde quedó la última búsqueda `$data =~ m/regexp/g`. El siguiente ejemplo con el depurador le muestra como trabaja. Escriba `perldoc -f pos` para obtener mas información sobre la función `pos`. Vea un ejemplo de uso:

   ```
   DB<1> $str = "world"
   #              01234567890123456789012345678901234
   DB<2> $data = "hello worldjust another statement"
   DB<3> $res = $data =~ m{$str}g
   DB<4> x pos($data) # la siguiente busqueda comienza en la 11
   0  11
   DB<5> $s = "hello"
   DB<6> $res = $data =~ m{$s}g
   DB<7> x pos($data) # No hay "hello" despues de world en $data
   0  undef
   DB<8> $res = $data =~ m{$str}g # Repetimos ...
   DB<9> x pos($data)
   0  11
   DB<10> pos($data) = 0  # Podemos reiniciar pos!
   DB<11> $res = $data =~ m{$s}g
   DB<12> x pos($data) # Ahora si se encuentra "hello"
   0  5
   ```

3. Empaquete las cadenas asumiendo que cada carácter ocupa un octeto o byte. Una posible solución al problema de alineamiento que se produce es rellenar con ceros los bytes que faltan hasta completar la última palabra ocupada por la cadena. Por ejemplo, si la longitud de palabra de la máquina objeto es de 4 bytes, la palabra *procesador* se cargaría asi:

   ```
     0    |    1    |    2
   0 1 2 3 | 4 5 6 7 | 8 9 10  11
   p r o c | e s a d | o r \0  \0
   ```

   Existen dos posibles formas de hacer esta empaquetado. Por ejemplo, si tenemos las cadenas `lucas` y `pedro` las podemos introducir en la cadena `$data` asi:

```
   0    |     1      |    2      |    3
0 1 2 3 | 4  5  6  7 | 8 9 10 11 | 12 13 14 15
l u c a | s \0 \0 \0 | p e  d  r |  o \0 \0 \0
```

o bien:

```
   0    |    1    |    2
0 1 2 3 | 4 5 6 7 | 8 9 10 11
l u c a | s p e d | r o \0 \0
```

La segunda forma aunque compacta más tiene la desventaja de hacer luego mas lento el código para el direccionamiento de las cadenas, ya que no empiezan en frontera de palabra. Utilice el primer modo.

Observe que este empaquetado introduce un efecto en lo que se considera un éxito en la búsqueda de una cadena `$str` en `$data`. ¿Que ocurre si `$str` aparece en `$data` como subcadena de una cadena previamente empaquetada ocupando una posición que no es frontera de palabra?

Cuando rellene con `\0` la cadena use el operador `x` (letra equis) para multiplicar número por cadena:

```
DB<1> $x = "hola"x4
DB<2> p $x
holaholaholahola
```

esto le evitará escribir un bucle.

4. Mantega atributos distintos en el nodo `STR` para el desplazamiento y longitud de `$str` en `$data` (en caracteres) y su dirección y tamaño final en memoria (en palabras).

33.14. Generación de Código: Máquina Pila

El generador de código emite el código para el lenguaje objeto, utilizando las direcciones calculadas en la fase anterior. Hemos simplificado esta fase considerando una máquina orientada a pila: una memoria de pila en la que se computan las expresiones, un segmento de datos en el que se guardan las variables y un segmento para guardar las cadenas. La estrategia utilizada es similar a la que vimos en el plegado de las constantes. Definimos un método para la traducción de cada clase de nodo del AAA. Entonces, para traducir el programa basta con llamar al correspondiente método:

```
  $tree = Syntax::Analysis::parser;
  ... # otras fases
  ########code generation
  local $target = ""; # target code
  $tree->translate;
  ...
```

en la cadena `$target` dejamos el código emitido. Cada método visita los hijos, traduciéndolos y añadiendo el código que fuera necesario para la traducción del nodo.

```
sub PROGRAM::translate {
  my $tree = shift;

  $target .= "DATA ". $data."\n" if $data;
  $tree->STS->translate;
}
```

Traducir la lista de sentencias es concatenar la traducción de cada una de las sentencias en la lista:

```
sub STATEMENTS::translate {
  my $statements = shift;
  my @statements = @{$statements};

  for my $s (@statements) {
    $s->translate;
  }
}
```

Si suponemos que disponemos en el ensamblador de nuestra máquina objeto de instrucciones `PRINT_INT` y `PRINT_STR` que imprimen el contenido de la expresión que esta en la cima de la pila, la traducción será:

```
sub PRINT::translate {
  my $self = shift;

  $self->EXPRESSION->translate;
  if ($self->EXPRESSION->TYPE == $int_type) { emit "PRINT_INT\n"; }
  else {emit "PRINT_STR\n"; }
}
```

Asi, si la sentencia era `P c`, donde `c` es del tipo cadena, se debería eventualmente llamar al método de traducción del identificador:

```
sub ID::translate {
  my $self = shift;

  my $id = $self->VAL;
  my $type =  Semantic::Analysis::get_type($id);
  if ($type == $int_type) {
    emit "LOAD ".$symbol_table{$id}->{ADDRESS}."\n";
  }
  else {
    emit "LOAD_STRING ".$symbol_table{$id}->{ADDRESS}."\n";
  }
}
```

la función `emit` simplemente concatena el código producido a la salida:

```
sub emit { $target .= shift; }
```

Para la traducción de una sentencia de asignación supondremos de la existencia de isntrucciones `STORE_INT` y `STORE_STRING`. La instrucción `STORE_STRING` asume que en la cima de la pila están la dirección y la cadena a almacenar.

```
sub ASSIGN::translate {
  my $self = shift;

  $self->RIGHT->translate;
  my $id = $self->LEFT;
  $id->translate;
  my $type =  Semantic::Analysis::get_type($id->VAL);
```

```
  if ($type == $int_type) {
    emit "STORE_INT\n";
  }
  else {
    emit "STORE_STRING\n";
  }
}
```

Si se está traduciendo una sentencia como `a = "hola"`, se acabará llamando al método de traducción asociado con la clase `STR` el cual actúa empujando la dirección y la longitud de la cadena en la pila:

```
sub STR::translate {
  my $self = shift;

  emit "PUSHSTR ".$self->OFFSET." ".$self->LENGTH."\n";
}
```

Así la traducción de este fuente:

```
$ cat test06.tutu
string a;
a = "hola";
p a
```

es:

```
$ ./main.pl test06.tutu test06.ok
$ cat test06.ok
DATA 2, hola
PUSHSTR 2 4
PUSHADDR 0
STORE_STRING
LOAD_STRING 0
PRINT_STR
```

El resto de los métodos de traducción es similar:

```
sub LEFTVALUE::translate {
  my $id = shift ->VAL;

  emit "PUSHADDR ".$symbol_table{$id}->{ADDRESS}."\n";
}

sub NUM::translate {
  my $self = shift;

  emit "PUSH ".$self->VAL."\n";
}

sub PLUS::translate {
  my $self = shift;

  $self->LEFT->translate;
  $self->RIGHT->translate;
  emit "PLUS\n";
```

```
}

sub TIMES::translate {
  my $self = shift;

  $self->LEFT->translate;
  $self->RIGHT->translate;
  emit "MULT\n";
}
```

Veamos un ejemplo. Dado el código fuente:

```
$ cat test02.tutu
int a,b; string c;
a = 2+3; b = 3*4; c = "hola"; p c;
c = "mundo"; p c; p 9+2; p a+1; p b+1
```

el código generado es:

```
$ ./main.pl test02.tutu test02.ok
$ cat test02.ok
DATA 4, holamundo
PUSH 5
PUSHADDR 0
STORE_INT
PUSH 12
PUSHADDR 1
STORE_INT
PUSHSTR 4 4
PUSHADDR 2
STORE_STRING
LOAD_STRING 2
PRINT_STR
PUSHSTR 8 5
PUSHADDR 2
STORE_STRING
LOAD_STRING 2
PRINT_STR
PUSH 11
PRINT_INT
LOAD 0
INC
PRINT_INT
LOAD 1
INC
PRINT_INT
```

33.15. Generación de Código: Máquina Basada en Registros

La máquina orientada a pila para la que generamos código en la sección 33.14 es un ejemplo de la clase de máquinas que es usada por la mayoría de los lenguajes interpretados: Perl, Python; java, etc.

En esta sección introduciremos una máquina basada en registros. Suponemos que la máquina tiene k registros $R_0 \dots R_{k-1}$. Las instrucciones toman dos argumentos, dejando el resultado en el primer argumento. Son las siguientes:

`LOADM Ri, [a]`	$R_i = M_a$
`LOADC Ri, c`	$R_i = c$
`STORE [a], Ri`	$M_a = R_i$
`ADDR Ri, Rj`	$R_i += R_j$
`ADDM Ri, [a]`	$R_i += M_a$
`ADDC Ri, c`	$R_i += c$
...	...

El problema es generar el código con el menor número de instrucciones posible, teniendo en cuenta la limitación existente de registros.

Supongamos que queremos traducir un subárbol $OP(t_1, t_2)$ y que la traducción del subárbol t_1 requiere r_1 registros y que la traducción de t_2 requiere r_2 registros, con $r_1 < r_2 \leq k$. Si realizamos primero la evaluación de t_1, debemos dejar el resultado en un registro que no podrá ser utilizado en la evaluación de t_2. Si $r_2 = k$, la evaluación de t_2 podría dar lugar a la necesidad de recurrir a almacenamiento temporal. Esta situación no se da si evaluamos primero t_2. En tal caso, dado que hay un registro en el que se guarda el resultado de t_2, quedan libres al menos $r_2 - 1$ registros. Como $r_2 - 1 \geq r_1$ se sigue que tenemos suficientes registros para traducir t_1. Como regla general es mejor evaluar primero el subárbol que mayores requerimientos de registros tiene.

La siguiente cuestión es como calcular los requerimientos en registros de una expresión dada. No consideraremos en esta fase límites en el número de registros disponibles. Obsérvese que si los requerimientos para los subárboles son distintos, $r_1 \neq r_2$ la traducción puede realizarse usando el máximo de ambos máx$\{r_1, r_2\}$ siguiendo la estrategia de traducir primero el que mayores requerimentos tenga. Si son iguales entonces se necesitan $r_1 + 1$ registros ya que es necesario un registro para guardar el resultado de la primera traducción.

Nótese que, como el juego de instrucciones para un operando puede tener como segundo argumento una dirección de memoria, los "segundos operandos" no necesitan registro. Por ejemplo, el árbol $PLUS(a, b)$ se traduce por

```
LOADM R0, a
PLUSM R0, b
```

Asi b no requiere registro, mientras que a si lo requiere. Por tanto, las hojas izquierdas requieren de registro mientras que las hojas derechas no.

Si t es un nodo de la forma $OP(t_1, t_2)$ el número de registros r_t requeridos por t viene dado por la fórmula:

$$r_t = \begin{cases} \text{máx}\{r_1, r_2\} & \text{si } r_1 \neq r_2 \\ r_1 + 1 & \text{si } r_1 = r_2 \end{cases}$$

Dotaremos a cada nodo del AST de un método `required_registers` que computa la demanda en registros de dicho nodo. Lo que haremos es introducir en la clase `Operation` de la cual heredan las operaciones binarias el correspondiente método `required_registers`:

```
package Operation;
our @ISA = ("Binary");

sub required_registers {
  my $self = shift;

  my $rl = $self->LEFT->required_registers('LEFT');
  my $rr = $self->RIGHT->required_registers('RIGHT');
  $self->{REQ_REG} = ($rl == $rr)? $rl+1: Aux::max($rl, $rr);
  return $self->REQ_REG;
}
```

El segundo argumento que recibe `required_registers` es su posición (izquierda o derecha) entre los hijos de su padre. dicha información no es usada en los nodos binarios. Su necesidad queda clara cuando se considera el cómputo del número de registros requeridos por las hojas.

El cómputo en las hojas corre a cargo del correspondiente método en la clase `Value`. Los nodos de tipo número (clase `NUM`), cadena (clase `STR`) y variable (clase `ID`) heredan de la clase `Value`.

```
package Value;
our @ISA = ("Leaf");

sub required_registers {
  my $self = shift;
  my $position = shift;

  $self->{REQ_REG} = ($position eq 'LEFT') ? 1 : 0;
  return $self->REQ_REG;
}
```

El atributo `REQ_REG` se computa para cada una de las sentencias del programa:

```
package STATEMENTS;

sub required_registers {
  my $self = shift;
  my @sts = @{$self};

  for (@sts) {
    $_->required_registers;
  }
}
```

Por supuesto los nodos `ASSIGN` y `PRINT` poseen sus propios métodos `required_registers`.

Una vez computados los requerimientos en registros de cada nódo, la generación de código para un nodo gestiona la asignación de registros usando una cola en la que se guardan los registros disponibles. Se siguen básicamente dos reglas para la traducción de un nodo `Operation`:

1. Realizar primero la traducción del hijo con mayores requerimientos y luego el otro
2. El resultado queda siempre en el registro que ocupa la primera posición en la cola

Hay cuatro casos a considerar: el primero es que el operando derecho sea una hoja. La generación de código para este caso es:

```
package Operation;
our @ISA = ("Binary");
...

sub gen_code {
  my $self = shift;

  if ($self->RIGHT->isa('Leaf')) {
    my $right = $self->RIGHT;
    my $a = $right->VAL;
    my $rightoperand = $right->gen_operand; # valor o dirección
    my $key = $right->key;                  # M, C, etc.
    $self->LEFT->gen_code;
```

```
    Aux::emit($self->nemonic."$key $RSTACK[0], $rightoperand # $a\n");
  }
  ...
}
```

La generación del nemónico se basa en tres métodos:

- El método `nemonic` devuelve el nemónico asociado con el nodo. Por ejemplo, para la clase `TIMES` el código es:

```
sub  nemonic {
  return "MULT";
}
```

- El método `key` devuelve el sufijo que hay que añadir para completar el nemónico, en términos de como sea el operando: `C` para los números, `M` para los identificadores, etc.
- El método `gen_operand` genera el operando. Así para las clases número e identificador su código es:

```
package NUM;
...
sub gen_operand {
  my $self = shift;

  return $self->VAL;
}
```

```
package ID;
...
sub gen_operand {
  my $self = shift;

  return $symbol_table{$self->VAL}->{ADDRESS},
}
```

El resto del código distingue tres casos, según sean r_1, r_2 y el número de registros disponibles. Los dos primeros casos desglosan la posibilidad de que uno de los dos subárboles pueda realizarse con el número de registros disponible ($\text{mín}\{r_1, r_2\} < k$). El tercer caso corresponde a que se necesiten temporales: $\text{mín}\{r_1, r_2\} \geq k$.

```
 1   ...
 2   if ($self->RIGHT->isa('Leaf')) { ...  }
 3   else { # Hijo derecho no es una hoja
 4     my ($t1, $t2) = ($self->LEFT, $self->RIGHT);
 5     my ($r1, $r2) = ($t1->REQ_REG, $t2->REQ_REG);
 6
 7     if ($r1 < Aux::min($r2, $NUM_REG)) {
 8       $t2->gen_code;
 9       my $R = shift @RSTACK;
10       $t1->gen_code;
11       Aux::emit($self->nemonic."R $RSTACK[0], $R\n");
12       push @RSTACK, $R;
13     }
14     ...
15   }
```

En este caso debemos realizar primero la traducción del hijo derecho. Salvando su resultado en `$R`. El registro es retirado de la cola y traducimos el lado izquierdo. El resultado ha quedado en el primer registro de la cola. Emitimos la operación, añadiendo el sufijo `R`, ya que se trata de una operación entre registros y posteriormente devolvemos el registro a la cola.

Ejercicio 33.15.1. *Responda a las siguientes preguntas:*

1. *Si en el código anterior sustituimos la línea 12*

   ```
   push @RSTACK, $R
   ```

   por

   ```
   unshift @RSTACK, $R
   ```

   ¿Seguiría funcionando el código?

2. *¿Podemos asegurar en este subcaso que el código generado para* `$t2` *(línea 8) se ha realizado integramente en los registros?*

Los otros dos casos tienen similar tratamiento:

```
  if ($self->RIGHT->isa('Leaf')) { ...  }
  else { ...
    if ($r1 < Aux::min($r2, $NUM_REG)) { ... }
    elsif (($r1 >= $r2) and ($r2 < $NUM_REG)) {
      $t1->gen_code;
      my $R = shift @RSTACK;
      $t2->gen_code;
      Aux::emit($self->nemonic."R $R, $RSTACK[0]\n");
      unshift @RSTACK, $R;
    }
    elsif (($r1 >= $NUM_REG) and ($r2 >= $NUM_REG)) {
      $t2->gen_code;
      Aux::emit("STORE $T, $RSTACK[0]\n");
      $T++;
      $t1->gen_code;
      $T--;
      Aux::emit($self->nemonic."M $RSTACK[0], $T\n");
    }
  }
}
```

Antes de comenzar a generar el código, la variable `$T` debe ser inicializada a un valor apropiado, de manera que se usen direcciones no ocupadas por los datos. Por ejemplo:

```
local $T =  $final_global_address+length($data);
```

El método `gen_code` sólo debería ser llamado sobre una hoja si se trata de una hoja izquierda (en cuyo caso el número de registros requeridos es uno):

```
package Value;
our @ISA = ("Leaf");
...

sub gen_code {
  my $self = shift;
  my $a = $self->VAL;

  if ($self->REQ_REG == 1) {
```

```
    if (ref($self) eq "NUM") { Aux::emit("LOADC $RSTACK[0], $a\n"); }
    else {
      my $address = $symbol_table{$a}->{ADDRESS};
      Aux::emit("LOADM $RSTACK[0], $address # $a\n");
    }
  }
  else {
    croak("gen_code visita hoja izquierda con REQ_REG = ".$self->REQ_REG);
  }
}
```

La pila de registros es inicializada al número de registros disponibles:

```
use constant LAST_REG => 1;
our @RSTACK = map "R$_", 0..LAST_REG; # Registros disponibles
```

Ejercicio 33.15.2. *Responda a las siguientes preguntas:*

- *¿Cuáles son los requerimientos de registros para un nodo de la clase ASSIGN?*
- *¿Cuáles son los requerimientos de registros para un nodo de la clase PRINT?*
- *¿Se puede lograr la funcionalidad proveída por el método* `required_registers` *usando* `match_and_transform`*?*

  ¿Sería necesario introducir modificaciones en `match_and_transform`*? Si es así, ¿Cuáles?.*

33.15.1. Práctica: Generación de Código

1. Complete la generación de código para la máquina basada en registros. Recuerde que debe escribir el método `required_registers` para las diferentes clases `Value`, `Operation`, `ASSIGN`, `PRINT`, `STATEMENTS`, etc. Asi mismo deberá escribir el método `gen_code` para las diversas clases: `Value`, `Operation`, `ASSIGN`, `PRINT`, `STATEMENTS`, etc. Recuerde que los temporales usados durante la generación de código deben ubicarse en una zona que no esté en uso.

2. En la sección anterior no se consideraba la generación de código para las cadenas. Amplíe y/o modifique el juego de instrucciones como considere conveniente. El siguiente ejemplo de traducción sugiere como puede ser la generación de código para las cadenas:

Fuente	Objeto

`string a,b;`
`a = "hola";`
`b = a;`
`p b`
	`1 LSTRG  R0, 4, 4`
	`2 STORES  0, R0 # a`
	`3 LOADS  R0, 0 # a`
	`4 STORES  2, R0 # b`
	`5 LOADS  R0, 2 # b`
	`6 PRNTS  R0`

   Asuma que los registros pueden contener dos direcciones de memoria (línea 1). La instrucción `LSTRG R0, a, b` carga las constantes (direcciones) `a` y `b` en el registro. La constante `"hola"` ocupa en la posición final en la que se colocan los contenidos de `$data` un desplazamiento de 4 y ocupa 4 palabras. Las instrucción `LOADS R, a` carga las dos palabras en las direcciones `a` y `a+1` en el registro `R`. La instrucción `STORES a, R` se encarga de que las dos palabras en la dirección `a` queden referenciando una cadena igual a la apuntada por el registro `R`. La instrucción `PRNTS` imprime la cadena apuntada por el registro. En una situación mas realista instrucciones como `STORES a, R` y `PRNTS` probablemente serían llamadas a funciones/servicios del sistema o de la librería para soporte en tiempo de ejecución asociada al lenguaje.

3. Se puede mejorar el código generado si hacemos uso de las propiedades algebraicas de los operadores. Por ejemplo, cuando se tiene un operador conmutativo que ha sido asociado a derechas, como ocurre en este programa fuente:

```
$ cat test18.tutu
int a,b,c;

a = a + (b + c)
```

El código producido por el compilador es:

```
LOADM R0, 0 # a
LOADM R1, 1 # b
PLUSM R1, 2 # c
PLUSR R0, R1
STORE  0, R0 # a
```

En este caso, la expresión `a + (b + c)` corresponde a un árbol que casa con el patrón árbol

$$PLUS(ID, t) \text{ and } \{r_t \geq 1\}$$

Donde r_t es el número de registros requeridos por t. En tales casos es posible sacar ventaja de la conmutatividad de la suma y transformar el árbol

$$PLUS(ID, t) \text{ and } \{r_t \geq 1\} \Longrightarrow PLUS(t, ID)$$

Observe que mientras el primer árbol requiere máx$\{2, r_t\}$ registros, el segundo requiere r_t registros, que en general es menor. Esta transformación invierte la traducción:

```
traduce(t)
ADDM $RSTACK[0], dirección de ID
```

que daría lugar a:

```
LOADM R0, 1 # b
PLUSM R0, 2 # c
PLUSM R0, 0 # a
STORE  0, R0 # a
```

la cual usa una instrucción y un registro menos.

Usando `match_and_transform` modifique el generador de código para que, después de la fase de cálculo del número de registros requeridos, aplique esta transformación sobre los nodos conmutativos cuyo hijo izquierdo sea un identificador y su hijo derecho requiera al menos un registro.

33.16. Optimización de Código

Aunque en esta fase se incluyen toda clase de optimizaciones, es aquí donde se hacen las optimizaciones de código dependientes del sistema objeto. Normalmente se recorre el código generado buscando secuencias de instrucciones que se puedan sustituir por otras cuya ejecución sea mas eficiente. El nombre *Peephole optimization* hace alusión a esta especie de "ventana de visión" que se desplaza sobre el código. En nuestro caso, supongamos que disponemos de una instrucción `INC` que permite incrementar eficientemente una expresión. Recorremos el código buscando por un patrón "sumar 1z lo reeemplazamos adecuadamente.

```
package Peephole::Optimization;

sub transform {
  $target = shift;
  $target =~ s/PUSH 1\nPLUS/INC/g;
}
```

Otro ejemplo de optimización peephole consiste en reemplazar las operaciones flotantes de división por una constante por la multiplicación por la inversa de la misma (aquí se pueden introducir diferencias en el resultado, debido a la inexactitud de las operaciones en punto flotante y a que, si se trata de un *compilador cruzado* la aritmética flotante en la máquina en la que se ejecuta el compilador puede ser diferente de la de la máquina que ejecutará el código objeto).

33.16.1. Práctica: Optimización Peephole

1. Optimice el código generado para la máquina de registros sustituyendo las operaciones de multiplicación y división enteras por una constante que sea potencia de dos (de la forma 2^n) por operaciones de desplazamiento. Repase el capítulo de expresiones regulares. Es posible que aquí quiera emplear una sustitución en la que la cadena de reemplazo sea evaluada sobre la marcha. Si es así, repase la sección 31.1.6. La siguiente sesión con el depurador pretende ilustrar la idea:

   ```
   $ perl -de 0
   DB<1> $a = "MUL R2, 16"
   DB<2> $a =~ s/MUL R(\d), (\d+)/($2 == 16)?"SHL R$1, 4":$&/e
   DB<3> p $a
   SHL R2, 4
   DB<5> $a = "MUL R2, 7"
   DB<6> $a =~ s/MUL R(\d), (\d+)/($2 == 16)?"SHL R$1, 4":$&/e
   DB<7> p $a
   MUL R2, 7
   DB<8>
   ```

2. El plegado de constantes realizado durante la optimización de código independiente de la máquina (véase la sección 33.11) es parcial. Si los árboles para el producto se hunden a izquierdas, una expresión como `a = a * 2 * 3` no será plegada, ya que produce un árbol de la forma

   $$t = TIMES(TIMES(a, 2), 3)$$

   Dado que el algoritmo no puede plegar $t/1$ tampoco plegará t. Busque en el código objeto secuencias de multiplicaciones por constantes y abrévielas en una. Haga lo mismo para las restantes operaciones.

3. Dado el siguiente programa:

```
$ cat test14.tutu
int a,b; a = 2; b = a*a+1
```

El código producido por el traductor para la máquina de registros es:

```
1 LOADC R0, 2
2 STORE  0, R0 # a
3 LOADM R0, 0 # a
4 MULTM R0, 0 # a
5 PLUSC R0, 1 # 1
6 STORE  1, R0 # b
```

Se ve que la instrucción de carga `LOADM R0, 0` de la línea 3 es innecesaria por cuanto el contenido de la variable `a` ya está en el registro `R0`, ya que fué cargada en el registro en la línea 2. Nótese que esta hipótesis no es necesariamente cierta si la línea 3 fuera el objetivo de un salto desde otro punto del programa. Esta condición se cumple cuando nos movemos dentro de un *bloque básico*: una secuencia de instrucciones que no contiene instrucciones de salto ni es el objetivo de instrucciones de salto, con la excepción de las instrucciones inicial y final. Mejore el código generado intentando detectar patrones de este tipo, eliminando la operación de carga correspondiente.

Capítulo 34

Análisis Sintáctico Ascendente en JavaScript

34.1. Conceptos Básicos para el Análisis Sintáctico

Suponemos que el lector de esta sección ha realizado con éxito un curso en teoría de autómatas y lenguajes formales. Las siguientes definiciones repasan los conceptos mas importantes.

Definición 34.1.1. *Dado un conjunto A, se define A^* el cierre de Kleene de A como: $A^* = \cup_{n=0}^{\infty} A^n$*

Se admite que $A^0 = \{\epsilon\}$, donde ϵ denota la palabra vacía, esto es la palabra que tiene longitud cero, formada por cero símbolos del conjunto base A.

Definición 34.1.2. *Una gramática G es una cuaterna $G = (\Sigma, V, P, S)$. Σ es el conjunto de terminales. V es un conjunto (disjunto de Σ) que se denomina conjunto de* variables sintácticas *o* categorías gramáticales*, P es un conjunto de pares de $V \times (V \cup \Sigma)^*$. En vez de escribir un par usando la notación $(A, \alpha) \in P$ se escribe $A \to \alpha$. Un elemento de P se denomina producción. Por último, S es un símbolo del conjunto V que se denomina símbolo de arranque.*

Definición 34.1.3. *Dada una gramática $G = (\Sigma, V, P, S)$ y $\mu = \alpha A \beta \in (V \cup \Sigma)^*$ una frase formada por variables y terminales y $A \to \gamma$ una producción de P, decimos que μ deriva en un paso en $\alpha\gamma\beta$. Esto es, derivar una cadena $\alpha A \beta$ es sustituir una variable sintáctica A de V por la parte derecha γ de una de sus reglas de producción. Se dice que μ deriva en n pasos en δ si deriva en $n-1$ pasos en una cadena $\alpha A \beta$ la cual deriva en un paso en δ. Se escribe entonces que $\mu \stackrel{*}{\Longrightarrow} \delta$. Una cadena deriva en 0 pasos en si misma.*

Definición 34.1.4. *Dada una gramática $G = (\Sigma, V, P, S)$ se denota por $L(G)$ o lenguaje generado por G al lenguaje:*

$$L(G) = \{x \in \Sigma^* : S \stackrel{*}{\Longrightarrow} x\}$$

Esto es, el lenguaje generado por la gramática G esta formado por las cadenas de terminales que pueden ser derivados desde el símbolo de arranque.

Definición 34.1.5. *Una derivación que comienza en el símbolo de arranque y termina en una secuencia formada por sólo terminales de Σ se dice* completa*.*

Una derivación $\mu \stackrel{*}{\Longrightarrow} \delta$ en la cual en cada paso $\alpha A x$ la regla de producción aplicada $A \to \gamma$ se aplica en la variable sintáctica mas a la derecha se dice una derivación a derechas

Una derivación $\mu \stackrel{*}{\Longrightarrow} \delta$ en la cual en cada paso $xA\alpha$ la regla de producción aplicada $A \to \gamma$ se aplica en la variable sintáctica mas a la izquierda se dice una derivación a izquierdas

Definición 34.1.6. *Observe que una derivación puede ser representada como un árbol cuyos nodos están etiquetados en $V \cup \Sigma$. La aplicación de la regla de producción $A \to \gamma$ se traduce en asignar como hijos del nodo etiquetado con A a los nodos etiquetados con los símbolos $X_1 \ldots X_n$ que constituyen la frase $\gamma = X_1 \ldots X_n$. Este árbol se llama* árbol sintáctico concreto *asociado con la derivación.*

Definición 34.1.7. *Observe que, dada una frase $x \in L(G)$ una derivación desde el símbolo de arranque da lugar a un árbol. Ese árbol tiene como raíz el símbolo de arranque y como hojas los terminales $x_1 \dots x_n$ que forman x. Dicho árbol se denomina* árbol de análisis sintáctico concreto *de x. Una derivación determina una forma de recorrido del árbol de análisis sintáctico concreto.*

Definición 34.1.8. *Una gramática G se dice ambigua si existe alguna frase $x \in L(G)$ con al menos dos árboles sintácticos. Es claro que esta definición es equivalente a afirmar que existe alguna frase $x \in L(G)$ para la cual existen dos derivaciones a izquierda (derecha) distintas.*

34.1.1. Ejercicio

Dada la gramática con producciones:

program → declarations statements | statements
declarations → declaration ';' declarations | declaration ';'
declaration → INT idlist | STRING idlist
statements → statement ';' statements | statement
statement → ID '=' expression | P expression
expression → term '+' expression | term
term → factor '*' term | factor
factor → '(' expression ')' | ID | NUM | STR
idlist → ID ',' idlist | ID

En esta gramática, Σ esta formado por los caracteres entre comillas simples y los símbolos cuyos identificadores están en mayúsculas. Los restantes identificadores corresponden a elementos de V. El símbolo de arranque es $S = program$.

Conteste a las siguientes cuestiones:

1. Describa con palabras el lenguaje generado.
2. Construya el árbol de análisis sintáctico concreto para cuatro frases del lenguaje.
3. Señale a que recorridos del árbol corresponden las respectivas derivaciones a izquierda y a derecha en el apartado 2.
4. ¿Es ambigua esta gramática?. Justifique su respuesta.

34.2. Ejemplo Simple en Jison

Gramática

```
%%

S   : A
    ;
A   : /* empty */
    | A x
    ;
```

basic2_lex.jison

```
[~/jison/examples/basic2_lex(develop)]$ cat basic2_lex.jison
/* description: Basic grammar that contains a nullable A nonterminal. */

%lex
%%
```

```
\s+               {/* skip whitespace */}
[a-zA-Z_]\w*      {return 'x';}

/lex

%%

S   : A
            { return $1+" identifiers"; }
    ;
A   : /* empty */
            {
               console.log("starting");
               $$ = 0;
            }
    | A x  {
               $$ = $1 + 1;
               console.log($$)
            }
    ;
```

index.html

```
$ cat basic2_lex.html
<!DOCTYPE HTML>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <title>Jison</title>
    <link rel="stylesheet" href="global.css" type="text/css" media="screen" charset="utf-8" />
  </head>
  <body>
    <h1>basic2_lex demo</h1>
    <div id="content">
      <script src="jquery/jquery.js"></script>
      <script src="basic2_lex.js"></script>
      <script src="main.js"></script>
      <p>
        <input type="text" value="x x x x" /> <button>parse</button>
        <span id="output"></span> <!-- Output goes here! -->
      </p>
    </div>
  </body>
</html>
```

Rakefile

```
$ cat Rakefile
# install package:
#     sudo npm install beautifier
#
#   more about beautifier:
#       https://github.com/rickeyski/node-beautifier
```

```
dec "compile the grammar basic2_lex_ugly.jison"
task :default => %w{basic2_lex_ugly.js} do
  sh "mv basic2_lex.js basic2_lex_ugly.js"
  sh "jsbeautify basic2_lex_ugly.js > basic2_lex.js"
  sh "rm -f basic2_lex_ugly.js"
end

file "basic2_lex_ugly.js" => %w{basic2_lex.jison} do
  sh "jison basic2_lex.jison -o basic2_lex.js"
end
```

1. node-beautifier

Véase También

1. JISON
2. Try Jison Examples
3. JavaScript 1.4 LR(1) Grammar 1999.
4. Creating a JavaScript Parser Una implementación de ECMAScript 5.1 usando Jison disponible en GitHub en https://github.com/cjihrig/jsparser. Puede probarse en: http://www.cjihrig.com/development/
5. Bison on JavaScript por Rolando Perez
6. Slogo a language written using Jison
7. List of languages that compile to JS
8. Prototype of a Scannerless, Generalized Left-to-right Rightmost (SGLR) derivation parser for JavaScript

global.css

```
[~/jison/examples/basic2_lex(develop)]$ cat global.css
html *
{
   font-size: large;
   /* The !important ensures that nothing can override what you've set in this style (unless i
   font-family: Arial;
}

.thumb {
    height: 75px;
    border: 1px solid #000;
    margin: 10px 5px 0 0;
  }

h1            { text-align: center; font-size: x-large; }
th, td        { vertical-align: top; text-align: left; }
/* #finaltable  * { color: white; background-color: black; }   */

/* #finaltable table { border-collapse:collapse; } */
/* #finaltable table, td { border:1px solid white; } */
#finaltable:hover td { background-color: blue; }
```

```
tr:nth-child(odd)    { background-color:#eee; }
tr:nth-child(even)    { background-color:#00FF66; }
input        { text-align: right;  border: none;       }     /* Align input to the right  */
textarea     { border: outset; border-color: white;       }
table        { border: inset; border-color: white; }
.hidden      { display: none; }
.unhidden    { display: block; }
table.center { margin-left:auto; margin-right:auto; }
#result      { border-color: red; }
tr.error       { background-color: red; }
pre.output   { background-color: white; }
span.repeated { background-color: red }
span.header { background-color: blue }
span.comments { background-color: orange }
span.blanks { background-color: green }
span.nameEqualValue { background-color: cyan }
span.error { background-color: red }

body
{
 background-color:#b0c4de;  /* blue */
}
```

34.2.1. Véase También

1. JISON
2. Try Jison Examples
3. JavaScript 1.4 LR(1) Grammar 1999.
4. Creating a JavaScript Parser Una implementación de ECAMScript 5.1 usando Jison disponible en GitHub en https://github.com/cjihrig/jsparser. Puede probarse en: http://www.cjihrig.com/development/
5. Slogo a language written using Jison
6. List of languages that compile to JS
7. Prototype of a Scannerless, Generalized Left-to-right Rightmost (SGLR) derivation parser for JavaScript

34.2.2. Práctica: Secuencia de Asignaciones Simples

Modifique este ejemplo para que el lenguaje acepte una secuencia de sentencias de asignación de la forma `ID = NUM` separadas por puntos y comas, por ejemplo `a = 4; b = 4.56; c = -8.57e34`. El analizador retorna un hash/objeto cuyas claves son los identificadores y cuyos valores son los números. Clone el repositorio en https://github.com/crguezl/jison-basic2.

Modifique los analizadores léxico y sintáctico de forma conveniente.

Añada acciones semánticas para que el analizador devuelva una tabla de símbolos con los identificadores y sus valores.

34.3. Ejemplo en Jison: Calculadora Simple

1. Enlace al fork del proyecto jison de crguezl (GitHub)

calculator.jison

```
[~/jison/examples/html_calc_example(develop)]$ cat calculator.jison

/* description: Parses end executes mathematical expressions. */

/* lexical grammar */
%lex
%%

\s+                   /* skip whitespace */
[0-9]+("."[0-9]+)?\b  return 'NUMBER'
"*"                   return '*'
"/"                   return '/'
"-"                   return '-'
"+"                   return '+'
"^"                   return '^'
"!"                   return '!'
"%"                   return '%'
"("                   return '('
")"                   return ')'
"PI"                  return 'PI'
"E"                   return 'E'
<<EOF>>               return 'EOF'
.                     return 'INVALID'

/lex

/* operator associations and precedence */

%left '+' '-'
%left '*' '/'
%left '^'
%right '!'
%right '%'
%left UMINUS

%start expressions

%% /* language grammar */

expressions
    : e EOF
        { typeof console !== 'undefined' ? console.log($1) : print($1);
          return $1; }
    ;

e
    : e '+' e
        {$$ = $1+$3;}
    | e '-' e
        {$$ = $1-$3;}
    | e '*' e
        {$$ = $1*$3;}
```

```
    | e '/' e
        {$$ = $1/$3;}
    | e '^' e
        {$$ = Math.pow($1, $3);}
    | e '!'
        {{
          $$ = (function fact (n) { return n==0 ? 1 : fact(n-1) * n })($1);
        }}
    | e '%'
        {$$ = $1/100;}
    | '-' e %prec UMINUS
        {$$ = -$2;}
    | '(' e ')'
        {$$ = $2;}
    | NUMBER
        {$$ = Number(yytext);}
    | E
        {$$ = Math.E;}
    | PI
        {$$ = Math.PI;}
    ;
```

main.js

```
[~/jison/examples/html_calc_example(develop)]$ cat main.js
$(document).ready(function () {
  $("button").click(function () {
    try {
      var result = calculator.parse($("input").val())
      $("span").html(result);
    } catch (e) {
      $("span").html(String(e));
    }
  });
});
```

calculator.html

```
[~/jison/examples/html_calc_example(develop)]$ cat calculator.html
<!DOCTYPE HTML>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <title>Calc</title>
    <link rel="stylesheet" href="global.css" type="text/css" media="screen" charset="utf-8" />
  </head>
  <body>
    <h1>Calculator demo</h1>
    <div id="content">
      <script src="jquery/jquery.js"></script>
      <script src="calculator.js"></script>
      <script src="main.js"></script>
      <p>
        <input type="text" value="PI*4^2 + 5" /> <button>equals</button>
```

```
        <span></span> <!-- Output goes here! -->
      </p>
    </div>
  </body>
</html>
```

Rakefile

```
[~/jisoncalc(clase)]$ cat Rakefile
task :default => %w{calcugly.js} do
  sh "jsbeautify calcugly.js > calculator.js"
  sh "rm -f calcugly.js"
end

file "calcugly.js" => %w{calculator.jison} do
  sh "jison calculator.jison calculator.l -o calculator.js; mv calculator.js calcugly.js"
end

task :testf do
  sh "open -a firefox test/test.html"
end

task :tests do
  sh "open -a safari test/test.html"
end
```

global.css

```
[~/jison/examples/html_calc_example(develop)]$ cat global.css
html *
{
   font-size: large;
   /* The !important ensures that nothing can override what you've set in this style (unless i
   font-family: Arial;
}

.thumb {
    height: 75px;
    border: 1px solid #000;
    margin: 10px 5px 0 0;
  }

h1            { text-align: center; font-size: x-large; }
th, td        { vertical-align: top; text-align: left; }
/* #finaltable  * { color: white; background-color: black; }   */

/* #finaltable table { border-collapse:collapse; } */
/* #finaltable table, td { border:1px solid white; } */
#finaltable:hover td { background-color: blue; }
tr:nth-child(odd)    { background-color:#eee; }
tr:nth-child(even)    { background-color:#00FF66; }
input        { text-align: right;  border: none;       }     /* Align input to the right  */
textarea     { border: outset; border-color: white;       }
table        { border: inset; border-color: white; }
```

```
.hidden      { display: none; }
.unhidden    { display: block; }
table.center { margin-left:auto; margin-right:auto; }
#result      { border-color: red; }
tr.error       { background-color: red; }
pre.output   { background-color: white; }
span.repeated { background-color: red }
span.header { background-color: blue }
span.comments { background-color: orange }
span.blanks { background-color: green }
span.nameEqualValue { background-color: cyan }
span.error { background-color: red }

body
{
 background-color:#b0c4de;  /* blue */
}
```

test/assert.html

```
$ cat test/assert.js
var output = document.getElementById('output');

function assert( outcome, description) {
  var li = document.createElement('li');
  li.className = outcome ? 'pass' : 'fail';
  li.appendChild(document.createTextNode(description));

  output.appendChild(li);
};
```

test/test.css

```
~/jisoncalc(clase)]$ cat test/test.css
.pass:before {
  content: 'PASS: ';
  color:  blue;
  font-weight: bold;
}

.fail:before {
  content: 'FAIL: ';
  color: red;
  font-weight: bold;

}
```

test/test.html

```
[~/jisoncalc(clase)]$ cat test/test.html
<!DOCTYPE HTML>
<html lang="en">
  <head>
    <meta charset="UTF-8">
```

```
    <title>Testing Our Simple Calculator</title>
    <link rel="stylesheet" href="test.css" />
    <script type="text/javascript" src="../calculator.js"></script>

  </head>
  <body>
    <h1>Testing Our Simple Calculator
    </h1>

    <ul id="output"></ul>
    <script type="text/javascript" src="______.js"></script>

    <script type="text/javascript">
      var r = __________.parse("a = 4*8");
      assert(________________, "a is 4*8");
      assert(_____________, "32 == 4*8");
      r = calculator.parse("a = 4;\nb=a+1;\nc=b*2");
      assert(____________, "4 is the first computed result ");
      assert(_____________, "a is 4");
      assert(_____________, "b is 5");
      assert(_____________, "c is 10");
    </script>
      See the NetTuts+ tutorial at <a href="http://net.tutsplus.com/tutorials/javascript-ajax/
  </body>
</html>
```

34.3.1. Práctica: Calculadora con Listas de Expresiones y Variables

Modifique la calculadora vista en la sección anterior 34.3 para que el lenguaje cumpla los siguientes requisitos:

- Extienda el lenguaje de la calculadora para que admita expresiones de asignación `a = 2*3`
- Extienda el lenguaje de la calculadora para que admita listas de sentencias `a = 2; b = a +1`
- El analizador devuelve la lista de expresiones evaluadas y la tabla de símbolos (con las parejas variable-valor).
- Emita un mensaje de error específico si se intentan modificar las constantes `PI` y `e`.
- Emita un mensaje de error específico si se intenta una división por cero
- Emita un mensaje de error específico si se intenta acceder para lectura a una variable no inicializada `a = c`
- El lenguaje debería admitir expresiones vacías, estos es secuencias consecutivas de puntos y comas sin producir error (`a = 4;;; b = 5`)
- Introduzca pruebas unitarias como las descritas en la sección 25.1 (*Quick Tip: Quick and Easy JavaScript Te*

34.4. Conceptos Básicos del Análisis LR

Los analizadores generados por `jison` entran en la categoría de analizadores *LR*. Estos analizadores construyen una derivación a derechas inversa (o *antiderivación*). De ahí la R en LR (del inglés *rightmost derivation*). El árbol sintáctico es construido de las hojas hacia la raíz, siendo el último paso en la antiderivación la construcción de la primera derivación desde el símbolo de arranque.

Empezaremos entonces considerando las frases que pueden aparecer en una derivación a derechas. Tales frases consituyen el *lenguaje de las formas sentenciales a rderechas* FSD:

Definición 34.4.1. *Dada una gramática* $G = (\Sigma, V, P, S)$ *no ambigua, se denota por* FSD *(lenguaje de las formas Sentenciales a Derechas) al lenguaje de las sentencias que aparecen en una derivación a derechas desde el símbolo de arranque.*

$$FSD = \left\{ \alpha \in (\Sigma \cup V)* : \exists S \underset{RM}{\overset{*}{\Longrightarrow}} \alpha \right\}$$

*Donde la notacion RM indica una derivación a derechas (*rightmost*). Los elementos de* FSD *se llaman "formas sentenciales derechas".*

Dada una gramática no ambigua $G = (\Sigma, V, P, S)$ y una frase $x \in L(G)$ el proceso de antiderivación consiste en encontrar la última derivación a derechas que dió lugar a x. Esto es, si $x \in L(G)$ es porque existe una derivación a derechas de la forma

$$S \overset{*}{\Longrightarrow} yAz \Longrightarrow ywz = x.$$

El problema es averiguar que regla $A \to w$ se aplicó y en que lugar de la cadena x se aplicó. En general, si queremos antiderivar una forma sentencial derecha $\beta\alpha w$ debemos averiguar por que regla $A \to \alpha$ seguir y en que lugar de la forma (después de β en el ejemplo) aplicarla.

$$S \overset{*}{\Longrightarrow} \beta Aw \Longrightarrow \beta\alpha w.$$

La pareja formada por la regla y la posición se denomina *handle*, *mango* o *manecilla* de la forma. Esta denominación viene de la visualización gráfica de la regla de producción como una mano que nos permite escalar hacia arriba en el árbol. Los "dedos" serían los símbolos en la parte derecha de la regla de producción.

Definición 34.4.2. *Dada una gramática* $G = (\Sigma, V, P, S)$ *no ambigua, y dada una forma sentencial derecha* $\alpha = \beta\gamma x$*, con* $x \in \Sigma^*$*, el* mango *o* handle *de* α *es la última producción/posición que dió lugar a* α*:*

$$S \underset{RM}{\overset{*}{\Longrightarrow}} \beta Bx \Longrightarrow \beta\gamma x = \alpha$$

Escribiremos: $handle(\alpha) = (B \to \gamma, \beta\gamma)$. La función *handle* tiene dos componentes: $handle_1(\alpha) = B \to \gamma$ y $handle_2(\alpha) = \beta\gamma$

Si dispusieramos de un procedimiento que fuera capaz de identificar el mango, esto es, de detectar la regla y el lugar en el que se posiciona, tendríamos un mecanismo para construir un analizador. Lo curioso es que, a menudo es posible encontrar un autómata finito que reconoce el lenguaje de los prefijos $\beta\gamma$ que terminan en el mango. Con mas precisión, del lenguaje:

Definición 34.4.3. *El conjunto de* prefijos viables *de una gramática* G *se define como el conjunto:*

$$PV = \left\{ \delta \in (\Sigma \cup V)* : \exists S \underset{RM}{\overset{*}{\Longrightarrow}} \alpha = \beta\gamma x \text{ y } \delta \text{ es un } prefijo \text{ de } handle_2(\alpha) = \beta\gamma \right\}$$

Esto es, es el lenguaje de los prefijos viables es el conjunto de frases que son prefijos de $handle_2(\alpha)) = \beta\gamma$, siendo α una forma sentencial derecha ($\alpha \in FSD$). Los elementos de PV se denominan prefijos viables.

Obsérvese que si se dispone de un autómata que reconoce PV entonces se dispone de un mecanismo para investigar el lugar y el aspecto que pueda tener el mango. Si damos como entrada la sentencia $\alpha = \beta\gamma x$ a dicho autómata, el autómata aceptará la cadena $\beta\gamma$ pero rechazará cualquier extensión del prefijo. Ahora sabemos que el mango será alguna regla de producción de G cuya parte derecha sea un sufijo de $\beta\gamma$.

Definición 34.4.4. *El siguiente autómata finito no determinista puede ser utilizado para reconocer el lenguaje de los prefijos viables PV:*

- *Alfabeto* $= V \cup \Sigma$
- *Los estados del autómata se denominan LR(0) items. Son parejas formadas por una regla de producción de la gramática y una posición en la parte derecha de la regla de producción. Por ejemplo,* $(E \to E + E, 2)$ *sería un LR(0) item para la gramática de las expresiones.*

  Conjunto de Estados:

  $Q = \{(A \to \alpha, n) : A \to \alpha \in P,\ n \leq |\alpha|\}$

  La notación $\mid \alpha \mid$ *denota la longitud de la cadena* $\mid \alpha \mid$*. En vez de la notación* $(A \to \alpha, n)$ *escribiremos:* $A \to \beta_{\uparrow}\gamma = \alpha$*, donde la flecha ocupa el lugar indicado por el número* $n =\mid \beta \mid$ *:*
- *La función de transición intenta conjeturar que partes derechas de reglas de producción son viables. El conjunto de estados actual del NFA representa el conjunto de pares (regla de producción, posición en la parte derecha) que tienen alguna posibilidad de ser aplicadas de acuerdo con la entrada procesada hasta el momento:*

$$\delta(A \to \alpha_{\uparrow} X\beta, X) = A \to \alpha X_{\uparrow}\beta \ \forall X \in V \cup \Sigma$$
$$\delta(A \to \alpha_{\uparrow} B\beta, \epsilon) = B \to_{\uparrow} \gamma \ \forall B \to \gamma \in P$$

- *Estado de arranque: Se añade la "superregla"* $S' \to S$ *a la gramática* $G = (\Sigma, V, P, S)$*. El LR(0) item* $S' \to_{\uparrow} S$ *es el estado de arranque.*
- *Todos los estados definidos (salvo el de muerte) son de aceptación.*

Denotaremos por $LR(0)$ a este autómata. Sus estados se denominan $LR(0) - items$. La idea es que este autómata nos ayuda a reconocer los prefijos viables PV.

Una vez que se tiene un autómata que reconoce los prefijos viables es posible construir un analizador sintáctico que construye una antiderivación a derechas. La estrategia consiste en "alimentar" el autómata con la forma sentencial derecha. El lugar en el que el autómata se detiene, rechazando indica el lugar exacto en el que termina el *handle* de dicha forma.

Ejemplo 34.4.1. *Consideremos la gramática:*

$$S \to a\ S\ b$$
$$S \to \epsilon$$

El lenguaje generado por esta gramática es $L(G) = \{a^n b^n\ :\ n \geq 0\}$ *Es bien sabido que el lenguaje* $L(G)$ *no es regular. La figura 35.1 muestra el* autómata finito no determinista con ϵ-transiciones *(*NFA*) que reconoce los prefijos viables de esta gramática, construido de acuerdo con el algoritmo 35.2.4.*

Véase https://github.com/crguezl/jison-aSb para una implementación en Jison de una variante de esta gramática.

Ejercicio 34.4.1. *Simule el comportamiento del autómata sobre la entrada aabb. ¿Donde rechaza? ¿En que estados está el autómata en el momento del rechazo?. ¿Qué etiquetas tienen? Haga también las trazas del autómata para las entradas aaSbb y aSb. ¿Que antiderivación ha construido el autómata con sus sucesivos rechazos? ¿Que terminales se puede esperar que hayan en la entrada cuando se produce el rechazo del autómata?*

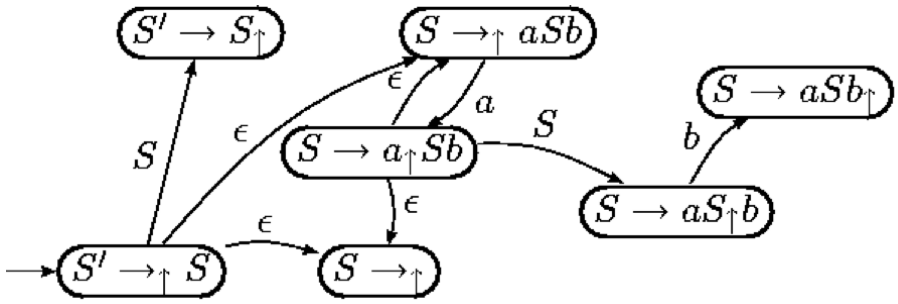


Figura 34.1: NFA que reconoce los prefijos viables

34.5. Construcción de las Tablas para el Análisis SLR

34.5.1. Los conjuntos de Primeros y Siguientes

Repasemos las nociones de conjuntos de *Primeros* y *siguientes*:

Definición 34.5.1. *Dada una gramática* $G = (\Sigma, V, P, S)$ *y una frase* $\alpha \in (V \cup \Sigma)^*$ *se define el conjunto* $FIRST(\alpha)$ *como:*

$$FIRST(\alpha) = \left\{ b \in \Sigma : \alpha \stackrel{*}{\Longrightarrow} b\beta \right\} \cup N(\alpha)$$

donde:

$$N(\alpha) = \begin{cases} \{\epsilon\} & si\ \alpha \stackrel{*}{\Longrightarrow} \epsilon \\ \emptyset & en\ otro\ caso \end{cases}$$

Definición 34.5.2. *Dada una gramática* $G = (\Sigma, V, P, S)$ *y una variable* $A \in V$ *se define el conjunto* $FOLLOW(A)$ *como:*

$$FOLLOW(A) = \left\{ b \in \Sigma : \exists\ S \stackrel{*}{\Longrightarrow} \alpha A b \beta \right\} \cup E(A)$$

donde

$$E(A) = \begin{cases} \{\$\} & si\ S \stackrel{*}{\Longrightarrow} \alpha A \\ \emptyset & en\ otro\ caso \end{cases}$$

Algoritmo 34.5.1. *Construcción de los conjuntos* $FIRST(X)$

1. *Si* $X \in \Sigma$ *entonces* $FIRST(X) = X$
2. *Si* $X \to \epsilon$ *entonces* $FIRST(X) = FIRST(X) \cup \{\epsilon\}$
3. *Si*$X \in V$ *y* $X \to Y_1 Y_2 \cdots Y_k \in P$ *entonces*

$i = 1;$
do
  $FIRST(X) = FIRST(X) \cup FIRST(Y_i) - \{\epsilon\};$
  $i++;$
mientras $(\epsilon \in FIRST(Y_i)\ and\ (i \leq k))$
si $(\epsilon \in FIRST(Y_k)\ and\ i > k)\ FIRST(X) = FIRST(X) \cup \{\epsilon\}$

Este algoritmo puede ser extendido para calcular $FIRST(\alpha)$ para $\alpha = X_1 X_2 \cdots X_n \in (V \cup \Sigma)^*$.

Algoritmo 34.5.2. *Construcción del conjunto* $FIRST(\alpha)$

$i = 1;$
$FIRST(\alpha) = \emptyset;$
do
  $FIRST(\alpha) = FIRST(\alpha) \cup FIRST(X_i) - \{\epsilon\};$
  $i++;$
mientras $(\epsilon \in FIRST(X_i)$ *and* $(i \leq n))$
si $(\epsilon \in FIRST(X_n)$ *and* $i > n)$ $FIRST(\alpha) = FIRST(X) \cup \{\epsilon\}$

Algoritmo 34.5.3. *Construcción de los conjuntos* $FOLLOW(A)$ *para las variables sintácticas* $A \in V$*:*
Repetir los siguientes pasos hasta que ninguno de los conjuntos $FOLLOW$ *cambie:*

1. $FOLLOW(S) = \{\$\}$ *(\$ representa el final de la entrada)*
2. *Si* $A \to \alpha B \beta$ *entonces*

$$FOLLOW(B) = FOLLOW(B) \cup (FIRST(\beta) - \{\epsilon\})$$

3. *Si* $A \to \alpha B$ *o bien* $A \to \alpha B \beta$ *y* $\epsilon \in FIRST(\beta)$ *entonces*

$$FOLLOW(B) = FOLLOW(B) \cup FOLLOW(A)$$

34.5.2. Construcción de las Tablas

Para la construcción de las tablas de un analizador SLR se construye el *autómata finito determinista* (*DFA*) (Q, Σ, δ, q_0) equivalente al NFA presentado en la sección 35.2 usando el *algoritmo de construcción del subconjunto*.

Como recordará, en la construcción del subconjunto, partiendo del estado de arranque q_0 del NFA con ϵ-transiciones se calcula su *clausura* $\overline{\{q_0\}}$ y las clausuras de los conjuntos de estados $\delta(\overline{\{q_0\}}, a)$ a los que transita. Se repite el proceso con los conjuntos resultantes hasta que no se introducen nuevos conjuntos-estado.

La clausura $\overline{A}$ de un subconjunto de estados del autómata A esta formada por todos los estados que pueden ser alcanzados mediante transiciones etiquetadas con la palabra vacía (denominadas ϵ transiciones) desde los estados de A. Se incluyen en $\overline{A}$, naturalmente los estados de A.

$$\overline{A} = \{q \in Q \ / \ \exists q' \in Q \ : \ \hat{\delta}(q', \epsilon) = q\}$$

Aquí $\hat{\delta}$ denota la *función de transición del autómata* extendida a cadenas de Σ^*.

$$\hat{\delta}(q, x) = \begin{cases} \delta(\hat{\delta}(q, y), a) & \text{si } x = ya \\ q & \text{si } x = \epsilon \end{cases} \tag{34.1}$$

En la práctica, y a partir de ahora así lo haremos, se prescinde de diferenciar entre δ y $\hat{\delta}$ usándose indistintamente la notación δ para ambas funciones.

La clausura puede ser computada usando una estructura de pila o aplicando la expresión recursiva dada en la ecuación 35.1.

Para el NFA mostrado en el ejemplo 35.2.1 el DFA construído mediante esta técnica es el que se muestra en la figura 35.2. Se ha utilizado el símbolo `#` como marcador. Se ha omitido el número 3 para que los estados coincidan en numeración con los generados por `jison` (véase el cuadro 35.1).

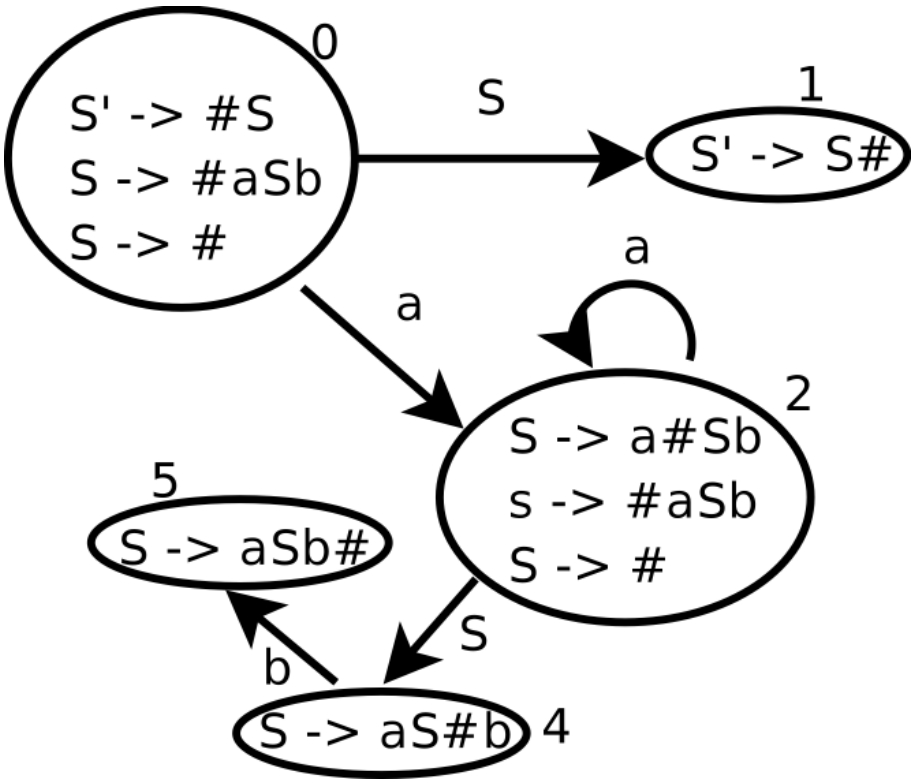


Figura 34.2: DFA equivalente al NFA de la figura 35.1

Un analizador sintáctico LR utiliza una tabla para su análisis. Esa tabla se construye a partir de la tabla de transiciones del DFA. De hecho, la tabla se divide en dos tablas, una llamada *tabla de saltos* o *tabla de gotos* y la otra *tabla de acciones*.

La tabla *goto* de un analizador *SLR* no es más que la tabla de transiciones del autómata DFA obtenido aplicando la construcción del subconjunto al NFA definido en 35.2.4. De hecho es la tabla de transiciones restringida a V (recuerde que el alfabeto del autómata es $V \cup \Sigma$). Esto es,

$$\delta_{|V \times Q} : V \times Q \to Q.$$
$$\text{donde se define } goto(i, A) = \delta(A, I_i)$$

La parte de la función de transiciones del DFA que corresponde a los terminales que no producen rechazo, esto es, $\delta_{|\Sigma \times Q} : \Sigma \times Q \to Q$ se adjunta a una tabla que se denomina *tabla de acciones*. La tabla de acciones es una tabla de doble entrada en los estados y en los símbolos de Σ. Las acciones de transición ante terminales se denominan *acciones de desplazamiento* o (*acciones shift*):

$$\delta_{|\Sigma \times Q} : \Sigma \times Q \to Q$$
$$\text{donde se define } action(i, a) = shift\ \delta(a, I_i)$$

Cuando un estado s contiene un LR(0)-item de la forma $A \to \alpha_{\uparrow}$, esto es, el estado corresponde a un posible rechazo, ello indica que hemos llegado a un final del prefijo viable, que hemos visto α y que, por tanto, es probable que $A \to \alpha$ sea el *handle* de la forma sentencial derecha actual. Por tanto, añadiremos en entradas de la forma (s, a) de la tabla de acciones una acción que indique que hemos encontrado el mango en la posición actual y que la regla asociada es $A \to \alpha$. A una acción de este tipo se la denomina *acción de reducción*.

La cuestión es, ¿para que valores de $a \in \Sigma$ debemos disponer que la acción para (s, a) es de reducción?

$$\text{Se define } action(i, a) = reduce\ A \to \alpha \text{ ¿Pero, para que } a \in \Sigma?$$

Podríamos decidir que ante cualquier terminal $a \in \Sigma$ que produzca un rechazo del autómata, pero podemos ser un poco mas selectivos. No cualquier terminal puede estar en la entrada en el momento en el que se produce la antiderivación o reducción. Observemos que si $A \to \alpha$ es el *handle* de γ es porque:

$$\exists S \underset{RM}{\overset{*}{\Longrightarrow}} \beta A b x \underset{RM}{\overset{*}{\Longrightarrow}} \beta \alpha b x = \gamma$$

Por tanto, cuando estamos reduciendo por $A \to \alpha$ los únicos terminales legales que cabe esperar en una reducción por $A \to \alpha$ son los terminales $b \in FOLLOW(A)$.

$$\text{Se define } action(i, b) = reduce\ A \to \alpha \text{ Para } b \in FOLLOW(A)$$

Dada una gramática $G = (\Sigma, V, P, S)$, podemos construir las tablas de acciones (*action table*) y transiciones (*gotos table*) mediante el siguiente algoritmo:

Algoritmo 34.5.4. *Construcción de Tablas* SLR

1. *Utilizando el Algoritmo de Construcción del Subconjunto, se construye el Autómata Finito Determinista (DFA) $(Q, V \cup \Sigma, \delta, I_0, F)$ equivalente al Autómata Finito No Determinista (NFA) definido en 35.2.4. Sea $C = \{I_1, I_2, \cdots I_n\}$ el conjunto de estados del DFA. Cada estado I_i es un conjunto de LR(0)-items o estados del NFA. Asociemos un índice i con cada conjunto I_i.*

2. *La tabla de* gotos *no es más que la función de transición del autómata restringida a las variables de la gramática:*

$$goto(i, A) = \delta(I_i, A) \text{ para todo } A \in V$$

3. *Las acciones para el estado I_i se determinan como sigue:*

   a) *Si $A \to \alpha_\uparrow a \beta \in I_i$, $\delta(I_i, a) = I_j$, $a \in \Sigma$ entonces:*

   $$action[i][a] = shift\ j$$

   b) *Si $S' \to S_\uparrow \in I_i$ entonces*

   $$action[i][\$] = accept$$

   c) *Para cualquier otro caso de la forma $A \to \alpha_\uparrow \in I_i$ distinto del anterior hacer*

   $$\forall a \in\ FOLLOW(A) :\ action[i][a] = reduce\ A \to \alpha$$

4. *Las entradas de la tabla de acción que queden indefinidas después de aplicado el proceso anterior corresponden a acciones de "error".*

Definición 34.5.3. *Si alguna de las entradas de la tabla resulta multievaluada, decimos que existe un conflicto y que la gramática no es* SLR.

1. *En tal caso, si una de las acciones es de 'reducción" y la otra es de 'desplazamiento", decimos que hay un* conflicto shift-reduce *o* conflicto de desplazamiento-reducción.

2. *Si las dos reglas indican una acción de reducción, decimos que tenemos un* conflicto reduce-reduce *o de* reducción-reducción.

Ejemplo 34.5.1. *Al aplicar el algoritmo 35.3.4 a la gramática 35.2.1*

1	$S \rightarrow a\ S\ b$
2	$S \rightarrow \epsilon$

partiendo del autómata finito determinista que se construyó en la figura 35.2 y calculando los conjuntos de primeros y siguientes

	FIRST	*FOLLOW*
S	*a*, ϵ	*b*, *$*

obtenemos la siguiente tabla de acciones SLR:

	a	*b*	*$*
0	*s2*	*r2*	*r2*
1			*aceptar*
2	*s2*	*r2*	*r2*
4		*s5*	
5		*r1*	*r1*

Las entradas denotadas con s n (s por shift) indican un desplazamiento al estado n, las denotadas con r n (r por reduce o reducción) indican una operación de reducción o antiderivación por la regla n. Las entradas vacías corresponden a acciones de error.

El método de análisis *LALR* usado por `jison` es una extensión del método SLR esbozado aqui. Supone un compromiso entre potencia (conjunto de gramáticas englobadas) y eficiencia (cantidad de memoria utilizada, tiempo de proceso). Veamos como `jison` aplica la construcción del subconjunto a la gramática del ejemplo 35.2.1. Para ello construimos el siguiente programa `jison`:

```
[~/srcPLgrado/aSb(develop)]$ cat -n aSb.jison
     1  %lex
     2  %%
     3  .                { return yytext; }
     4  /lex
     5  %%
     6  P: S             { return $1; }
     7  ;
     8  S: /* empty */   { console.log("empty");    $$ = ''; }
     9     | 'a' S 'b'   { console.log("S -> aSb"); $$ = $1+$2+$3; }
    10  ;
    11  %%
```

y lo compilamos con `jison`. Estas son las opciones disponibles:

```
nereida:[~/PLgradoBOOK(eps)]$ jison --help

Usage: jison [file] [lexfile] [options]

file        file containing a grammar
lexfile     file containing a lexical grammar

Options:
   -o FILE, --outfile FILE       Filename and base module name of the generated parser
   -t, --debug                   Debug mode
   -t TYPE, --module-type TYPE   The type of module to generate (commonjs, amd, js)
   -V, --version                 print version and exit
```

Desafortunadamente carece de la típica opción `-v` que permite generar las tablas de análisis. Podemos intentar usar `bison`, pero, obviamente, `bison` protesta ante la entrada:

```
[~/srcPLgrado/aSb(develop)]$ bison -v aSb.jison
aSb.jison:1.1-4: invalid directive: ‘%lex’
aSb.jison:3.1: syntax error, unexpected identifier
aSb.jison:4.1: invalid character: ‘/’
```

El error es causado por la presencia del analizador léxico empotrado en el fichero `aSb.jison`. Si suprimimos provisionalmente las líneas del analizador léxico empotrado, `bison` es capaz de analizar la gramática:

```
[~/srcPLgrado/aSb(develop)]$ bison -v aSb.jison
[~/srcPLgrado/aSb(develop)]$ ls -ltr | tail -1
-rw-rw-r--  1 casiano  staff    926 19 mar 13:29 aSb.output
```

Que tiene los siguientes contenidos:

```
[~/srcPLgrado/aSb(develop)]$ cat -n aSb.output
     1  Grammar
     2
     3      0 $accept: P $end
     4
     5      1 P: S
     6
     7      2 S: /* empty */
     8      3  | 'a' S 'b'
     9
    10
    11  Terminals, with rules where they appear
    12
    13  $end (0) 0
    14  'a' (97) 3
    15  'b' (98) 3
    16  error (256)
    17
    18
    19  Nonterminals, with rules where they appear
    20
    21  $accept (5)
    22      on left: 0
    23  P (6)
    24      on left: 1, on right: 0
    25  S (7)
    26      on left: 2 3, on right: 1 3
    27
    28
    29  state 0
    30
    31      0 $accept: . P $end
    32
    33      'a'  shift, and go to state 1
    34
    35      $default  reduce using rule 2 (S)
    36
    37      P  go to state 2
    38      S  go to state 3
    39
```

```
40
41  state 1
42
43      3 S: 'a' . S 'b'
44
45      'a'  shift, and go to state 1
46
47      $default  reduce using rule 2 (S)
48
49      S  go to state 4
50
51
52  state 2
53
54      0 $accept: P . $end
55
56      $end  shift, and go to state 5
57
58
59  state 3
60
61      1 P: S .
62
63      $default  reduce using rule 1 (P)
64
65
66  state 4
67
68      3 S: 'a' S . 'b'
69
70      'b'  shift, and go to state 6
71
72
73  state 5
74
75      0 $accept: P $end .
76
77      $default  accept
78
79
80  state 6
81
82      3 S: 'a' S 'b' .
83
84      $default  reduce using rule 3 (S)
```

El contenido del fichero `aSb.output` se muestra en la tabla 35.1. Los números de referencia a las producciones en las acciones de reducción vienen dados por:

```
0: $start -> S $end
1: S -> /* empty */
2: S -> 'a' S 'b'
```

Observe que el final de la entrada se denota por `$end` y el marcador en un LR-item por un punto. Fíjese en el estado 2: En ese estado están también los items

`S -> . 'a' S 'b'` y `S -> .`

sin embargo no se explicitan por que se entiende que su pertenencia es consecuencia directa de aplicar la operación de clausura. Los LR items cuyo marcador no está al principio se denominan *items núcleo*.

Estado 0	Estado 1	Estado 2

`$start -> . S $end`
`'a'shift 2`
`$default`
`reduce 1 (S)`
`S go to state 1`
	`$start -> S . $end`
	`$end shift 3`
		`S -> 'a' . S 'b'`
		`'a'shift 2`
		`$default`
		`reduce 1 (S)`
		`S go to state 4`

Estado 3	Estado 4	Estado 5

`$start -> S $end .`
`$default accept`
	`S -> 'a' S . 'b'`
	`'b'shift 5`
		`S -> 'a' S 'b' .`
		`$default`
		`reduce 2 (S)`

Cuadro 34.1: Tablas generadas por `bison`. El estado 3 resulta de transitar con $

Ejercicio 34.5.1. *Compare la tabla 35.1 resultante de aplicar* `jison` *con la que obtuvo en el ejemplo 35.3.1.*

34.6. Algoritmo de Análisis LR

Asi pues la tabla de transiciones del autómata nos genera dos tablas: la tabla de acciones y la de saltos. El algoritmo de análisis sintáctico *LR* en el que se basa *jison* utiliza una pila y dos tablas para analizar la entrada. Como se ha visto, la tabla de acciones contiene cuatro tipo de acciones:

1. Desplazar (*shift*)
2. Reducir (*reduce*)
3. Aceptar
4. Error

El algoritmo utiliza una pila en la que se guardan los estados del autómata. De este modo se evita tener que "comenzar" el procesado de la forma sentencial derecha resultante después de una reducción (antiderivación).

Algoritmo 34.6.1. *Análizador LR*

```
push(s0);
 b = yylex();
 for( ; ; ;) {
   s = top(0); a = b;
   switch (action[s][a]) {
     case "shift t" :
       push(t);
```

```
      b = yylex();
      break;
    case "reduce A ->alpha" :
      eval(Sub{A -> alpha}->(top(|alpha|-1).attr, ... , top(0).attr));
      pop(|alpha|);
      push(goto[top(0)][A]);
      break;
    case "accept" : return (1);
    default : yyerror("syntax error");
  }
}
```

Como es habitual, $|x|$ denota la longitud de la cadena x. La función `top(k)` devuelve el elemento que ocupa la posición `k` desde el *top* de la pila (esto es, está a profundidad `k`). La función `pop(k)` extrae `k` elementos de la pila. La notación `state.attr` hace referencia al atributo asociado con cada estado. Denotamos por `sub_{reduce A -> alpha}` el código de la acción asociada con la regla $A \to \alpha$.

Todos los analizadores LR comparten, salvo pequeñas exepciones, el mismo algoritmo de análisis. Lo que más los diferencia es la forma en la que construyen las tablas. En `jison` la construcción de las tablas de *acciones* y *gotos* se realiza mediante el algoritmo *LALR*.

34.7. Práctica: Traducción de Infijo a Postfijo

Modifique el programa Jison realizado en la práctica 34.3.1 para traducir de infijo a postfijo. Añada los operadores de comparación e igualdad. Por ejemplo

Infijo	Postfijo
`a = 3+2*4`	`3 2 4 * + &a =`
`b = a == 11`	`a 11 == &b =`

En estas traducciones la notación `&a` indica la dirección de la variable `a` y `a` indica el valor almacenado en la variable `a`.

Añada sentencias `if` ... `then` e `if` ... `then` ... `else`

Para realizar la traducción de estas sentencias añada instrucciones `jmp label` y `jmpz label` (por *jump if zero*) y etiquetas:

Infijo	Postfijo

`a = (2+5)*3;`
`if a == 0 then b = 5 else b = 3;`
`c = b + 1;`
	`2`
	`5`
	`+`
	`3`
	`*`
	`&a`
	`=`
	`a`
	`0`
	`==`
	`jmpz else1`
	`5`
	`&b`
	`=`
	`jmp endif0`
	`:else1`
	`3`
	`&b`
	`=`
	`:endif0`
	`b`
	`1`
	`+`
	`&c`
	`=`

Introduzca pruebas unitarias como las descritas en la sección 25.1 (*Quick Tip: Quick and Easy JavaScript Testing*

```
[~/srcPLgrado/jisoninfix2postfix(master)]$ cat test/test.html
<!DOCTYPE HTML>
<html lang="en">
  <head>
    <meta charset="UTF-8">
    <title>Testing Our Simple Translator</title>
    <link rel="stylesheet" href="test.css" />
    <script type="text/javascript" src="../calculator.js"></script>

  </head>
  <body>
    <h1>Testing Our Simple Translator
    </h1>

    <ul id="output"></ul>
    <script type="text/javascript" src="assert.js"></script>

    <script type="text/javascript">

      var r = calculator.parse("a = 4*8");
      assert( /4\s*8\s*[*]\s*a\s*=\s*/.exec(r), "a is 4*8");

      r = calculator.parse("a=4;b=a+1");
```

```
      r = r.replace(/\s+/g,'');
      var expected = "4a=a1+b=";
      assert( r == expected, "a = 4;\nb=a+1 translated");


      var r = calculator.parse("if a > 0 then b = 1 else b = 2");
      r = r.replace(/\s+/g,'');
      expected = "a 0 > jmpz else1 1 b = jmp endif0 :else1 2 b = :endif0".
                 replace(/\s+/g,'');
      assert( r == expected, "'if a > 0 then b = 1 else b = 2' translated");
    </script>
      See the NetTuts+ tutorial at <a href="http://net.tutsplus.com/tutorials/javascript-ajax/
  </body>
</html>
```

34.8. El módulo Generado por `jison`

34.8.1. Version

En esta sección estudiamos el analizador generado por Jison:

```
[~/Dropbox/src/javascript/PLgrado/jison-aSb(develop)]$ jison --version
0.4.2
```

34.8.2. Gramática Inicial

Veamos el módulo generado por jison para esta gramática:

```
[~/srcPLgrado/aSb(develop)]$ cat aSb.jison
%lex
%%
.               { return yytext; }
/lex
%%
S: /* empty */  { console.log("empty"); }
   | 'a' S 'b'  { console.log("S -> aSb"); }
;
%%
```

34.8.3. Tablas

Esta es la primera parte del parser generado:

```
/* parser generated by jison 0.4.2 */
var aSb = (function() {
    var parser = {
        trace: function trace() {},
        yy: {},
        symbols_: {
            "$accept": 0, /* super-arranque $accept -> S */
            "$end": 1     /* end of input */
            "error": 2, /* numero para el símbolo 'error' */
            "S": 3,     /* numero para el símbolo 'S' */
            "a": 4,
            "b": 5,
```

```
        },
        /* array inverso de terminales */
        terminals_: {   /* numero -> terminal */
            2: "error",
            4: "a",
            5: "b"
        },
        productions_:
        [0,
/* 1 */     [3, 0], /* S : vacio        simbolo,longitud de la parte derecha */
/* 2 */     [3, 3]  /* S : a S b        simbolo,longitud */
        ],
```

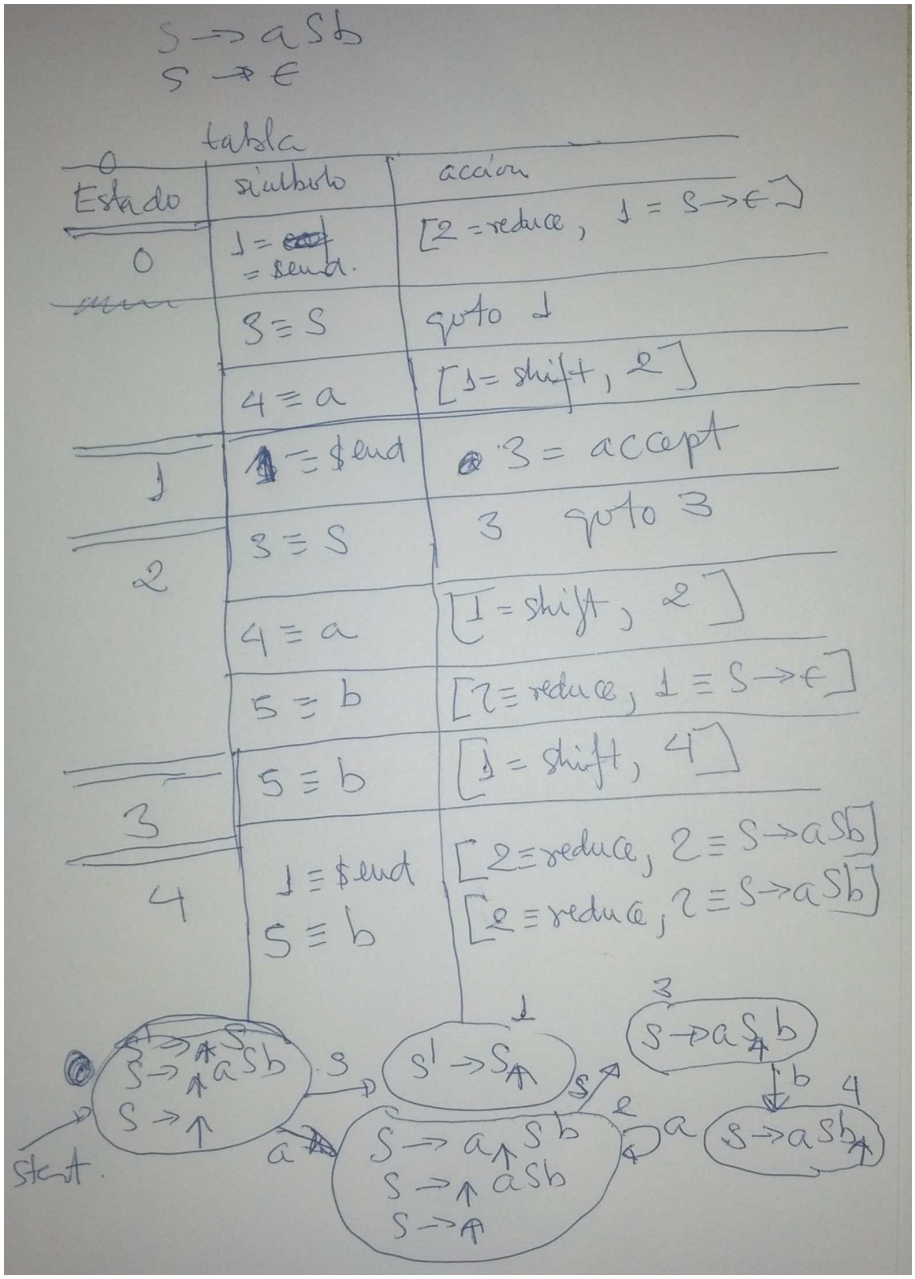


Figura 34.3: DFA construido por Jison

34.8.4. Acciones Semánticas

Cada vez que se produce una acción de reducción esta función es llamada:

```
performAction: function anonymous(yytext, yyleng, yylineno, yy, yystate, $$, _$) {

    var $0 = $$.length - 1;
    switch (yystate) { /* yystate: numero de regla de producción */
        case 1:
            console.log("empty");
            break;
        case 2:
            console.log("S -> aSb");
            break;
    }
},
```

- Parece que cuando se llama a este método `this` refiere a un objeto `yyval`. Este es el punto de llamada a la acción semántica dentro del parser generado por Jison. Puede encontrarse dentro del parser en el caso de un `switch` que corresponde a la acción de reducción:

  ```
  r = this.performAction.call(yyval, yytext, yyleng, yylineno, this.yy, action[1], vstack,
  ```

  El método `call` nos permite invocar una función como si fuera un método de algún otro objeto. Véase la sección 8.7.3.

  Este objeto `yyval` tiene dos atributos: `$` y `_$`.

  - El atributo `$` se corresponde con `$$` de la gramática (atributo de la variable sintactica en la parte izquierda)
  - El atributo `_$` guarda información sobre la posición del último token leído.

- `yytext` parece contener el texto asociado con el token actual
- `yyleng` es la longitud del token actual
- `yylineno` es la línea actual (empezando en 0)
- `yy` es un objeto con dos atributos `lexer` y `parser`
- `yystate` es el estado actual
- `$$` parece ser un array/pila conteniendo los valores de los atributos asociados con los estados de la pila (`vstack` ¿Por value stack?)
- Asi pues `$0` es el índice en `$0` del último elemento de `$$`. Por ejemplo, una acción semántica como:

  ```
  $$ = $1 + $2 + $3;
  ```

  Se traduce por:

  ```
  this.$ = $$[$0 - 2] + $$[$0 - 1] + $$[$0];
  ```

- `_$` Es un array con la información sobre la localización de los simbolos (`lstack` ¿Por location stack?)

34.8.5. Tabla de Acciones y GOTOs

```
        table: [{
/* 0 */     1: [2, 1],    /* En estado 0 viendo $end(1) reducir por S : vacio */
            3: 1,         /* En el estado 0 viendo S(3) ir al estado 1 */
            4: [1, 2]     /* Estado 0 viendo a(4) shift(1) al estado 2 */
        }, {
/* 1 */     1: [3]        /* En 1 viendo $end(1) aceptar */
        }, {
/* 2 */     3: 3,         /* En 2 viendo S ir a 3 */
            4: [1, 2],    /* En 2 viendo a(4) shift a 2 */
            5: [2, 1]     /* En 2 viendo b(5) reducir por regla 1: S -> vacio */
        }, {
/* 3 */     5: [1, 4]     /* En 3 viendo b(5) shift a 4 */
        }, {
/* 4 */     1: [2, 2],    /* En 4  viendo $end(1) reducir(2) por la 2: S -> aSb */
            5: [2, 2]     /* En 4 viendo b(5) reducir por la 2: S-> aSb */
        }],
```

- La tabla es un array de objetos
- El índice de la tabla es el estado. En el ejemplo tenemos 5 estados
- El objeto/hash que es el valor contiene las acciones ante los símbolos.

1. Los atributos/claves son los símbolos, los valores las acciones
2. Las acciones son de dos tipos:
   a) El número del estado al que se transita mediante la tabla `goto` cuando el símbolo es una variable sintactica
   b) Un par `[tipo de acción, estado o regla]`. Si el `tipo de acción` es 1 indica un shift al `estado` con ese número. Si el `tipo de acción` es 2 indica una reducción por la `regla` con ese número.
3. Por ejemplo `table[0]` es

```
{
    1: [2, 1],    /* En estado 0 viendo $end(1) reducir(2) por S : vacio */
    3: 1,         /* En el estado 0 viendo S(3) ir (goto) al estado 1 */
    4: [1, 2]     /* Estado 0 viendo a(4) shift(1) al estado 2 */
}
```

34.8.6. defaultActions

```
defaultActions: {},
```

- `defaultActions` contiene las acciones por defecto.
- Después de la construcción de la tabla, Jison identifica para cada estado la reducción que tiene el conjunto de lookaheads mas grande. Para reducir el tamaño del parser, Jison puede decidir suprimir dicho conjunto y asiganr esa reducción como acción del parser por defecto. Tal reducción se conoce como *reducción por defecto*.
- Esto puede verse en este segmento del código del parser:

```
while (true) {
      state = stack[stack.length - 1];
      if (this.defaultActions[state]) {
```

```
                action = this.defaultActions[state];
            } else {
                if (symbol === null || typeof symbol == "undefined") {
                    symbol = lex();
                }
                action = table[state] && table[state][symbol];
            }
            ...
        }
```

34.8.7. Reducciones

```
parse: function parse(input) {
    ...
    while (true) {
        state = stack[stack.length - 1];
        if (this.defaultActions[state]) {
            action = this.defaultActions[state];
        } else {
            if (symbol === null || typeof symbol == "undefined") {
                symbol = lex(); /* obtener siguiente token */
            }
            action = table[state] && table[state][symbol];
        }
        if (typeof action === "undefined" || !action.length || !action[0]) {
          ... // error
        }
        if (action[0] instanceof Array && action.length > 1) {
            throw new Error("Parse Error: multiple actions possible at state: ..."
        }
        switch (action[0]) {
            case 1:                                     // shift
                ...
                break;
            case 2:                                     // reduce
                len = this.productions_[action[1]][1]; // longitud de la producción
                yyval.$ = vstack[vstack.length - len];
                yyval._$ = {                            // datos de la posición
                    first_line: lstack[lstack.length - (len || 1)].first_line,
                    last_line: lstack[lstack.length - 1].last_line,
                    first_column: lstack[lstack.length - (len || 1)].first_column,
                    last_column: lstack[lstack.length - 1].last_column
                };
                ...
                r = this.performAction.call(yyval, yytext, yyleng, yylineno, this.yy, action[1
                if (typeof r !== "undefined") {
                    return r; /* un return de algo distinto de undefined nos saca del parser *
                }
                if (len) {                                  /* retirar de las pilas */
                    stack = stack.slice(0, - 1 * len * 2);  /* simbolo, estado, simbolo, estad
                    vstack = vstack.slice(0, - 1 * len);    /* retirar atributos */
                    lstack = lstack.slice(0, - 1 * len);    /* retirar localizaciones */
                }
                stack.push(this.productions_[action[1]][0]); /* empujemos el símbolo */
```

```
                vstack.push(yyval.$);                       /* empujemos valor semantico */
                lstack.push(yyval._$);                      /* empujemos localización */
                newState = table[stack[stack.length - 2]][stack[stack.length - 1]];
                stack.push(newState);                       /* empujemos goto[top][A]*/
                break;
            case 3: // accept
                return true;
        }
    }
    return true;
}
```

34.8.8. Desplazamientos/Shifts

```
parse: function parse(input) {
    ...
    while (true) {
        state = stack[stack.length - 1];     /* estado en el top de la pila */
        if (this.defaultActions[state]) {    /* definida la acción por defecto? */
            action = this.defaultActions[state];
        } else {
            if (symbol === null || typeof symbol == "undefined") {
                symbol = lex();              /* obtener token */
            }
            action = table[state] && table[state][symbol]; /* obtener la acción para el estado
        }
        if (typeof action === "undefined" || !action.length || !action[0]) {
            ... /* error */
        }
        if (action[0] instanceof Array && action.length > 1) {
            throw new Error("Parse Error: multiple actions possible at state: " + state + ", t
        }
        switch (action[0]) {
            case 1:
                stack.push(symbol);                 /* empujamos token */
                vstack.push(this.lexer.yytext);     /* empujamos el atributo del token */
                lstack.push(this.lexer.yylloc);     /* salvamos la localización del token */
                stack.push(action[1]);              /* salvamos el estado */
                symbol = null;
                if (!preErrorSymbol) {              /* si no hay errores ... */
                    yyleng = this.lexer.yyleng;     /* actualizamos los atributos */
                    yytext = this.lexer.yytext;     /* del objeto */
                    yylineno = this.lexer.yylineno;
                    yyloc = this.lexer.yylloc;
                    if (recovering > 0) recovering--; /* las cosas van mejor si hubieron error
                } else {
                    symbol = preErrorSymbol;
                    preErrorSymbol = null;
                }
                break;
            case 2:
                ...
                break;
            case 3:
```

```
                return true;
            }
        }
        return true;
    }
```

34.8.9. Manejo de Errores

```
    while (true) {
        state = stack[stack.length - 1];
        if (this.defaultActions[state]) { action = this.defaultActions[state]; }
        else {
            if (symbol === null || typeof symbol == "undefined") { symbol = lex(); }
            action = table[state] && table[state][symbol];
        }
        if (typeof action === "undefined" || !action.length || !action[0]) {
            var errStr = "";
            if (!recovering) { /* recovering = en estado de recuperación de un error */
                expected = [];                      /* computemos los tokens esperados */
                for (p in table[state])             /* si el estado "state" transita con p */
                  if (this.terminals_[p] && p > 2) { /* y "p" es un terminal no especial */
                      expected.push("'" + this.terminals_[p] + "'"); /* entonces es esperado */
                  }
                if (this.lexer.showPosition) { /* si esta definida la función showPosition */
                    errStr = "Parse error on line " + (yylineno + 1) +
                             ":\n" + this.lexer.showPosition() +
                             "\nExpecting " + expected.join(", ") +
                             ", got '" +
                             (this.terminals_[symbol] || symbol) + /* terminals_ es el array inver
                             "'";                                  /* numero -> terminal
                } else { /* ¡monta la cadena como puedas! */
                    errStr = "Parse error on line " + (yylineno + 1) +
                             ": Unexpected " +
                             (symbol == 1 ? "end of input" : "'" +
                             (this.terminals_[symbol] || symbol) + "'");
                }
                this.parseError(errStr, {    /* genera la excepción */
                    text: this.lexer.match,  /* hash/objeto conteniendo los detalles del */
                    token: this.terminals_[symbol] || symbol,                  /* error */
                    line: this.lexer.yylineno,
                    loc: yyloc,
                    expected: expected
                });
            }
        }
        if (action[0] instanceof Array && action.length > 1) {
            throw new Error("Parse Error: multiple actions possible at state: " + state + ", token
        }
        ...
    }
```

La función `parseError` genera una excepción:

```
        parseError: function parseError(str, hash) {
```

```
        throw new Error(str); /* El hash contiene info sobre el error: token, linea, etc.
    },
```

- `parseError` es llamada cada vez que ocurre un error sintáctico. `str` contiene la cadena con el mensaje de error del tipo: `Expecting something, got other thing'`. `hash` contiene atributos como `expected`: el array de tokens esperados; `line` la línea implicada, `loc` una descripción de la localización detallada del punto/terminal en el que ocurre el error; etc.

34.8.10. Analizador Léxico

El analizador léxico:

```
/* generated by jison-lex 0.1.0 */
var lexer = (function() {
    var lexer = {
        EOF: 1,
        parseError: function parseError(str, hash) { /* manejo de errores léxicos */ },
        setInput: function(input) { /* inicializar la entrada para el analizadorléxico */},
        input: function() { /* ... */ },
        unput: function(ch) { /* devolver al flujo de entrada */ },
        more: function() { /* ... */ },
        less: function(n) { /* ... */ },
        pastInput: function() { /* ... */ },
        upcomingInput: function() { /* ... */ },
        showPosition: function() { /* ... */ },
        next: function() {
                if (this.done) { return this.EOF; }
                if (!this._input) this.done = true;

                var token, match, tempMatch, index, col, lines;
                if (!this._more) { this.yytext = ''; this.match = ''; }
                var rules = this._currentRules();
                for (var i = 0; i < rules.length; i++) {
                    tempMatch = this._input.match(this.rules[rules[i]]);
                    if (tempMatch && (!match || tempMatch[0].length > match[0].length)) {
                        match = tempMatch;
                        index = i;
                        if (!this.options.flex) break;
                    }
                }
                if (match) {
                    lines = match[0].match(/(?:\r\n?|\n).*/g);
                    if (lines) this.yylineno += lines.length;
                    this.yylloc = {
                        first_line: this.yylloc.last_line,
                        last_line: this.yylineno + 1,
                        first_column: this.yylloc.last_column,
                        last_column:
                          lines ? lines[lines.length - 1].length -
                                  lines[lines.length - 1].match(/\r?\n?/)[0].length
                                  :
                                  this.yylloc.last_column + match[0].length
                    };
                    this.yytext += match[0];
```

```
            this.match += match[0];
            this.matches = match;
            this.yyleng = this.yytext.length;
            if (this.options.ranges) {
                this.yylloc.range = [this.offset, this.offset += this.yyleng];
            }
            this._more = false;
            this._input = this._input.slice(match[0].length);
            this.matched += match[0];
            token = this.performAction.call(
                        this,
                        this.yy,
                        this,
                        rules[index],
                        this.conditionStack[this.conditionStack.length - 1]
                    );
            if (this.done && this._input) this.done = false;
            if (token) return token;
            else return;
        }
        if (this._input === "") { return this.EOF; }
        else {
            return this.parseError(
                     'Lexical error on line ' + (this.yylineno + 1) +
                      '. Unrecognized text.\n' + this.showPosition(),
                      { text: "", token: null, line: this.yylineno }
                   );
        }
    },
lex: function lex() {
    var r = this.next();
    if (typeof r !== 'undefined') {
        return r;
    } else {
        return this.lex();
    }
},
begin: function begin(condition) { },
popState: function popState() { },
_currentRules: function _currentRules() { },
topState: function() { },
pushState: function begin(condition) { },
options: {},
performAction: function anonymous(yy, yy_, $avoiding_name_collisions, YY_START)
{
    var YYSTATE = YY_START;
    switch ($avoiding_name_collisions) {
        case 0:
            return yy_.yytext;
            break;
    }
},
rules: [/^(?:.)/], /* lista de expresiones regulares */
```

```
        conditions: { /* ... */ }
  }
};
```

34.8.11. Exportación

Si no ha sido exportado ya ...

```
if (typeof require !== 'undefined' && typeof exports !== 'undefined') {
    exports.parser = aSb;         /* hacemos accesible el objeto aSb  */
    exports.Parser = aSb.Parser;
```

El objeto `aSb.Parser` representa al parser. Este es el código que lo crea.

```
      function Parser() {
          this.yy = {};
      }
      Parser.prototype = parser;
      parser.Parser = Parser;
      return new Parser;
  })();
```

También se exporta una función `parse`:

```
    exports.parse = function() {
        return aSb.parse.apply(aSb, arguments);
    };
```

y una función `main`:

```
    exports.main = function commonjsMain(args) {
        if (!args[1]) {
            console.log('Usage: ' + args[0] + ' FILE');
            process.exit(1);
        }
        var source = require('fs').readFileSync(require('path').normalize(args[1]), "utf8");
        return exports.parser.parse(source);
    };
    if (typeof module !== 'undefined' && require.main === module) {
        exports.main(process.argv.slice(1));
    }
}
```

Esto permite ejecutar el módulo directamente:

```
[~/Dropbox/src/javascript/PLgrado/jison-aSb(develop)]$ node aSb.js input.ab
empty
S -> aSb
S -> aSb
[~/Dropbox/src/javascript/PLgrado/jison-aSb(develop)]$ cat input.ab
aabb

~/Dropbox/src/javascript/PLgrado/jison-aSb(develop)]$ node debug aSb.js input.ab
< debugger listening on port 5858
connecting... ok
break in aSb.js:2
  1 /* parser generated by jison 0.4.2 */
```

```
  2 var aSb = (function() {
  3     var parser = {
  4         trace: function trace() {},
debug> n
break in aSb.js:390
 388     return new Parser;
 389 })();
 390 if (typeof require !== 'undefined' && typeof exports !== 'undefined') {
 391     exports.parser = aSb;
 392     exports.Parser = aSb.Parser;

debug> repl
Press Ctrl + C to leave debug repl
>
> typeof require
'function'
> typeof exports
'object'
> aSb
{ yy: {} }
> aSb.Parser
[Function]
^C
debug> sb(396)
 395     };
debug> c
break in aSb.js:396
 394         return aSb.parse.apply(aSb, arguments);
 395     };
*396     exports.main = function commonjsMain(args) {
 397         if (!args[1]) {
 398             console.log('Usage: ' + args[0] + ' FILE');
debug> n
break in aSb.js:404
 402         return exports.parser.parse(source);
 403     };
 404     if (typeof module !== 'undefined' && require.main === module) {
 405         exports.main(process.argv.slice(1));
 406     }
debug> repl
Press Ctrl + C to leave debug repl
> process.argv.slice(1)
[ '/Users/casiano/Dropbox/src/javascript/PLgrado/jison-aSb/aSb.js',
  'input.ab' ]
> typeof module
'object'
> require.main
{ id: '.',
  exports:
   { parser: { yy: {} },
     Parser: [Function],
     parse: [Function],
     main: [Function] },
  parent: null,
```

```
  filename: '/Users/casiano/Dropbox/src/javascript/PLgrado/jison-aSb/aSb.js',
  loaded: false,
  children: [],
  paths:
   [ '/Users/casiano/Dropbox/src/javascript/PLgrado/jison-aSb/node_modules',
     '/Users/casiano/Dropbox/src/javascript/PLgrado/node_modules',
     '/Users/casiano/Dropbox/src/javascript/node_modules',
     '/Users/casiano/Dropbox/src/node_modules',
     '/Users/casiano/Dropbox/node_modules',
     '/Users/casiano/node_modules',
     '/Users/node_modules',
     '/node_modules' ] }
^C
debug> n
break in aSb.js:405
 403     };
 404     if (typeof module !== 'undefined' && require.main === module) {
 405         exports.main(process.argv.slice(1));
 406     }
 407 }
debug> n
< empty
< S -> aSb
< S -> aSb
break in aSb.js:409
 407 }
 408
 409 });
debug> c
program terminated
debug>
```

34.9. Precedencia y Asociatividad

Recordemos que si al construir la tabla LALR, alguna de las entradas de la tabla resulta multievaluada, decimos que existe un conflicto. Si una de las acciones es de 'reducción" y la otra es de 'desplazamiento", se dice que hay un *conflicto shift-reduce* o *conflicto de desplazamiento-reducción*. Si las dos reglas indican una acción de reducción, decimos que tenemos un *conflicto reduce-reduce* o de *reducción-reducción*. En caso de que no existan indicaciones específicas *jison* resuelve los conflictos que aparecen en la construcción de la tabla utilizando las siguientes reglas:

1. Un conflicto *reduce-reduce* se resuelve eligiendo la producción que se listó primero en la especificación de la gramática.

2. Un conflicto *shift-reduce* se resuelve siempre en favor del *shift*

Las declaraciones de precedencia y asociatividad mediante las palabras reservadas `%left` , `%right` , `%nonassoc` se utilizan para modificar estos criterios por defecto. La declaración de `token` s mediante la palabra reservada `%token` no modifica la precedencia. Si lo hacen las declaraciones realizadas usando las palabras `left` , `right` y `nonassoc` .

1. Los *tokens* declarados en la misma línea tienen igual precedencia e igual asociatividad. La precedencia es mayor cuanto mas abajo su posición en el texto. Así, en el ejemplo de la calculadora en la sección 35.1, el *token* `*` tiene mayor precedencia que `+` pero la misma que `/`.

2. La precedencia de una regla $A \rightarrow \alpha$ se define como la del terminal mas a la derecha que aparece en α. En el ejemplo, la producción

```
expr : expr '+' expr
```

   tiene la precedencia del *token* `+`.

3. Para decidir en un conflicto *shift-reduce* se comparan la precedencia de la regla con la del terminal que va a ser desplazado. Si la de la regla es mayor se reduce si la del *token* es mayor, se desplaza.

4. Si en un conflicto *shift-reduce* ambos la regla y el terminal que va a ser desplazado tiene la misma precedencia *jison* considera la asociatividad, si es asociativa a izquierdas, reduce y si es asociativa a derechas desplaza. Si no es asociativa, genera un mensaje de error.
   Obsérvese que, en esta situación, la asociatividad de la regla y la del *token* han de ser por fuerza, las mismas. Ello es así, porque en *jison* los *tokens* con la misma precedencia se declaran en la misma línea y sólo se permite una declaración por línea.

5. *Por tanto es imposible declarar dos* tokens *con diferente asociatividad y la misma precedencia.*

6. Es posible modificar la precedencia "natural" de una regla, calificándola con un *token* específico. para ello se escribe a la derecha de la regla `prec token`, donde `token` es un *token* con la precedencia que deseamos. Vea el uso del *token* `dummy` en el siguiente ejercicio.

Para ilustrar las reglas anteriores usaremos el siguiente programa `jison`:

```
$ cat -n Precedencia.yp
     1  %token NUMBER
     2  %left '@'
     3  %right '&'  dummy
     4  %%
     5  list
     6      :
     7      | list '\n'
     8      | list e
     9      ;
    10
    11  e : NUMBER
    12    | e '&' e
    13    | e '@' e %prec dummy
    14    ;
    15
    16  %%
```

El código del programa cliente es el siguiente:

```
$ cat -n useprecedencia.pl
cat -n useprecedencia.pl
     1  #!/usr/bin/perl -w
     2  use strict;
     3  use Precedencia;
     4
     5  sub Error {
     6    exists $_[0]->YYData->{ERRMSG}
     7    and do {
     8      print $_[0]->YYData->{ERRMSG};
```

```
 9      delete $_[0]->YYData->{ERRMSG};
10      return;
11    };
12    print "Syntax error.\n";
13  }
14
15  sub Lexer {
16    my($parser)=shift;
17
18    defined($parser->YYData->{INPUT})
19    or  $parser->YYData->{INPUT} = <STDIN>
20    or  return('',undef);
21
22    $parser->YYData->{INPUT}=~s/^[ \t]//;
23
24    for ($parser->YYData->{INPUT}) {
25        s/^([0-9]+(?:\.[0-9]+)?)//
26                and return('NUMBER',$1);
27        s/^(.)//s
28                and return($1,$1);
29    }
30  }
31
32  my $debug_level = (@ARGV)? oct(shift @ARGV): 0x1F;
33  my $parser = Precedencia->new();
34  $parser->YYParse( yylex => \&Lexer, yyerror => \&Error, yydebug => $debug_level );
```

Observe la llamada al analizador en la línea 34. Hemos añadido el parámetro con nombre *yydebug* con argumento `yydebug => $debug_level` (véase la sección 35.6 para ver los posibles valores de depuración).

Compilamos a continuación el módulo usando la opción `-v` para producir información sobre los conflictos y las tablas de salto y de acciones:

```
yapp -v -m Precedencia Precedencia.yp
$ ls -ltr |tail -2
-rw-r--r--  1 lhp lhp   1628 2004-12-07 13:21 Precedencia.pm
-rw-r--r--  1 lhp lhp   1785 2004-12-07 13:21 Precedencia.output
```

La opción `-v` genera el fichero `Precedencia.output` el cual contiene información detallada sobre el autómata:

```
$ cat -n Precedencia.output
     1  Conflicts:
     2  ----------
     3  Conflict in state 8 between rule 6 and token '@' resolved as reduce.
     4  Conflict in state 8 between rule 6 and token '&' resolved as shift.
     5  Conflict in state 9 between rule 5 and token '@' resolved as reduce.
     6  Conflict in state 9 between rule 5 and token '&' resolved as shift.
     7
     8  Rules:
     9  ------
    10  0:      $start -> list $end
    11  1:      list -> /* empty */
    12  2:      list -> list '\n'
    13  3:      list -> list e
```

```
14  4:      e -> NUMBER
15  5:      e -> e '&' e
16  6:      e -> e '@' e
17  ...
```

¿Porqué se produce un conflicto en el estado 8 entre la regla 6 (`e -> e '@' e`) y el terminal `'@'`?. Editando el fichero `Precedencia.output` podemos ver los contenidos del estado 8:

```
85  State 8:
86
87          e -> e . '&' e  (Rule 5)
88          e -> e . '@' e  (Rule 6)
89          e -> e '@' e .  (Rule 6)
90
91          '&'     shift, and go to state 7
92
93          $default        reduce using rule 6 (e)
```

El item de la línea 88 indica que debemos desplazar, el de la línea 89 que debemos reducir por la regla 6. ¿Porqué `jison` resuelve el conflicto optando por reducir? ¿Que prioridad tiene la regla 6? ¿Que asociatividad tiene la regla 6? La declaración en la línea 13 modifica la precedencia y asociatividad de la regla:

```
13    | e '@' e %prec dummy
```

de manera que la regla pasa a tener la precedencia y asociatividad de `dummy`. Recuerde que habíamos declarado `dummy` como asociativo a derechas:

```
 2  %left '@'
 3  %right '&'  dummy
```

¿Que ocurre? Que `dummy` tiene mayor prioridad que `'@'` y por tanto la regla tiene mayor prioridad que el terminal: por tanto se reduce.

¿Que ocurre cuando el terminal en conflicto es `'&'`? En ese caso la regla y el terminal tienen la misma prioridad. Se hace uso de la asociatividad a derechas que indica que el conflicto debe resolverse desplazando.

Ejercicio 34.9.1. *Explique la forma en que* `jison` *resuelve los conflictos que aparecen en el estado 9. Esta es la información sobre el estado 9:*

```
State 9:

e -> e . '&' e (Rule 5)
e -> e '&' e . (Rule 5)
e -> e . '@' e (Rule 6)
'&'shift, and go to state 7
$default reduce using rule 5 (e)
```

Veamos un ejemplo de ejecución:

```
$ ./useprecedencia.pl
----------------------------------------
In state 0:
Stack:[0]
Don't need token.
Reduce using rule 1 (list,0): Back to state 0, then go to state 1.
```

Lo primero que ocurre es una reducción por la regla en la que `list` produce vacío. Si miramos el estado 0 del autómata vemos que contiene:

```
20 State 0:
21
22   $start -> . list $end (Rule 0)
23
24   $default  reduce using rule 1 (list)
25
26   list  go to state 1
```

A continuación se transita desde 0 con `list` y se consume el primer terminal:

```
----------------------------------------
In state 1:
Stack:[0,1]
2@3@4
Need token. Got >NUMBER<
Shift and go to state 5.
----------------------------------------
In state 5:
Stack:[0,1,5]
Don't need token.
Reduce using rule 4 (e,1): Back to state 1, then go to state 2.
----------------------------------------
```

En el estado 5 se reduce por la regla `e -> NUMBER`. Esto hace que se retire el estado 5 de la pila y se transite desde el estado 1 viendo el símbolo `e`:

```
In state 2:
Stack:[0,1,2]
Need token. Got >@<
Shift and go to state 6.
----------------------------------------
In state 6:
Stack:[0,1,2,6]
Need token. Got >NUMBER<
Shift and go to state 5.
----------------------------------------
In state 5:
Stack:[0,1,2,6,5]
Don't need token.
Reduce using rule 4 (e,1): Back to state 6, then go to state 8.
----------------------------------------
In state 8:
Stack:[0,1,2,6,8]
Need token. Got >@<
Reduce using rule 6 (e,3): Back to state 1, then go to state 2.
----------------------------------------
...
Accept.
```

Obsérvese la resolución del conflicto en el estado 8

La presencia de conflictos, aunque no siempre, en muchos casos es debida a la introducción de ambiguedad en la gramática. Si el conflicto es de desplazamiento-reducción se puede resolver explicitando alguna regla que rompa la ambiguedad. Los conflictos de reducción-reducción suelen producirse por un diseño erróneo de la gramática. En tales casos, suele ser mas adecuado modificar la gramática.

34.10. Depuración en `jison`

34.11. Construcción del Árbol Sintáctico

El siguiente ejemplo usa `jison` para construir el árbol sintáctico de una expresión en infijo:

34.12. Esquemas de Traducción

Un *esquema de traducción* es una gramática independiente del contexto en la cuál se han asociado atributos a los símbolos de la gramática. Un atributo queda caracterizado por un identificador o nombre y un tipo o clase. Además se han insertado acciones, esto es, código Perl/Python/C, ... en medio de las partes derechas. En ese código es posible referenciar los atributos de los símbolos de la gramática como variables del lenguaje subyacente.

Recuerde que el orden en que se evalúan los fragmentos de código es el de un recorrido primero-profundo del árbol de análisis sintáctico. Mas específicamente, considerando a las acciones como hijos-hoja del nodo, el recorrido que realiza un esquema de traducción es:

34.13. Definición Dirigida por la Sintáxis

Una *definición dirigida por la sintáxis* es un pariente cercano de los esquemas de traducción. En una definición dirigida por la sintáxis una gramática $G = (V, \Sigma, P, S)$ se aumenta con nuevas características:

- A cada símbolo $S \in V \cup \Sigma$ de la gramática se le asocian cero o mas atributos. Un atributo queda caracterizado por un identificador o nombre y un tipo o clase. A este nivel son *atributos formales*, como los parámetros formales, en el sentido de que su realización se produce cuando el nodo del árbol es creado.

- A cada regla de producción $A \to X_1 X_2 \ldots X_n \in P$ se le asocian un conjunto de *reglas de evaluación de los atributos* o *reglas semánticas* que indican que el atributo en la parte izquierda de la regla semántica depende de los atributos que aparecen en la parte derecha de la regla. El atributo que aparece en la parte izquierda de la regla semántica puede estar asociado con un símbolo en la parte derecha de la regla de producción.

- Los atributos de cada símbolo de la gramática $X \in V \cup \Sigma$ se dividen en dos grupos disjuntos: *atributos sintetizados* y *atributos heredados*. Un atributo de X es un *atributo heredado* si depende de atributos de su padre y hermános en el árbol. Un *atributo sintetizado* es aquél tal que el valor del atributo depende de los valores de los atributos de los hijos, es decir en tal caso X ha de ser una variable sintáctica y los atributos en la parte derecha de la regla semántica deben ser atributos de símbolos en la parte derecha de la regla de producción asociada.

- Los atributos predefinidos se denominán *atributos intrínsecos*. Ejemplos de atributos intrínsecos son los atributos sintetizados de los terminales, los cuáles se han computado durante la fase de análisis léxico. También son atributos intrínsecos los atributos heredados del símbolo de arranque, los cuales son pasados como parámetros al comienzo de la computación.

La diferencia principal con un esquema de traducción está en que no se especifica el orden de ejecución de las reglas semánticas. Se asume que, bien de forma manual o automática, se resolverán las dependencias existentes entre los atributos determinadas por la aplicación de las reglas semánticas, de manera que serán evaluados primero aquellos atributos que no dependen de ningún otro, despues los que dependen de estos, etc. siguiendo un esquema de ejecución que viene guiado por las dependencias existentes entre los datos.

Aunque hay muchas formas de realizar un evaluador de una definición dirigida por la sintáxis, conceptualmente, tal evaluador debe:

1. Construir el árbol de análisis sintáctico para la gramática y la entrada dadas.

2. Analizar las reglas semánticas para determinar los atributos, su clase y las dependencias entre los mismos.

3. Construir el *grafo de dependencias* de los atributos, el cual tiene un nodo para cada ocurrencia de un atributo en el árbol de análisis sintáctico etiquetado con dicho atributo. El grafo tiene una arista entre dos nodos si existe una dependencia entre los dos atributos a través de alguna regla semántica.

4. Supuesto que el grafo de dependencias determina un *orden parcial* (esto es cumple las propiedades reflexiva, antisimétrica y transitiva) construir un *orden topológico* compatible con el orden parcial.

5. Evaluar las reglas semánticas de acuerdo con el orden topológico.

Una definición dirigida por la sintáxis en la que las reglas semánticas no tienen efectos laterales se denomina una *gramática atribuída*.

Si la definición dirigida por la sintáxis puede ser realizada mediante un esquema de traducción se dice que es *L-atribuída*. Para que una definición dirigida por la sintáxis sea L-atribuída deben cumplirse que cualquiera que sea la regla de producción $A \to X_1 \dots X_n$, los atributos heredados de X_j pueden depender únicamente de:

1. Los atributos de los símbolos a la izquierda de X_j

2. Los atributos heredados de A

Nótese que las restricciones se refieren a los atributos heredados. El cálculo de los atributos sintetizados no supone problema para un esquema de traducción. Si la gramática es LL(1), resulta fácil realizar una definición L-atribuída en un analizador descendente recursivo predictivo.

Si la definición dirigida por la sintáxis sólo utiliza atributos sintetizados se denomina *S-atribuída*. Una definición S-atribuída puede ser fácilmente trasladada a un programa `jison`.

34.14. Manejo en `jison` de Atributos Heredados

Jison no soporta acciones en medio de una regla.

34.15. Acciones en Medio de una Regla y Atributos Heredados

Jison no soporta acciones en medio de una regla.

34.16. Recuperación de Errores

Jison no soporta recuperación de errores.

34.17. Consejos a seguir al escribir un programa `jison`

Cuando escriba un programa `jison` asegurese de seguir los siguientes consejos:

1. Coloque el punto y coma de separación de reglas en una línea aparte. Un punto y coma "pegado" al final de una regla puede confundirse con un terminal de la regla.

2. Si hay una regla que produce vacío, coloquela en primer lugar y acompáñela de un comentario resaltando ese hecho.

3. Nunca escriba dos reglas de producción en la misma línea.

4. Sangre convenientemente todas las partes derechas de las reglas de producción de una variable, de modo que queden alineadas.

5. Ponga nombres representativos a sus variables sintácticas. No llame `Z` a una variable que representa el concepto "lista de parámetros", llámela `ListaDeParametros`.

6. Es conveniente que declare los terminales simbólicos, esto es, aquellos que llevan un identificador asociado. Si no llevan prioridad asociada o no es necesaria, use una declaración `%token`. De esta manera el lector de su programa se dará cuenta rápidamente que dichos identificadores no se corresponden con variables sintácticas. Por la misma razón, si se trata de terminales asociados con caracteres o cadenas no es tan necesario que los declare, a menos que, como en el ejemplo de la calculadora para `'+'` y `'*'`, sea necesario asociarles una precedencia.

7. Es importante que use la opción `-v` para producir el fichero `.output` conteniendo información detallada sobre los conflictos y el autómata. Cuando haya un conflicto shift-reduce no resuelto busque en el fichero el estado implicado y vea que LR(0) items $A \to \alpha_\uparrow$ y $B \to \beta_\uparrow\gamma$ entran en conflicto.

8. Si según el informe de `jison` el conflicto se produce ante un terminal a, es porque $a \in FOLLOW(A)$ y $a \in FIRST(\gamma)$. Busque las causas por las que esto ocurre y modifique su gramática con vistas a eliminar la presencia del terminal a en uno de los dos conjuntos implicados o bien establezca reglas de prioridad entre los terminales implicados que resuelvan el conflicto.

9. Nótese que cuando existe un conflicto de desplazamiento reducción entre $A \to \alpha_\uparrow$ y $B \to \beta_\uparrow\gamma$, el programa `jison` contabiliza un error por cada terminal $a \in FOLLOW(A) \cap FIRST(\gamma)$. Por esta razón, si hay 16 elementos en $FOLLOW(A) \cap FIRST(\gamma)$, el analizador `jison` informará de la existencia de 16 conflictos *shift-reduce*, cuando en realidad se trata de uno sólo. No desespere, los conflictos "auténticos" suelen ser menos de los que `jison` anuncia.

10. Si necesita declarar variables globales, inicializaciones, etc. que afectan la conducta global del analizador, escriba el código correspondiente en la cabecera del analizador, protegido por los delimitadores `%{` y `%}`. Estos delimitadores deberán aparecer en una línea aparte. Por ejemplo:

```
%{
our contador = 0;
%}

%token NUM
...
%%
```

Capítulo 35

Análisis Sintáctico Ascendente en Perl

35.1. `Parse::Yapp`: Ejemplo de Uso

El generador de analizadores sintácticos `Parse::Yapp` es un analizador LALR inspirado en `yacc` . El módulo `Parse::Yapp` no viene con la distribución de Perl, por lo que es necesario bajarlo desde CPAN, en la dirección

`\http://search.cpan.org/~fdesar/Parse-Yapp-1.05/lib/Parse/Yapp.p`

o bien en nuestros servidores locales, por ejemplo en el mismo directorio en que se guarda la versión HTML de estos apuntes encontrará una copia de Parse-Yapp-1.05.tar.gz. La versión a la que se refiere este capítulo es la 1.05.

Para ilustrar su uso veamos un ejemplo en el que se genera una sencilla calculadora numérica. Los contenidos del programa `yapp` los hemos guardado en un fichero denominado `Calc.yp` (el código completo figura en el apéndice en la página **??**)

```
 1  #
 2  # Calc.yp
 3  #
 4  # Parse::Yapp input grammar example.
 5  #
 6  # This file is PUBLIC DOMAIN
 7  #
 8  #
```

Se pueden poner comentarios tipo Perl o tipo C (`/* ... */`) a lo largo del fichero.

```
 9  %right  '='
10  %left   '-' '+'
11  %left   '*' '/'
12  %left   NEG
13  %right  '^'
14
15  %%
16  input:  #empty
17          |   input line  { push(@{$_[1]},$_[2]); $_[1] }
18  ;
```

las declaraciones `%left` y `%right` expresan la asociatividad y precedencia de los terminales, permitiendo decidir que árbol construir en caso de ambiguedad. *Los terminales declarados en líneas posteriores tienen mas prioridad que los declarados en las líneas anteriores.* Véase la sección 35.7 para mas detalles.

Un programa yapp consta de tres partes: la cabeza, el cuerpo y la cola. Cada una de las partes va separada de las otras por el símbolo %% en una línea aparte. Así, el %% de la línea 15 separa la cabeza del cuerpo. En la cabecera se colocan el código de inicialización, las declaraciones de terminales, las reglas de precedencia, etc. El cuerpo contiene las reglas de la gramática y las acciones asociadas. Por último, la cola de un program yapp contiene las rutinas de soporte al código que aparece en las acciones asi como, posiblemente, rutinas para el análisis léxico y el tratamiento de errores.

En Parse::Yapp las acciones son convertidas en subrutinas anónimas. Mas bien en métodos anónimos. Así pues el primer argumento de la subrutina se identifica con una referencia al analizador ($_[0]). Los restantes parámetros se corresponden con los *atributos de los símbolos* en la parte derecha de la regla de producción ($_[1] ...). Por ejemplo, el código en la línea 21 imprime el atributo asociado con la variable sintáctica expr, que en este caso es su valor numérico.

La línea 17 indica que el atributo asociado con la variable sintáctica input es una referencia a una pila y que el atributo asociado con la variable sintáctica line debe empujarse en la pila. De hecho, el atributo asociado con line es el valor de la expresión. Asi pues el atributo retornado por input es una referencia a una lista conteniendo los valores de las expresiones evaluadas.

Para saber mas sobre las estructuras internas de yapp para la representación de las acciones asociadas con las reglas véase la sección 35.4.

```
19
20  line:  '\n'       { 0 }
21      |  exp '\n'   { print "$_[1]\n"; $_[1] }
22      |  error '\n' { $_[0]->YYErrok; 0 }
23  ;
```

El terminal error en la línea 22 esta asociado con la aparición de un error. El tratamiento es el mismo que en yacc. Cuando se produce un error en el análisis, yapp emite un mensaje de error y produce "mágicamente" un terminal especial denominado error. A partir de ahí permanecerá silencioso, consumiendo terminales hasta encontrar uno de los terminales que le hemos indicado en las reglas de recuperación de errores, en este caso, cuando encuentre un retorno de carro. Como se ha dicho, en Parse::Yapp el primer argumento de la acción denota al analizador sintáctico. Así pues el código $_[0]->YYErrok es una llamada al método YYErrok del analizador. Este método funciona como la macro yyerrok de yacc, indicando que la presencia del retorno del carro (\n) la podemos considerar un signo seguro de que nos hemos recuperado del error. A partir de este momento, yapp volverá a emitir mensajes de error. Para saber más sobre la recuperación de errores en yapp léase la sección 35.15.

```
24
25  exp:        NUM
```

La acción por defecto es retornar $_[1]. Por tanto, en este caso el valor retornado es el asociado a NUM.

```
26          |   VAR           { $_[0]->YYData->{VARS}{$_[1]} }
```

El método YYData provee acceso a un hash que contiene los datos que están siendo analizados. En este caso creamos una entrada VARS que es una referencia a un hash en el que guardamos las variables. Este hash es la tabla de símbolos de la calculadora.

```
27          |   VAR '=' exp   { $_[0]->YYData->{VARS}{$_[1]}=$_[3] }
28          |   exp '+' exp   { $_[1] + $_[3] }
29          |   exp '-' exp   { $_[1] - $_[3] }
30          |   exp '*' exp   { $_[1] * $_[3] }
```

Hay numerosas ambiguedades en esta gramática. Por ejemplo,

- ¿Como debo interpretar la expresión `4 - 5 - 2`? ¿Como `(4 - 5) - 2`? ¿o bien `4 - (5 - 2)`? La respuesta la da la asignación de asociatividad a los operadores que hicimos en las líneas 9-13. Al declarar como asociativo a izquierdas al terminal `-` (línea 10) hemos resuelto este tipo de ambiguedad. Lo que estamos haciendo es indicarle al analizador que a la hora de elegir entre los árboles abstractos $-(-(4,5),2)$ y $-(4,-(5,2))$ elija siempre el árbol que se hunde a izquierdas.
- ¿Como debo interpretar la expresión `4 - 5 * 2`? ¿Como `(4 - 5) * 2`? ¿o bien `4 - (5 * 2)`? Al declarar que `*` tiene mayor prioridad que `-` estamos resolviendo esta otra fuente de ambiguedad. Esto es así pues `*` fué declarado en la línea 11 y `-` en la 10.

```
31          |   exp '/' exp   {
32                              $_[3]
33                              and return($_[1] / $_[3]);
34                              $_[0]->YYData->{ERRMSG}
35                                =   "Illegal division by zero.\n";
36                              $_[0]->YYError;
37                              undef
38                            }
```

En la regla de la división comprobamos que el divisor es distinto de cero. Si es cero inicializamos el atributo `ERRMSG` en la zona de datos con el mensaje de error apropiado. Este mensaje es aprovechado por la subrutina de tratamiento de errores (véase la subrutina `_Error` en la zona de la cola). La subrutina `_Error` es llamada automáticamente por `yapp` cada vez que ocurre un error sintáctico. Esto es asi por que en la llamada al analizador se especifican quienes son las diferentes rutinas de apoyo:

```
my $result = $self->YYParse( yylex => \&_Lexer,
                             yyerror => \&_Error,
                             yydebug => 0x0 );
```

Por defecto, *una regla de producción tiene la prioridad del último terminal que aparece en su parte derecha*. Una regla de producción puede ir seguida de una directiva `%prec` la cual le da una prioridad explícita. Esto puede ser de gran ayuda en ciertos casos de ambiguedad.

```
39          |   '-' exp %prec NEG   { -$_[2] }
```

¿Cual es la ambiguedad que surge con esta regla? Una de las ambiguedad de esta regla esta relacionada con el doble significado del menos como operador unario y binario: hay frases como `-y-z` que tiene dos posibles interpretaciones: Podemos verla como `(-y)-z` o bien como `-(y-z)`. Hay dos árboles posibles. El analizador, cuando este analizando la entrada `-y-z` y vea el segundo `-` deberá escoger uno de los dos árboles. ¿Cuál?. El conflicto puede verse como una "lucha" entre la regla `exp: '-' exp` la cual interpreta la frase como `(-y)-z` y la segunda aparición del terminal `-` el cuál "quiere entrar" para que gane la regla `exp: exp '-' exp` y dar lugar a la interpretación `-(y-z)`. En este caso, las dos reglas $E \to -E$ y $E \to E - E$ tienen, en principio la prioridad del terminal `-`, el cual fué declarado en la línea 10. La prioridad expresada explícitamente para la regla por la declaración `%prec NEG` de la línea 39 hace que la regla tenga la prioridad del terminal `NEG` (línea 12) y por tanto mas prioridad que el terminal `-`. Esto hará que `yapp` finalmente opte por la regla `exp: '-' exp`.

La declaración de `^` como asociativo a derechas y con un nivel de prioridad alto resuelve las ambiguedades relacionadas con este operador:

```
40          |   exp '^' exp         { $_[1] ** $_[3] }
41          |   '(' exp ')'         { $_[2] }
42  ;
```

Después de la parte de la gramática, y separada de la anterior por el símbolo `%%`, sigue la parte en la que se suelen poner las rutinas de apoyo. Hay al menos dos rutinas de apoyo que el analizador

sintáctico requiere le sean pasados como argumentos: la de manejo de errores y la de análisis léxico. El método `Run` ilustra como se hace la llamada al método de análisis sintáctico generado, utilizando la técnica de llamada con argumentos con nombre y pasándole las referencias a las dos subrutinas (en Perl, es un convenio que si el nombre de una subrutina comienza por un guión bajo es que el autor la considera privada):

```
...

sub Run {
    my($self)=shift;
    my $result = $self->YYParse( yylex => \&_Lexer,
                                 yyerror => \&_Error,
                                 yydebug => 0x0 );
    my @result = @$result;
    print "@result\n";
}
```

La subrutina de manejo de errores `_Error` imprime el mensaje de error proveído por el usuario, el cual, si existe, fué guardado en `$_[0]->YYData->{ERRMSG}`.

```
43  # rutinas de apoyo
44  %%
45
46  sub _Error {
47    exists $_[0]->YYData->{ERRMSG}
48      and do {
49          print $_[0]->YYData->{ERRMSG};
50          delete $_[0]->YYData->{ERRMSG};
51          return;
52      };
53    print "Syntax error.\n";
54  }
55
```

A continuación sigue el método que implanta el análisis léxico `_Lexer`. En primer lugar se comprueba la existencia de datos en `parser->YYData->{INPUT}`. Si no es el caso, los datos se tomarán de la entrada estándar:

```
56  sub _Lexer {
57    my($parser)=shift;
58
59    defined($parser->YYData->{INPUT})
60    or  $parser->YYData->{INPUT} = <STDIN>
61    or  return('',undef);
62
```

Cuando el analizador léxico alcanza el final de la entrada debe devolver la pareja `('',undef)`.

Eliminamos los blancos iniciales (lo que en inglés se conoce por *trimming*):

```
63    $parser->YYData->{INPUT}=~s/^[ \t]//;
64
```

A continuación vamos detectando los números, identificadores y los símbolos individuales. El bucle `for ($parser->YYData->{INPUT})` se ejecuta mientras la cadena en `$parser->YYData->{INPUT}` no sea vacía, lo que ocurrirá cuando todos los terminales hayan sido consumidos.

```
65    for ($parser->YYData->{INPUT}) {
66      s/^([0-9]+(?:\.[0-9]+)?)//
67        and return('NUM',$1);
68      s/^([A-Za-z][A-Za-z0-9_]*)//
69        and return('VAR',$1);
70      s/^(.)//s
71        and return($1,$1);
72    }
73  }
```

Ejercicio 35.1.1. *1. ¿Quién es la variable índice en la línea 65?*

2. ¿Sobre quién ocurre el binding en las líneas 66, 68 y 70?

3. ¿Cual es la razón por la que `$parser->YYData->{INPUT}` *se ve modificado si no aparece como variable para el binding en las líneas 66, 68 y 70?*

Construimos el módulo `Calc.pm` a partir del fichero `Calc.yp` especificando la gramática, usando un fichero `Makefile`:

```
> cat Makefile
Calc.pm: Calc.yp
        yapp -m Calc Calc.yp
> make
yapp -m Calc Calc.yp
```

Esta compilación genera el fichero `Calc.pm` conteniendo el analizador:

```
> ls -ltr
total 96
-rw-r-----    1 pl       users        1959 Oct 20  1999 Calc.yp
-rw-r-----    1 pl       users          39 Nov 16 12:26 Makefile
-rwxrwx--x    1 pl       users          78 Nov 16 12:30 usecalc.pl
-rw-rw----    1 pl       users        5254 Nov 16 12:35 Calc.pm
```

El script `yapp` es un *frontend* al módulo `Parse::Yapp`. Admite diversas formas de uso:

- `yapp [options]` *`grammar`*`[.yp]`

  El sufijo `.yp` es opcional.

- `yapp -`*`V`*

  Nos muestra la versión:

  ```
  $ yapp -V
  This is Parse::Yapp version 1.05.
  ```

- `yapp -`*`h`*

  Nos muestra la ayuda:

  ```
  $ yapp -h

  Usage:  yapp [options] grammar[.yp]
    or    yapp -V
    or    yapp -h

     -m module   Give your parser module the name <module>
  ```

```
                default is <grammar>
-v              Create a file <grammar>.output describing your parser
-s              Create a standalone module in which the driver is included
-n              Disable source file line numbering embedded in your parser
-o outfile      Create the file <outfile> for your parser module
                Default is <grammar>.pm or, if -m A::Module::Name is
                specified, Name.pm
-t filename     Uses the file <filename> as a template for creating the parser
                module file.  Default is to use internal template defined
                in Parse::Yapp::Output
-b shebang      Adds '#!<shebang>' as the very first line of the output file

grammar         The grammar file. If no suffix is given, and the file
                does not exists, .yp is added

-V              Display current version of Parse::Yapp and gracefully exits
-h              Display this help screen
```

La opción *-m module* da el nombre al paquete o espacio de nombres o clase encapsulando el analizador. Por defecto toma el nombre de la gramática. En el ejemplo podría haberse omitido.

La opción *-o outfile* da el nombre del fichero de salida. Por defecto toma el nombre de la gramática, seguido del sufijo `.pm`. sin embargo, si hemos especificado la opción *-m A::Module::Name* el valor por defecto será *Name.pm*.

Veamos los contenidos del ejecutable `usecalc.pl` el cuál utiliza el módulo generado por `yapp`:

```
> cat usecalc.pl
#!/usr/local/bin/perl5.8.0 -w

use Calc;

$parser = new Calc();
$parser->Run;
```

Al ejecutar obtenemos:

```
$ ./usecalc3.pl
2+3
5
4*8
32
^D
5 32
```

Pulsamos al final `Ctrl-D` para generar el final de fichero. El analizador devuelve la lista de valores computados la cual es finalmente impresa.

¿En que orden ejecuta `YYParse` las acciones? La respuesta es que el analizador generado por `yapp` construye una derivación a derechas inversa y ejecuta las acciones asociadas a las reglas de producción que se han aplicado. Así, para la frase `3+2` la antiderivación es:

$$NUM + NUM \overset{NUM \leftarrow E}{\Longleftarrow} E + NUM \overset{NUM \leftarrow E}{\Longleftarrow} E + E \overset{E+E \leftarrow E}{\Longleftarrow} E$$

por tanto las acciones ejecutadas son las asociadas con las correspondientes reglas de producción:

1. La acción de la línea 25:

```
25  exp:        NUM { $_[1]; } # acción por defecto
```

Esta instancia de `exp` tiene ahora como atributo `3`.

2. De nuevo la acción de la línea 25:

```
25  exp:        NUM { $_[1]; } # acción por defecto
```

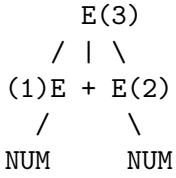
Esta nueva instancia de `exp` tiene como atributo `2`.

3. La acción asociada con $E \to E + E$, en la línea 28:

```
28          |   exp '+' exp   { $_[1] + $_[3] }
```

La nueva instancia (nodo) `exp` tiene como atributo `5 = 3 + 2`.

Obsérvese que la antiderivación a derechas da lugar a un recorrido ascendente y a izquierdas del árbol:



Los números entre paréntesis indican el orden de visita de las producciones.

35.2. Conceptos Básicos

Los analizadores generador por `yapp` entran en la categoría de analizadores *LR*. Estos analizadores construyen una derivación a derechas inversa (o *antiderivación*). De ahí la R en LR (del inglés *rightmost derivation*). El árbol sintáctico es construido de las hojas hacia la raíz, siendo el último paso en la antiderivación la construcción de la primera derivación desde el símbolo de arranque.

Empezaremos entonces considerando las frases que pueden aparecer en una derivación a derechas. Tales frases consituyen el lenguaje FSD:

Definición 35.2.1. *Dada una gramática $G = (\Sigma, V, P, S)$ no ambigua, se denota por FSD (lenguaje de las formas Sentenciales a Derechas) al lenguaje de las sentencias que aparecen en una derivación a derechas desde el símbolo de arranque.*

$$FSD = \left\{ \alpha \in (\Sigma \cup V)* : \exists S \underset{RM}{\overset{*}{\Longrightarrow}} \alpha \right\}$$

*Donde la notacion RM indica una derivación a derechas (*rightmost*). Los elementos de FSD se llaman "formas sentenciales derechas".*

Dada una gramática no ambigua $G = (\Sigma, V, P, S)$ y una frase $x \in L(G)$ el proceso de antiderivación consiste en encontrar la última derivación a derechas que dió lugar a x. Esto es, si $x \in L(G)$ es porque existe una derivación a derechas de la forma

$$S \overset{*}{\Longrightarrow} yAz \Longrightarrow ywz = x.$$

El problema es averiguar que regla $A \to w$ se aplicó y en que lugar de la cadena x se aplicó. En general, si queremos antiderivar una forma sentencial derecha $\beta\alpha w$ debemos averiguar por que regla $A \to \alpha$ seguir y en que lugar de la forma (después de β en el ejemplo) aplicarla.

$$S \overset{*}{\Longrightarrow} \beta Aw \Longrightarrow \beta\alpha w.$$

La pareja formada por la regla y la posición se denomina mango o manecilla de la forma. Esta denominación viene de la visualización gráfica de la regla de producción como una mano que nos permite escalar hacia arriba en el árbol. Los "dedos" serían los símbolos en la parte derecha de la regla de producción.

Definición 35.2.2. *Dada una gramática $G = (\Sigma, V, P, S)$ no ambigua, y dada una forma sentencial derecha $\alpha = \beta\gamma x$, con $x \in \Sigma^*$, el* mango *o* handle *de α es la última producción/posición que dió lugar a α:*

$$S \underset{RM}{\overset{*}{\Longrightarrow}} \beta B x \Longrightarrow \beta\gamma x = \alpha$$

Escribiremos: $handle(\alpha) = (B \to \gamma, \beta\gamma)$. La función *handle* tiene dos componentes: $handle_1(\alpha) = B \to \gamma$ y $handle_2(\alpha) = \beta\gamma$

Si dispusieramos de un procedimiento que fuera capaz de identificar el mango, esto es, de detectar la regla y el lugar en el que se posiciona, tendríamos un mecanismo para construir un analizador. Lo curioso es que, a menudo es posible encontrar un autómata finito que reconoce el lenguaje de los prefijos $\beta\gamma$ que terminan en el mango. Con mas precisión, del lenguaje:

Definición 35.2.3. *El conjunto de* prefijos viables *de una gramática G se define como el conjunto:*

$$PV = \left\{ \delta \in (\Sigma \cup V)* : \exists S \underset{RM}{\overset{*}{\Longrightarrow}} \alpha \text{ y } \delta \text{ es un } prefijo \text{ de } handle_2(\alpha) \right\}$$

Esto es, es el lenguaje de los prefijos viables es el conjunto de frases que son prefijos de $handle_2(\alpha)) = \beta\gamma$, siendo α una forma sentencial derecha ($\alpha \in FSD$). Los elementos de PV se denominan prefijos viables.

Obsérvese que si se dispone de un autómata que reconoce PV entonces se dispone de un mecanismo para investigar el lugar y el aspecto que pueda tener el mango. Si damos como entrada la sentencia $\alpha = \beta\gamma x$ a dicho autómata, el autómata aceptará la cadena $\beta\gamma$ pero rechazará cualquier extensión del prefijo. Ahora sabemos que el mango será alguna regla de producción de G cuya parte derecha sea un sufijo de $\beta\gamma$.

Definición 35.2.4. *El siguiente autómata finito no determinista puede ser utilizado para reconocer el lenguaje de los prefijos viables PV:*

- $Alfabeto = V \cup \Sigma$
- *Los estados del autómata se denominan LR(0) items. Son parejas formadas por una regla de producción de la gramática y una posición en la parte derecha de la regla de producción. Por ejemplo, $(E \to E + E, 2)$ sería un LR(0) item para la gramática de las expresiones.*

  Conjunto de Estados:

  $Q = \{(A \to \alpha, n) : A \to \alpha \in P,\ n \leq |\alpha|\}$

  La notación $\mid \alpha \mid$ denota la longitud de la cadena $\mid \alpha \mid$. En vez de la notación $(A \to \alpha, n)$ escribiremos: $A \to \beta_{\uparrow}\gamma = \alpha$, donde la flecha ocupa el lugar indicado por el número $n = \mid \beta \mid$:
- *Función de transición:*

  $\delta(A \to \alpha_{\uparrow} X\beta, X) = A \to \alpha X_{\uparrow}\beta\ \forall X \in V \cup \Sigma$

  $\delta(A \to \alpha_{\uparrow} B\beta, \epsilon) = B \to \gamma \forall B \in V$
- *Estado de arranque: Se añade la "superregla" $S' \to S$ a la gramática $G = (\Sigma, V, P, S)$. El LR(0) item $S' \to_{\uparrow} S$ es el estado de arranque.*
- *Todos los estados definidos (salvo el de muerte) son de aceptación.*

Denotaremos por $LR(0)$ a este autómata. Sus estados se denominan $LR(0) - items$. La idea es que este autómata nos ayuda a reconocer los prefijos viables PV.

Una vez que se tiene un autómata que reconoce los prefijos viables es posible construir un analizador sintáctico que construye una antiderivación a derechas. La estrategia consiste en "alimentar" el autómata con la forma sentencial derecha. El lugar en el que el autómata se detiene, rechazando indica el lugar exacto en el que termina el *handle* de dicha forma.

Ejemplo 35.2.1. *Consideremos la gramática:*

$$S \to a\ S\ b$$
$$S \to \epsilon$$

El lenguaje generado por esta gramática es $L(G) = \{a^n b^n\ :\ n \geq 0\}$ *Es bien sabido que el lenguaje* $L(G)$ *no es regular. La figura 35.1 muestra el* autómata finito no determinista con ϵ-transiciones *(*NFA*) que reconoce los prefijos viables de esta gramática, construido de acuerdo con el algoritmo 35.2.4.*

Figura 35.1: NFA que reconoce los prefijos viables

Ejercicio 35.2.1. *Simule el comportamiento del autómata sobre la entrada aabb. ¿Donde rechaza? ¿En que estados está el autómata en el momento del rechazo?. ¿Qué etiquetas tienen? Haga también las trazas del autómata para las entradas aaSbb y aSb. ¿Que antiderivación ha construido el autómata con sus sucesivos rechazos? ¿Que terminales se puede esperar que hayan en la entrada cuando se produce el rechazo del autómata?*

35.3. Construcción de las Tablas para el Análisis SLR

35.3.1. Los conjuntos de Primeros y Siguientes

Repasemos las nociones de conjuntos de *Primeros* y *siguientes*:

Definición 35.3.1. *Dada una gramática* $G = (\Sigma, V, P, S)$ *y un símbolo* $\alpha \in (V \cup \Sigma)^*$ *se define el conjunto* $FIRST(\alpha)$ *como:*

$$FIRST(\alpha) = \left\{ b \in \Sigma : \alpha \stackrel{*}{\Longrightarrow} b\beta \right\} \cup N(\alpha)$$

donde:

$$N(\alpha) = \begin{cases} \{\epsilon\} & si\ \alpha \stackrel{*}{\Longrightarrow} \epsilon \\ \emptyset & en\ otro\ caso \end{cases}$$

Definición 35.3.2. *Dada una gramática* $G = (\Sigma, V, P, S)$ *y una variable* $A \in V$ *se define el conjunto* $FOLLOW(A)$ *como:*

$$FOLLOW(A) = \left\{ b \in \Sigma : \exists\ S \stackrel{*}{\Longrightarrow} \alpha A b \beta \right\} \cup E(A)$$

donde

$$E(A) = \begin{cases} \{\$\} & si\ S \stackrel{*}{\Longrightarrow} \alpha A \\ \emptyset & en\ otro\ caso \end{cases}$$

Algoritmo 35.3.1. *Construcción de los conjuntos* $FIRST(X)$

1. *Si* $X \in \Sigma$ *entonces* $FIRST(X) = X$
2. *Si* $X \to \epsilon$ *entonces* $FIRST(X) = FIRST(X) \cup \{\epsilon\}$

3. Si $X \in V$ *y* $X \to Y_1 Y_2 \cdots Y_k \in P$ *entonces*

$i = 1;$
do
$\quad FIRST(X) = FIRST(X) \cup FIRST(Y_i) - \{\epsilon\};$
$\quad i++;$
mientras $(\epsilon \in FIRST(Y_i)$ *and* $(i \leq k))$
si $(\epsilon \in FIRST(Y_k)$ *and* $i > k)$ $FIRST(X) = FIRST(X) \cup \{\epsilon\}$

Este algoritmo puede ser extendido para calcular $FIRST(\alpha)$ para $\alpha = X_1 X_2 \cdots X_n \in (V \cup \Sigma)^*$.

Algoritmo 35.3.2. *Construcción del conjunto* $FIRST(\alpha)$

$i = 1;$
$FIRST(\alpha) = \emptyset;$
do
$\quad FIRST(\alpha) = FIRST(\alpha) \cup FIRST(X_i) - \{\epsilon\};$
$\quad i++;$
mientras $(\epsilon \in FIRST(X_i)$ *and* $(i \leq n))$
si $(\epsilon \in FIRST(X_n)$ *and* $i > n)$ $FIRST(\alpha) = FIRST(X) \cup \{\epsilon\}$

Algoritmo 35.3.3. *Construcción de los conjuntos* $FOLLOW(A)$ *para las variables sintácticas* $A \in V$*:*
Repetir los siguientes pasos hasta que ninguno de los conjuntos $FOLLOW$ *cambie:*

1. $FOLLOW(S) = \{\$\}$ *(*$ *representa el final de la entrada)*

2. Si $A \to \alpha B \beta$ *entonces*

$$FOLLOW(B) = FOLLOW(B) \cup (FIRST(\beta) - \{\epsilon\})$$

3. Si $A \to \alpha B$ *o bien* $A \to \alpha B \beta$ *y* $\epsilon \in FIRST(\beta)$ *entonces*

$$FOLLOW(B) = FOLLOW(B) \cup FOLLOW(A)$$

35.3.2. Construcción de las Tablas

Para la construcción de las tablas de un analizador SLR se construye el *autómata finito determinista* (*DFA*) (Q, Σ, δ, q_0) equivalente al NFA presentado en la sección 35.2 usando el *algoritmo de construcción del subconjunto.*

Como recordará, en la construcción del subconjunto, partiendo del estado de arranque q_0 del NFA con ϵ-transiciones se calcula su *clausura* $\overline{\{q_0\}}$ y las clausuras de los conjuntos de estados $\overline{\delta(\overline{\{q_0\}}, a)}$ a los que transita. Se repite el proceso con los conjuntos resultantes hasta que no se introducen nuevos conjuntos-estado.

La clausura $\overline{A}$ de un subconjunto de estados del autómata A esta formada por todos los estados que pueden ser alcanzados mediante transiciones etiquetadas con la palabra vacía (denominadas ϵ transiciones) desde los estados de A. Se incluyen en $\overline{A}$, naturalmente los estados de A.

$$\overline{A} = \{q \in Q \ / \ \exists q' \in Q \ : \ \hat{\delta}(q, \epsilon) = q\}$$

Aquí $\hat{\delta}$ denota la *función de transición del autómata* extendida a cadenas de Σ^*.

$$\hat{\delta}(q, x) = \begin{cases} \delta(\hat{\delta}(q, y), a) & \text{si } x = ya \\ q & \text{si } x = \epsilon \end{cases} \tag{35.1}$$

Figura 35.2: DFA equivalente al NFA de la figura 35.1

En la práctica, y a partir de ahora así lo haremos, se prescinde de diferenciar entre δ y $\hat{\delta}$ usándose indistintamente la notación δ para ambas funciones.

La clausura puede ser computada usando una estructura de pila o aplicando la expresión recursiva dada en la ecuación 35.1.

Para el NFA mostrado en el ejemplo 35.2.1 el DFA construído mediante esta técnica es el que se muestra en la figura 35.2. Se ha utilizado el símbolo `#` como marcador. Se ha omitido el número 3 para que los estados coincidan en numeración con los generados por `yapp` (véase el cuadro 35.1).

Un analizador sintáctico LR utiliza una tabla para su análisis. Esa tabla se construye a partir de la tabla de transiciones del DFA. De hecho, la tabla se divide en dos tablas, una llamada *tabla de saltos* o *tabla de gotos* y la otra *tabla de acciones*.

La tabla *goto* de un analizador *SLR* no es más que la tabla de transiciones del autómata DFA obtenido aplicando la construcción del subconjunto al NFA definido en 35.2.4. De hecho es la tabla de transiciones restringida a V (recuerde que el alfabeto del autómata es $V \cup \Sigma$). Esto es,

$$\delta_{|V\times Q} : V \times Q \to Q.$$
$$\text{donde se define } goto(i, A) = \delta(A, I_i)$$

La parte de la función de transiciones del DFA que corresponde a los terminales que no producen rechazo, esto es, $\delta_{|\Sigma\times Q} : \Sigma \times Q \to Q$ se adjunta a una tabla que se denomina *tabla de acciones*. La tabla de acciones es una tabla de doble entrada en los estados y en los símbolos de Σ. Las acciones de transición ante terminales se denominan *acciones de desplazamiento* o (*acciones shift*):

$$\delta_{|\Sigma\times Q} : \Sigma \times Q \to Q$$
$$\text{donde se define } action(i, a) = \delta(a, I_i)$$

Cuando un estado s contiene un LR(0)-item de la forma $A \to \alpha_{\uparrow}$, esto es, el estado corresponde a un posible rechazo, ello indica que hemos llegado a un final del prefijo viable, que hemos visto α y que, por tanto, es probable que $A \to \alpha$ sea el *handle* de la forma sentencial derecha actual. Por tanto, añadiremos en entradas de la forma (s, a) de la tabla de acciones una acción que indique que hemos encontrado el mango en la posición actual y que la regla asociada es $A \to \alpha$. A una acción de este tipo se la denomina *acción de reducción*.

La cuestión es, ¿para que valores de $a \in \Sigma$ debemos disponer que la acción para (s, a) es de reducción? Podríamos decidir que ante cualquier terminal $a \in \Sigma$ que produzca un rechazo del autómata, pero podemos ser un poco mas selectivos. No cualquier terminal puede estar en la entrada en el momento en el que se produce la antiderivación o reducción. Observemos que si $A \to \alpha$ es el *handle* de γ es porque:

$$\exists S \underset{RM}{\overset{*}{\Longrightarrow}} \beta A b x \underset{RM}{\overset{*}{\Longrightarrow}} \beta\alpha b x = \gamma$$

Por tanto, cuando estamos reduciendo por $A \to \alpha$ los únicos terminales legales que cabe esperar en una reducción por $A \to \alpha$ son los terminales $b \in FOLLOW(A)$.

Dada una gramática $G = (\Sigma, V, P, S)$, podemos construir las tablas de acciones (*action table*) y transiciones (*gotos table*) mediante el siguiente algoritmo:

Algoritmo 35.3.4. *Construcción de Tablas* SLR

1. *Utilizando el Algoritmo de Construcción del Subconjunto, se construye el Autómata Finito Determinista (DFA)* $(Q, V \cup \Sigma, \delta, I_0, F)$ *equivalente al Autómata Finito No Determinista (NFA) definido en 35.2.4. Sea* $C = \{I_1, I_2, \cdots I_n\}$ *el conjunto de estados del DFA. Cada estado* I_i *es un conjunto de LR(0)-items o estados del NFA. Asociemos un índice* i *con cada conjunto* I_i*.*

2. *La tabla de* gotos *no es más que la función de transición del autómata restringida a las variables de la gramática:*

$$goto(i, A) = \delta(I_i, A) \text{ para todo } A \in V$$

3. *Las acciones para el estado* I_i *se determinan como sigue:*

   a) *Si* $A \to \alpha_{\uparrow} a \beta \in I_i$, $\delta(I_i, a) = I_j$, $a \in \Sigma$ *entonces:*

   $$action[i][a] = shift\ j$$

   b) *Si* $S' \to S_{\uparrow} \in I_i$ *entonces*

   $$action[i][\$] = accept$$

   c) *Para cualquier otro caso de la forma* $A \to \alpha_{\uparrow} \in I_i$ *distinto del anterior hacer*

   $$\forall a \in\ FOLLOW(A) :\ action[i][a] = reduce\ A \to \alpha$$

4. *Las entradas de la tabla de acción que queden indefinidas después de aplicado el proceso anterior corresponden a acciones de "error".*

Definición 35.3.3. *Si alguna de las entradas de la tabla resulta multievaluada, decimos que existe un conflicto y que la gramática no es* SLR*.*

1. *En tal caso, si una de las acciones es de 'reducción" y la otra es de 'desplazamiento", decimos que hay un* conflicto shift-reduce *o* conflicto de desplazamiento-reducción*.*

2. *Si las dos reglas indican una acción de reducción, decimos que tenemos un* conflicto reduce-reduce *o de* reducción-reducción*.*

Ejemplo 35.3.1. *Al aplicar el algoritmo 35.3.4 a la gramática 35.2.1*

1	$S \to a\ S\ b$
2	$S \to \epsilon$

partiendo del autómata finito determinista que se construyó en la figura 35.2 y calculando los conjuntos de primeros y siguientes

	FIRST	*FOLLOW*
S	*a*, ϵ	*b*, *$*

obtenemos la siguiente tabla de acciones SLR:

	a	*b*	*$*
0	*s2*	*r2*	*r2*
1			*aceptar*
2	*s2*	*r2*	*r2*
4		*s5*	
5		*r1*	*r1*

Las entradas denotadas con s n (s por shift) indican un desplazamiento al estado n, las denotadas con r n (r por reduce o reducción) indican una operación de reducción o antiderivación por la regla n. Las entradas vacías corresponden a acciones de error.

El método de análisis *LALR* usado por `yapp` es una extensión del método SLR esbozado aqui. Supone un compromiso entre potencia (conjunto de gramáticas englobadas) y eficiencia (cantidad de memoria utilizada, tiempo de proceso). Veamos como `yapp` aplica la construcción del subconjunto a la gramática del ejemplo 35.2.1. Para ello construimos el siguiente programa `yapp`:

```
$ cat -n aSb.yp
     1  %%
     2  S:  # empty
     3      |   'a' S 'b'
     4  ;
     5  %%
     ......
```

y compilamos, haciendo uso de la opción `-v` para que `yapp` produzca las tablas en el fichero `aSb.output`:

```
$ ls -l aSb.*
-rw-r--r--  1 lhp lhp  738 2004-12-19 09:52 aSb.output
-rw-r--r--  1 lhp lhp 1841 2004-12-19 09:52 aSb.pm
-rw-r--r--  1 lhp lhp  677 2004-12-19 09:46 aSb.yp
```

El contenido del fichero `aSb.output` se muestra en la tabla 35.1. Los números de referencia a las producciones en las acciones de reducción vienen dados por:

```
0: $start -> S $end
1: S -> /* empty */
2: S -> 'a' S 'b'
```

Observe que el final de la entrada se denota por `$end` y el marcador en un LR-item por un punto. Fíjese en el estado 2: En ese estado están también los items

`S -> . 'a' S 'b'` y `S -> .`

sin embargo no se explicitan por que se entiende que su pertenencia es consecuencia directa de aplicar la operación de clausura. Los LR items cuyo marcador no está al principio se denominan *items núcleo.*

Puede encontrar el listado completo de las tablas en `aSb.output` en el apéndice que se encuentra en la página **??**.

Ejercicio 35.3.1. *Compare la tabla 35.1 resultante de aplicar* `yapp` *con la que obtuvo en el ejemplo 35.3.1.*

35.4. El módulo Generado por yapp

La ejecución de la orden `yapp -m Calc Calc.yp` produce como salida el módulo `Calc.pm` el cual contiene las tablas LALR(1) para la gramática descrita en `Calc.yp`. Estas tablas son las que dirigen al analizador LR. Puede ver el código completo del módulo en el apéndice que se encuentra en la página **??**. La estructura del módulo `Calc.pm` es como sigue:

```
1 package Calc;
2 use vars qw ( @ISA );
3 use strict;
```

Estado 0	Estado 1	Estado 2

`$start -> . S $end`
`'a'shift 2`
`$default`
`reduce 1 (S)`
`S go to state 1`
	`$start -> S . $end`
	`$end shift 3`
		`S -> 'a' . S 'b'`
		`'a'shift 2`
		`$default`
		`reduce 1 (S)`
		`S go to state 4`

Estado 3	**Estado 4**	**Estado 5**

`$start -> S $end .`
`$default accept`
	`S -> 'a' S . 'b'`
	`'b'shift 5`
		`S -> 'a' S 'b' .`
		`$default`
		`reduce 2 (S)`

Cuadro 35.1: Tablas generadas por `yapp`. El estado 3 resulta de transitar con $

```
 4 @ISA= qw ( Parse::Yapp::Driver );
 5 use Parse::Yapp::Driver;
 6
 7 sub new {
 8    my($class)=shift;
 9    ref($class) and $class=ref($class);
10
11     my($self)=$class->SUPER::new(
12        yyversion => '1.05',
13        yystates => [
..          ...
32        ], # estados
33        yyrules  => [
..        # ... mas reglas
70        ], # final de las reglas
71        @_); # argumentos pasados
72   bless($self,$class);
73 }
```

La clase `Calc` hereda de `Parse::Yapp::Driver`, pero el objeto creado será bendecido en la clase `Calc` (Línea 4, véanse también la figura 35.3 y la línea 72 del fuente). Por tanto, el constructor llamado en la línea 11 es el de `Parse::Yapp::Driver`. Se utiliza la estrategia de llamada con parámetros con nombre. El valor para la clave `yystates` es una referencia anónima al array de estados y el valor para la clave `yyrules` es una referencia anónima a las reglas.

```
10
11     my($self)=$class->SUPER::new(
12        yyversion => '1.05',
13        yystates => [
14         {#State 0
15           DEFAULT => -1, GOTOS => { 'input' => 1 }
16         },
17         {#State 1
18           ACTIONS => {
19             'NUM' => 6, '' => 4, "-" => 2, "(" => 7,
```

```
20              'VAR' => 8, "\n" => 5, 'error' => 9
21            },
22            GOTOS => { 'exp' => 3, 'line' => 10 }
23          },
24          # ...  mas estados
25          {#State 27
26            ACTIONS => {
27              "-" => 12, "+" => 13, "/" => 15, "^" => 16,
28              "*" => 17
29            },
30            DEFAULT => -8
31          }
32        ], # estados
```

Se ve que un estado se pasa como un hash anónimo indexado en las acciones y los saltos.

Para consultar los números asociados con las reglas de producción vea el apéndice en la página **??** conteniendo el fichero `Calc.output`.

A continuación vienen las reglas:

```
33        yyrules  => [
34        [#Rule 0
35              '$start', 2, undef ],
36        [#Rule 1
37              'input', 0, undef ],
38        [#Rule 2
39              'input', 2, sub
40 #line 17 "Calc.yp"
41              { push(@{$_[1]},$_[2]); $_[1] }
42        ],
43        [#Rule 3
44              'line', 1, sub
45 #line 20 "Calc.yp"
46              { $_[1] }
47        ],
48        [#Rule 4
49              'line', 2, sub
50 #line 21 "Calc.yp"
51              { print "$_[1]\n" }
52        ],
53   # ... mas reglas
54     [#Rule 11
55             'exp', 3, sub
56 #line 30 "Calc.yp"
57 { $_[1] * $_[3] }
58    ],
59   [#Rule 12
60             'exp', 3, sub
61 #line 31 "Calc.yp"
62   {
63     $_[3] and return($_[1] / $_[3]);
64     $_[0]->YYData->{ERRMSG} =   "Illegal division by zero.\n";
65     $_[0]->YYError;
66     undef
67   }
```

```
68     ],
69   # ... mas reglas
70   ], # final de las reglas
```

Las reglas son arrays anónimos conteniendo el nombre de la regla o variable sintáctica (`exp`), el número de símbolos en la parte derecha y la subrutina anónima con el código asociado.

Vemos como la acción es convertida en una subrutina anónima. Los argumentos de dicha subrutina son los atributos semánticos asociados con los símbolos en la parte derecha de la regla de producción. El valor retornado por la acción/subrutina es el valor asociado con la reducción.

Para hacer que el compilador Perl diagnostique los errores relativos al fuente `Calc.yp` se usa una directiva `#line`.

```
71   @_);
72   bless($self,$class);
73 }
74
```

la bendición con dos argumentos hace que el objeto pertenezca a la clase `Calc`. A continuación siguen las subrutinas de soporte:

```
75 #line 44 "Calc.yp"
76
77
78 sub _Error {
79   # ...
80 }
81
82 sub _Lexer {
83   my($parser)=shift;
84   # ...
85 }
86
87 sub Run {
88     my($self)=shift;
89     $self->YYParse( yylex => \&_Lexer, yyerror => \&_Error );
90 }
91
92 my($calc)=new Calc;
93 $calc->Run;
94
95 1;
```

35.5. Algoritmo de Análisis LR

Asi pues la tabla de transiciones del autómata nos genera dos tablas: la tabla de acciones y la de saltos. El algoritmo de análisis sintáctico *LR* en el que se basa *yapp* utiliza una pila y dos tablas para analizar la entrada. Como se ha visto, la tabla de acciones contiene cuatro tipo de acciones:

1. Desplazar (*shift*)
2. Reducir (*reduce*)
3. Aceptar
4. Error

El algoritmo utiliza una pila en la que se guardan los estados del autómata. De este modo se evita tener que "comenzar" el procesado de la forma sentencial derecha resultante después de una reducción (antiderivación).

Algoritmo 35.5.1. *Análizador LR*

```
push(s0);
 b = yylex();
 for( ; ; ;) {
   s = top(0); a = b;
   switch (action[s][a]) {
     case "shift t" :
       push(t);
       b = yylex();
       break;
     case "reduce A ->alpha" :
       eval(Sub{A -> alpha}->(top(|alpha|-1).attr, ... , top(0).attr));
       pop(|alpha|);
       push(goto[top(0)][A]);
       break;
     case "accept" : return (1);
     default : yyerror("syntax error");
   }
 }
```

Como es habitual, $|x|$ denota la longitud de la cadena x. La función `top(k)` devuelve el elemento que ocupa la posición `k` desde el *top* de la pila (esto es, está a profundidad `k`). La función `pop(k)` extrae `k` elementos de la pila. La notación `state.attr` hace referencia al atributo asociado con cada estado. Denotamos por `sub_{reduce A -> alpha}` el código de la acción asociada con la regla $A \to \alpha$.

Todos los analizadores LR comparten, salvo pequeñas exepciones, el mismo algoritmo de análisis. Lo que más los diferencia es la forma en la que construyen las tablas. En `yapp` la construcción de las tablas de *acciones* y *gotos* se realiza mediante el algoritmo *LALR*.

35.6. Depuración en `yapp`

Es posible añadir un parámetro en la llamada a `YYParse` con nombre *yydebug* y valor el nivel de depuración requerido. Ello nos permite observar la conducta del analizador. Los posibles valores de depuración son:

Bit	Información de Depuración
0x01	Lectura de los terminales
0x02	Información sobre los estados
0x04	Acciones (shifts, reduces, accept ...)
0x08	Volcado de la pila
0x10	Recuperación de errores

Veamos un ejemplo de salida para la gramática que se describe en la página **??** cuando se llama con:

```
$self->YYParse( yylex => \&_Lexer, yyerror => \&_Error, yydebug => 0x1F )
```

```
1  $ ./use_aSb.pl
2  ----------------------------------------
3  In state 0:
```

```
 4  Stack:[0]
 5  ab  # el usuario ha escrito esto
 6  Need token. Got >a<
 7  Shift and go to state 2.
 8  ----------------------------------------
 9  In state 2:
10  Stack:[0,2]
11  Need token. Got >b<
12  Reduce using rule 1 (S,0): S -> epsilon
13  Back to state 2, then go to state 4.
14  ----------------------------------------
15  In state 4:
16  Stack:[0,2,4]
17  Shift and go to state 5.
18  ----------------------------------------
19  In state 5:
20  Stack:[0,2,4,5]
21  Don't need token.
22  Reduce using rule 2 (S,3): S -> a S b
23  Back to state 0, then go to state 1.
24  ----------------------------------------
25  In state 1:
26  Stack:[0,1]
27  Need token. Got ><
28  Shift and go to state 3.
29  ----------------------------------------
30  In state 3:
31  Stack:[0,1,3]
32  Don't need token.
33  Accept.
```

35.7. Precedencia y Asociatividad

Recordemos que si al construir la tabla LALR, alguna de las entradas de la tabla resulta multievaluada, decimos que existe un conflicto. Si una de las acciones es de 'reducción" y la otra es de 'desplazamiento", se dice que hay un *conflicto shift-reduce* o *conflicto de desplazamiento-reducción*. Si las dos reglas indican una acción de reducción, decimos que tenemos un *conflicto reduce-reduce* o de *reducción-reducción*. En caso de que no existan indicaciones específicas *yapp* resuelve los conflictos que aparecen en la construcción de la tabla utilizando las siguientes reglas:

1. Un conflicto *reduce-reduce* se resuelve eligiendo la producción que se listó primero en la especificación de la gramática.

2. Un conflicto *shift-reduce* se resuelve siempre en favor del *shift*

Las declaraciones de precedencia y asociatividad mediante las palabras reservadas `%left` , `%right` , `%nonassoc` se utilizan para modificar estos criterios por defecto. La declaración de `token` s mediante la palabra reservada `%token` no modifica la precedencia. Si lo hacen las declaraciones realizadas usando las palabras `left` , `right` y `nonassoc` .

1. Los *tokens* declarados en la misma línea tienen igual precedencia e igual asociatividad. La precedencia es mayor cuanto mas abajo su posición en el texto. Así, en el ejemplo de la calculadora en la sección 35.1, el *token* `*` tiene mayor precedencia que `+` pero la misma que `/`.

2. La precedencia de una regla $A \rightarrow \alpha$ se define como la del terminal mas a la derecha que aparece en α. En el ejemplo, la producción

```
expr : expr '+' expr
```

   tiene la precedencia del *token* `+`.

3. Para decidir en un conflicto *shift-reduce* se comparan la precedencia de la regla con la del terminal que va a ser desplazado. Si la de la regla es mayor se reduce si la del *token* es mayor, se desplaza.

4. Si en un conflicto *shift-reduce* ambos la regla y el terminal que va a ser desplazado tiene la misma precedencia *yapp* considera la asociatividad, si es asociativa a izquierdas, reduce y si es asociativa a derechas desplaza. Si no es asociativa, genera un mensaje de error.
   Obsérvese que, en esta situación, la asociatividad de la regla y la del *token* han de ser por fuerza, las mismas. Ello es así, porque en *yapp* los *tokens* con la misma precedencia se declaran en la misma línea y sólo se permite una declaración por línea.

5. *Por tanto es imposible declarar dos* tokens *con diferente asociatividad y la misma precedencia.*

6. Es posible modificar la precedencia "natural" de una regla, calificándola con un *token* específico. para ello se escribe a la derecha de la regla `prec token`, donde `token` es un *token* con la precedencia que deseamos. Vea el uso del *token* `dummy` en el siguiente ejercicio.

Para ilustrar las reglas anteriores usaremos el siguiente programa `yapp`:

```
$ cat -n Precedencia.yp
     1  %token NUMBER
     2  %left '@'
     3  %right '&'  dummy
     4  %%
     5  list
     6      :
     7      | list '\n'
     8      | list e
     9      ;
    10
    11  e : NUMBER
    12    | e '&' e
    13    | e '@' e %prec dummy
    14    ;
    15
    16  %%
```

El código del programa cliente es el siguiente:

```
$ cat -n useprecedencia.pl
cat -n useprecedencia.pl
     1  #!/usr/bin/perl -w
     2  use strict;
     3  use Precedencia;
     4
     5  sub Error {
     6    exists $_[0]->YYData->{ERRMSG}
     7    and do {
     8      print $_[0]->YYData->{ERRMSG};
```

```
 9      delete $_[0]->YYData->{ERRMSG};
10      return;
11    };
12    print "Syntax error.\n";
13  }
14
15  sub Lexer {
16    my($parser)=shift;
17
18    defined($parser->YYData->{INPUT})
19    or  $parser->YYData->{INPUT} = <STDIN>
20    or  return('',undef);
21
22    $parser->YYData->{INPUT}=~s/^[ \t]//;
23
24    for ($parser->YYData->{INPUT}) {
25        s/^([0-9]+(?:\.[0-9]+)?)//
26                and return('NUMBER',$1);
27        s/^(.)//s
28                and return($1,$1);
29    }
30  }
31
32  my $debug_level = (@ARGV)? oct(shift @ARGV): 0x1F;
33  my $parser = Precedencia->new();
34  $parser->YYParse( yylex => \&Lexer, yyerror => \&Error, yydebug => $debug_level );
```

Observe la llamada al analizador en la línea 34. Hemos añadido el parámetro con nombre *yydebug* con argumento `yydebug => $debug_level` (véase la sección 35.6 para ver los posibles valores de depuración).

Compilamos a continuación el módulo usando la opción `-v` para producir información sobre los conflictos y las tablas de salto y de acciones:

```
yapp -v -m Precedencia Precedencia.yp
$ ls -ltr |tail -2
-rw-r--r--  1 lhp lhp   1628 2004-12-07 13:21 Precedencia.pm
-rw-r--r--  1 lhp lhp   1785 2004-12-07 13:21 Precedencia.output
```

La opción `-v` genera el fichero `Precedencia.output` el cual contiene información detallada sobre el autómata:

```
$ cat -n Precedencia.output
     1  Conflicts:
     2  ----------
     3  Conflict in state 8 between rule 6 and token '@' resolved as reduce.
     4  Conflict in state 8 between rule 6 and token '&' resolved as shift.
     5  Conflict in state 9 between rule 5 and token '@' resolved as reduce.
     6  Conflict in state 9 between rule 5 and token '&' resolved as shift.
     7
     8  Rules:
     9  ------
    10  0:      $start -> list $end
    11  1:      list -> /* empty */
    12  2:      list -> list '\n'
    13  3:      list -> list e
```

```
14  4:      e -> NUMBER
15  5:      e -> e '&' e
16  6:      e -> e '@' e
17  ...
```

¿Porqué se produce un conflicto en el estado 8 entre la regla 6 (`e -> e '@' e`) y el terminal `'@'`?. Editando el fichero `Precedencia.output` podemos ver los contenidos del estado 8:

```
85  State 8:
86
87          e -> e . '&' e  (Rule 5)
88          e -> e . '@' e  (Rule 6)
89          e -> e '@' e .  (Rule 6)
90
91          '&'     shift, and go to state 7
92
93          $default        reduce using rule 6 (e)
```

El item de la línea 88 indica que debemos desplazar, el de la línea 89 que debemos reducir por la regla 6. ¿Porqué `yapp` resuelve el conflicto optando por reducir? ¿Que prioridad tiene la regla 6? ¿Que asociatividad tiene la regla 6? La declaración en la línea 13 modifica la precedencia y asociatividad de la regla:

```
13    | e '@' e %prec dummy
```

de manera que la regla pasa a tener la precedencia y asociatividad de `dummy`. Recuerde que habíamos declarado `dummy` como asociativo a derechas:

```
 2  %left '@'
 3  %right '&'  dummy
```

¿Que ocurre? Que `dummy` tiene mayor prioridad que `'@'` y por tanto la regla tiene mayor prioridad que el terminal: por tanto se reduce.

¿Que ocurre cuando el terminal en conflicto es `'&'`? En ese caso la regla y el terminal tienen la misma prioridad. Se hace uso de la asociatividad a derechas que indica que el conflicto debe resolverse desplazando.

Ejercicio 35.7.1. *Explique la forma en que* `yapp` *resuelve los conflictos que aparecen en el estado 9. Esta es la información sobre el estado 9:*

```
State 9:

e -> e . '&' e (Rule 5)
e -> e '&' e . (Rule 5)
e -> e . '@' e (Rule 6)
'&'shift, and go to state 7
$default reduce using rule 5 (e)
```

Veamos un ejemplo de ejecución:

```
$ ./useprecedencia.pl
----------------------------------------
In state 0:
Stack:[0]
Don't need token.
Reduce using rule 1 (list,0): Back to state 0, then go to state 1.
```

Lo primero que ocurre es una reducción por la regla en la que `list` produce vacío. Si miramos el estado 0 del autómata vemos que contiene:

```
20 State 0:
21
22   $start -> . list $end (Rule 0)
23
24   $default  reduce using rule 1 (list)
25
26   list  go to state 1
```

A continuación se transita desde 0 con `list` y se consume el primer terminal:

```
----------------------------------------
In state 1:
Stack:[0,1]
2@3@4
Need token. Got >NUMBER<
Shift and go to state 5.
----------------------------------------
In state 5:
Stack:[0,1,5]
Don't need token.
Reduce using rule 4 (e,1): Back to state 1, then go to state 2.
----------------------------------------
```

En el estado 5 se reduce por la regla `e -> NUMBER`. Esto hace que se retire el estado 5 de la pila y se transite desde el estado 1 viendo el símbolo `e`:

```
In state 2:
Stack:[0,1,2]
Need token. Got >@<
Shift and go to state 6.
----------------------------------------
In state 6:
Stack:[0,1,2,6]
Need token. Got >NUMBER<
Shift and go to state 5.
----------------------------------------
In state 5:
Stack:[0,1,2,6,5]
Don't need token.
Reduce using rule 4 (e,1): Back to state 6, then go to state 8.
----------------------------------------
In state 8:
Stack:[0,1,2,6,8]
Need token. Got >@<
Reduce using rule 6 (e,3): Back to state 1, then go to state 2.
----------------------------------------
...
Accept.
```

Obsérvese la resolución del conflicto en el estado 8

La presencia de conflictos, aunque no siempre, en muchos casos es debida a la introducción de ambiguedad en la gramática. Si el conflicto es de desplazamiento-reducción se puede resolver explicitando alguna regla que rompa la ambiguedad. Los conflictos de reducción-reducción suelen producirse por un diseño erróneo de la gramática. En tales casos, suele ser mas adecuado modificar la gramática.

35.8. Generación interactiva de analizadores Yapp

En el siguiente código, la subrutina `create_yapp_package` nos muestra como crear un analizador `Yapp` en tiempo de ejecución. Las dos líneas:

```
  my $p = new Parse::Yapp(input => $grammar);
  $p = $p->Output(classname => $name);
```

crean una cadena en `$p` conteniendo el código de la clase que implanta el analizador. Todo el truco está en hacer

```
  eval $p;
```

para tener el paquete a mano:

```
$ cat left.pl
#!/usr/local/bin/perl5.8.0 -w
#use strict;
use Parse::Yapp;

sub lex{
    my($parser)=shift;

    return('',undef) unless $parser->YYData->{INPUT};
    for ($parser->YYData->{INPUT}) {
        s/^\s*//;
        s/^(.)//;
        my $ret = $1;
        return($ret, $ret);
    }
}

sub yapp {
  my $grammar = shift
     or die "Must specify a grammar as first argument";
  my $name = shift
     or die "Must specify the name of the class as second argument";

  my $p = new Parse::Yapp(input => $grammar) or die "Bad grammar.";
  $p = $p->Output(classname => $name) or die "Can't generate parser.";

  eval $p;
  $@ and die "Error while compiling your parser: $@\n";
}

######## main #########
my $grammar = q {
%left '*'
%%
S:  A
;

A:  A '*' A  { "($_[1] $_[2] $_[3])" }
 |  B
;
```

```
B:  'a' | 'b' | 'c' | 'd'
;

%%
};

&yapp($grammar, "Example");
my $p = new Example(yylex => \&lex, yyerror => sub {});

print "Expresion: ";
$p->YYData->{INPUT} = <>;
$p->YYData->{INPUT} =~ s/\s*$//;

my $out=$p->YYParse;
print "out = $out\n";
```

Sigue un ejemplo de ejecución:

```
$ ./left.pl
Expresion: a*b*c*d
out = (((a * b) * c) * d)
```

35.9. Construcción del Árbol Sintáctico

El siguiente ejemplo usa `yapp` para construir el árbol sintáctico de una expresión en infijo:

```
$ cat -n Infixtree_bless.yp
 1  #
 2  # Infixtree.yp
 3  #
 4
 5  %{
 6  use Data::Dumper;
 7  %}
 8  %right  '='
 9  %left   '-' '+'
10  %left   '*' '/'
11  %left   NEG
12
13  %%
14  input:  #empty
15          |   input line
16  ;
17
18  line:     '\n'         { $_[1] }
19          | exp '\n'     { print Dumper($_[1]); }
20          | error '\n'   { $_[0]->YYErrok }
21  ;
22
23  exp:        NUM
24          |   VAR                { $_[1] }
25          |   VAR '=' exp        { bless [$_[1], $_[3]], 'ASSIGN' }
26          |   exp '+' exp        { bless [$_[1], $_[3] ], 'PLUS'}
27          |   exp '-' exp        { bless [$_[1], $_[3] ], 'MINUS'}
```

```
28          |   exp '*' exp         { bless [$_[1], $_[3]], 'TIMES' }
29          |   exp '/' exp         { bless [$_[1], $_[3]], 'DIVIDE' }
30          |   '-' exp %prec NEG   { bless [$_[2]], 'NEG' }
31          |   '(' exp ')'         { $_[2] }
32  ;
33
34  %%
35
36  sub _Error {
37          exists $_[0]->YYData->{ERRMSG}
38      and do {
39          print $_[0]->YYData->{ERRMSG};
40          delete $_[0]->YYData->{ERRMSG};
41          return;
42      };
43      print "Syntax error.\n";
44  }
45
46  sub _Lexer {
47      my($parser)=shift;
48
49          defined($parser->YYData->{INPUT})
50      or  $parser->YYData->{INPUT} = <STDIN>
51      or  return('',undef);
52
53      $parser->YYData->{INPUT}=~s/^[ \t]//;
54
55      for ($parser->YYData->{INPUT}) {
56          s/^([0-9]+(?:\.[0-9]+)?)//
57                  and return('NUM',$1);
58          s/^([A-Za-z][A-Za-z0-9_]*)//
59                  and return('VAR',$1);
60          s/^(.)//s
61                  and return($1,$1);
62      }
63  }
64
65  sub Run {
66      my($self)=shift;
67      $self->YYParse( yylex => \&_Lexer, yyerror => \&_Error );
68  }
```

Para compilar hacemos:

```
$ yapp -m Infixtree Infixtree_bless.yp
```

El guión que usa el analizador anterior es similar al que vimos en la sección 35.1:

```
$ cat -n ./useinfixtree.pl
     1  #!/usr/bin/perl -w
     2
     3  use Infixtree;
     4
     5  $parser = new Infixtree();
     6  $parser->Run;
```

Veamos un ejemplo de ejecución:

```
$ ./useinfixtree.pl
a = 2+3
$VAR1 = bless( [
                 'a',
                 bless( [
                          '2',
                          '3'
                        ], 'PLUS' )
               ], 'ASSIGN' );
b = a*4+a
$VAR1 = bless( [
                 'b',
                 bless( [
                          bless( [
                                   'a',
                                   '4'
                                 ], 'TIMES' ),
                          'a'
                        ], 'PLUS' )
               ], 'ASSIGN' );
```

35.10. Acciones en Medio de una Regla

A veces necesitamos insertar una acción en medio de una regla. Una acción en medio de una regla puede hacer referencia a los atributos de los símbolos que la preceden (usando `$n`), pero no a los que le siguen.

Cuando se inserta una acción $\{action_1\}$ para su ejecución en medio de una regla $A \to \alpha\beta$:

$$A \to \alpha \{action_1\} \beta \{action_2\}$$

`yapp` crea una variable sintáctica temporal T e introduce una nueva regla:

1. $A \to \alpha T \beta \{action_2\}$
2. $T \to \epsilon \{action_1\}$

Las acciones en mitad de una regla cuentan como un símbolo mas en la parte derecha de la regla. Asi pues, en una acción posterior en la regla, se deberán referenciar los atributos de los símbolos, teniendo en cuenta este hecho.

Las acciones en mitad de la regla pueden tener un atributo. Las acciones posteriores en la regla se referirán a él como `$_[n]`, siendo `n` su número de orden en la parte derecha.

35.11. Esquemas de Traducción

Un *esquema de traducción* es una gramática independiente del contexto en la cuál se han asociado atributos a los símbolos de la gramática. Un atributo queda caracterizado por un identificador o nombre y un tipo o clase. Además se han insertado acciones, esto es, código Perl/Python/C, ... en medio de las partes derechas. En ese código es posible referenciar los atributos de los símbolos de la gramática como variables del lenguaje subyacente.

Recuerde que el orden en que se evalúan los fragmentos de código es el de un recorrido primero-profundo del árbol de análisis sintáctico. Mas específicamente, considerando a las acciones como hijos-hoja del nodo, el recorrido que realiza un esquema de traducción es:

```
 1     sub esquema_de_traduccion {
 2       my $node = shift;
 3
 4       for my $child ($node->children) { # de izquierda a derecha
 5         if ($child->isa('ACTION') {
 6           $child->execute;
 7         }
 8         else { esquema_de_traduccion($child) }
 9       }
10     }
```

Obsérvese que, como el bucle de la línea 4 recorre a los hijos de izquierda a derecha, se debe dar la siguiente condición para que un esquema de traducción funcione:

Para cualquier regla de producción aumentada con acciones, de la forma

$$A \to X_1 \dots X_j \{\ \texttt{action(\$A\{b\}, \$X}_1\texttt{\{c\}... X}_n\texttt{\{d\})}\ \} X_{j+1} \dots X_n$$

debe ocurrir que los atributos evaluados en la acción insertada después de X_j dependan de atributos y variables que fueron computadas durante la visita de los hermanos izquierdos o de sus ancestros. En particular no deberían depender de atributos asociados con las variables $X_{j+1} \dots X_n$. Ello no significa que no sea correcto evaluar atributos de $X_{j+1} \dots X_n$ en esa acción.

35.12. Definición Dirigida por la Sintáxis

Una *definición dirigida por la sintáxis* es un pariente cercano de los esquemas de traducción. En una definición dirigida por la sintáxis una gramática $G = (V, \Sigma, P, S)$ se aumenta con nuevas características:

- A cada símbolo $S \in V \cup \Sigma$ de la gramática se le asocian cero o mas atributos. Un atributo queda caracterizado por un identificador o nombre y un tipo o clase. A este nivel son *atributos formales*, como los parámetros formales, en el sentido de que su realización se produce cuando el nodo del árbol es creado.

- A cada regla de producción $A \to X_1 X_2 \dots X_n \in P$ se le asocian un conjunto de *reglas de evaluación de los atributos* o *reglas semánticas* que indican que el atributo en la parte izquierda de la regla semántica depende de los atributos que aparecen en la parte derecha de la regla. El atributo que aparece en la parte izquierda de la regla semántica puede estar asociado con un símbolo en la parte derecha de la regla de producción.

- Los atributos de cada símbolo de la gramática $X \in V \cup \Sigma$ se dividen en dos grupos disjuntos: *atributos sintetizados* y *atributos heredados*. Un atributo de X es un *atributo heredado* si depende de atributos de su padre y hermános en el árbol. Un *atributo sintetizado* es aquél tal que el valor del atributo depende de los valores de los atributos de los hijos, es decir en tal caso X ha de ser una variable sintáctica y los atributos en la parte derecha de la regla semántica deben ser atributos de símbolos en la parte derecha de la regla de producción asociada.

- Los atributos predefinidos se denominán *atributos intrínsecos*. Ejemplos de atributos intrínsecos son los atributos sintetizados de los terminales, los cuáles se han computado durante la fase de análisis léxico. También son atributos intrínsecos los atributos heredados del símbolo de arranque, los cuales son pasados como parámetros al comienzo de la computación.

La diferencia principal con un esquema de traducción está en que no se especifica el orden de ejecución de las reglas semánticas. Se asume que, bien de forma manual o automática, se resolverán las dependencias existentes entre los atributos determinadas por la aplicación de las reglas semánticas, de manera que serán evaluados primero aquellos atributos que no dependen de ningún otro, despues los

que dependen de estos, etc. siguiendo un esquema de ejecución que viene guiado por las dependencias existentes entre los datos.

Aunque hay muchas formas de realizar un evaluador de una definición dirigida por la sintáxis, conceptualmente, tal evaluador debe:

1. Construir el árbol de análisis sintáctico para la gramática y la entrada dadas.
2. Analizar las reglas semánticas para determinar los atributos, su clase y las dependencias entre los mismos.
3. Construir el *grafo de dependencias* de los atributos, el cual tiene un nodo para cada ocurrencia de un atributo en el árbol de análisis sintáctico etiquetado con dicho atributo. El grafo tiene una arista entre dos nodos si existe una dependencia entre los dos atributos a través de alguna regla semántica.
4. Supuesto que el grafo de dependencias determina un *orden parcial* (esto es cumple las propiedades reflexiva, antisimétrica y transitiva) construir un *orden topológico* compatible con el orden parcial.
5. Evaluar las reglas semánticas de acuerdo con el orden topológico.

Una definición dirigida por la sintáxis en la que las reglas semánticas no tienen efectos laterales se denomina una *gramática atribuída*.

Si la definición dirigida por la sintáxis puede ser realizada mediante un esquema de traducción se dice que es *L-atribuída*. Para que una definición dirigida por la sintáxis sea L-atribuída deben cumplirse que cualquiera que sea la regla de producción $A \to X_1 \dots X_n$, los atributos heredados de X_j pueden depender únicamente de:

1. Los atributos de los símbolos a la izquierda de X_j
2. Los atributos heredados de A

Nótese que las restricciones se refieren a los atributos heredados. El cálculo de los atributos sintetizados no supone problema para un esquema de traducción. Si la gramática es LL(1), resulta fácil realizar una definición L-atribuída en un analizador descendente recursivo predictivo.

Si la definición dirigida por la sintáxis sólo utiliza atributos sintetizados se denomina *S-atribuída*. Una definición S-atribuída puede ser fácilmente trasladada a un programa `yapp`.

35.13. Manejo en `yapp` de Atributos Heredados

Supongamos que `yapp` esta inmerso en la construcción de la antiderivación a derechas y que la forma sentencial derecha en ese momento es:

$$X_m \dots X_1 X_0 Y_1 \dots Y_n a_1 \dots a_0$$

y que el mango es $B \to Y_1 \dots Y_n$ y en la entrada quedan por procesar $a_1 \dots a_0$.

Es posible acceder en `yapp` a los valores de los atributos de los estados en la pila del analizador que se encuentran "por debajo" o si se quiere "a la izquierda" de los estados asociados con la regla por la que se reduce. Para ello se usa una llamada al método `YYSemval`. La llamada es de la forma `$_[0]->YYSemval( index )`, donde `index` es un entero. Cuando se usan los valores `1 ...n` devuelve lo mismo que `$_[1]`, ...`$_[n]`. Esto es `$_[1]` es el atributo asociado con Y_1 y `$_[n]` es el atributo asociado con Y_n. Cuando se usa con el valor 0 devolverá el valor del atributo asociado con el símbolo que esta a la izquierda del mango actual, esto es el atributo asociado con X_0, si se llama con -1 el que está dos unidades a la izquierda de la variable actual, esto es, el asociado con X_1 etc. Así `$_[-m]` denota el atributo de X_m.

Esta forma de acceder a los atributos es especialmente útil cuando se trabaja con *atributos heredados*. Esto es, cuando un atributo de un nodo del árbol sintáctico se computa en términos de valores

de atributos de su padre y/o sus hermanos. Ejemplos de atributos heredados son la clase y tipo en la declaración de variables. Supongamos que tenemos el siguiente *esquema de traducción* para calcular la clase (C) y tipo (T) en las declaraciones (D) de listas (L) de identificadores:

```
D →  C T { $L{c} = $C{c}; $L{t} = $T{t} } L
C →  global { $C{c} = "global" }
C →  local { $C{c} = "local" }
T →  integer { $T{t} = "integer" }
T →  float { $T{t} = "float" }
L →  { $L1{t} = $L{t}; $L1{c} = $L{c}; } L1 ','
     id { set_class($id{v}, $L{c}); set_type($id{v}, $L{t}); }
L →  id { set_class($id{v}, $L{c}); set_type($id{v}, $L{t}); }
```

Los atributos `c` y `t` denotan respectivamente la clase y el tipo.

Ejercicio 35.13.1. *Evalúe el esquema de traducción para la entrada* `global float x,y`*. Represente el árbol de análisis, las acciones incrustadas y determine el orden de ejecución.*

Olvide por un momento la notación usada en las acciones y suponga que se tratara de acciones `yapp`*. ¿En que orden construye* `yapp` *el árbol y en que orden ejecutará las acciones?*

A la hora de transformar este esquema de traducción en un programa `yapp` es importante darse cuenta que en cualquier derivación a derechas desde D, cuando se reduce por una de las reglas

$$\mathrm{L} \to \mathrm{id} \mid \mathrm{L}_1 \text{ ',' } \mathrm{id}$$

el símbolo a la izquierda de L es T y el que esta a la izquierda de T es C. Considere, por ejemplo la derivación a derechas:

$$\begin{aligned}\mathrm{D} \Longrightarrow \mathrm{C\ T\ L} \Longrightarrow \mathrm{C\ T\ L,\ id} \Longrightarrow \mathrm{C\ T\ L,\ id,\ id} \Longrightarrow \mathrm{C\ T\ id,\ id,\ id} \Longrightarrow \\ \Longrightarrow \mathrm{C\ float\ id,\ id,\ id} \Longrightarrow \mathrm{local\ float\ id,\ id,\ id}\end{aligned}$$

Observe que el orden de recorrido de `yapp` es:

$$\begin{aligned}\mathrm{local\ float\ id,\ id,\ id} \Longleftarrow \mathrm{C\ float\ id,\ id} \Longleftarrow \mathrm{C\ T\ id,\ id,\ id} \Longleftarrow \\ \Longleftarrow \mathrm{C\ T\ L,\ id,\ id} \Longleftarrow \mathrm{C\ T\ L,\ id} \Longleftarrow \mathrm{C\ T\ L} \Longleftarrow \mathrm{D}\end{aligned}$$

en la antiderivación, cuando el mango es una de las dos reglas para listas de identificadores, L → id y L → L, id es decir durante las tres ultimas antiderivaciones:

$$\mathrm{C\ T\ L,\ id,\ id} \Longleftarrow \mathrm{C\ T\ L,\ id} \Longleftarrow \mathrm{C\ T\ L} \Longleftarrow \mathrm{D}$$

las variables a la izquierda del mango son T y C. Esto ocurre siempre. Estas observaciones nos conducen al siguiente programa `yapp`:

```
$ cat -n Inherited.yp
 1  %token FLOAT INTEGER
 2  %token GLOBAL
 3  %token LOCAL
 4  %token NAME
 5
 6  %%
 7  declarationlist
 8    : # vacio
 9    | declaration ';' declarationlist
10    ;
```

```
11
12  declaration
13    : class type namelist { ; }
14    ;
15
16  class
17    : GLOBAL
18    | LOCAL
19    ;
20
21  type
22    : FLOAT
23    | INTEGER
24    ;
25
26  namelist
27    : NAME
28       { printf("%s de clase %s, tipo %s\n",
29               $_[1], $_[0]->YYSemval(-1),$_[0]->YYSemval(0)); }
30    | namelist ',' NAME
31       { printf("%s de clase %s, tipo %s\n",
32                $_[3], $_[0]->YYSemval(-1),$_[0]->YYSemval(0)); }
33    ;
34  %%
```

A continuación escribimos el programa que usa el módulo generado por `yapp`:

```
$ cat -n useinherited.pl
 1  #!/usr/bin/perl -w
 2  use strict;
 3  use Inherited;
 4
 5  sub Error {
 6    exists $_[0]->YYData->{ERRMSG}
 7    and do {
 8      print $_[0]->YYData->{ERRMSG};
 9      delete $_[0]->YYData->{ERRMSG};
10      return;
11    };
12    print "Error sintáctico\n";
13  }
14
15  { # hagamos una clausura con la entrada
16    my $input;
17    local $/ = undef;
18    print "Entrada (En Unix, presione CTRL-D para terminar):\n";
19    $input = <stdin>;
20
21    sub scanner {
22
23      { # Con el redo del final hacemos un bucle "infinito"
24        if ($input =~ m|\G\s*INTEGER\b|igc) {
25          return ('INTEGER', 'INTEGER');
26        }
```

```
27        elsif ($input =~ m|\G\s*FLOAT\b|igc) {
28          return ('FLOAT', 'FLOAT');
29        }
30        elsif ($input =~ m|\G\s*LOCAL\b|igc) {
31          return ('LOCAL', 'LOCAL');
32        }
33        elsif ($input =~ m|\G\s*GLOBAL\b|igc) {
34          return ('GLOBAL', 'GLOBAL');
35        }
36        elsif ($input =~ m|\G\s*([a-z_]\w*)\b|igc) {
37          return ('NAME', $1);
38        }
39        elsif ($input =~ m/\G\s*([,;])/gc) {
40          return ($1, $1);
41        }
42        elsif ($input =~ m/\G\s*(.)/gc) {
43          die "Caracter invalido: $1\n";
44        }
45        else {
46          return ('', undef); # end of file
47        }
48        redo;
49      }
50    }
51  }
52
53  my $debug_level = (@ARGV)? oct(shift @ARGV): 0x1F;
54  my $parser = Inherited->new();
55  $parser->YYParse( yylex => \&scanner, yyerror => \&Error, yydebug => $debug_level );
```

En las líneas de la 15 a la 51 esta nuestro analizador léxico. La entrada se lee en una variable local cuyo valor permanece entre llamadas: hemos creado una clausura con la variable `$input` (véase la sección [10] para mas detalles sobre el uso de clausuras en Perl). Aunque la variable `$input` queda inaccesible desde fuera de la clausura, persiste entre llamadas como consecuencia de que la subrutina `scanner` la utiliza.

A continuación sigue un ejemplo de ejecución:

```
$ ./useinherited.pl 0
Entrada (En Unix, presione CTRL-D para terminar):
global integer x, y, z;
local float a,b;
x de clase GLOBAL, tipo INTEGER
y de clase GLOBAL, tipo INTEGER
z de clase GLOBAL, tipo INTEGER
a de clase LOCAL, tipo FLOAT
b de clase LOCAL, tipo FLOAT
```

Ejercicio 35.13.2. *El siguiente programa* `yapp` *calcula un árbol de análisis abstracto para la gramática del ejemplo anterior:*

```
%token FLOAT INTEGER
%token GLOBAL
%token LOCAL
%token NAME
```

```
%%
declarationlist
  : /* vacio */                    { bless [], 'declarationlist' }
  | declaration ';' declarationlist { push @{$_[3]}, $_[1]; $_[3] }
  ;

declaration
  : class type namelist
      {
        bless {class => $_[1], type => $_[2], namelist => $_[3]}, 'declaration';
      }
  ;

class
  : GLOBAL  { bless { GLOBAL => 0}, 'class' }
  | LOCAL   { bless { LOCAL => 1}, 'class' }
  ;

type
  : FLOAT   { bless { FLOAT => 2}, 'type' }
  | INTEGER { bless { INTEGER => 3}, 'type' }
  ;

namelist
  : NAME
     { bless [ $_[1]], 'namelist' }
  | namelist ',' NAME
      { push @{$_[1]}, $_[3]; $_[1] }
  ;
%%
```

sigue un ejemplo de ejecución:

```
$ ./useinherited3.pl
Entrada (En Unix, presione CTRL-D para terminar):
global float x,y;
$VAR1 = bless( [
  bless( {
    'namelist' => bless( [ 'x', 'y' ], 'namelist' ),
    'type' => bless( { 'FLOAT' => 2 }, 'type' ),
    'class' => bless( { 'GLOBAL' => 0 }, 'class' )
  }, 'declaration' )
], 'declarationlist' );
```

Extienda el programa del ejemplo para que la gramática incluya las acciones del esquema de traducción. Las acciones se tratarán como un terminal `CODE` *y serán devueltas por el analizador léxico. Su atributo asociado es el texto del código. El programa* `yapp` *deberá devolver el árbol abstracto extendido con las acciones-terminales. La parte mas difícil de este problema consiste en "reconocer" el código Perl incrustado. La estrategia seguir consiste en contar el número de llaves que se abren y se cierran. Cuando el contador alcanza cero es que hemos llegado al final del código Perl incrustado. Esta estrategia tiene una serie de problemas. ¿Sabría decir cuáles? (sugerencia: repase la sección 35.19.3 o vea como* `yapp` *resuelve el problema).*

35.14. Acciones en Medio de una Regla y Atributos Heredados

La estrategia utilizada en la sección 35.13 funciona si podemos predecir la posición del atributo en la pila del analizador. En el ejemplo anterior los atributos clase y tipo estaban siempre, cualquiera que fuera la derivación a derechas, en las posiciones 0 y -1. Esto no siempre es asi. Consideremos la siguiente *definición dirigida por la sintáxis*:

S → a A C	`$C{i} = $A{s}`
S → b A B C	`$C{i} = $A{s}`
C → c	`$C{s} = $C{i}`
A → a	`$A{s} = "a"`
B → b	`$B{s} = "b"`

Ejercicio 35.14.1. *Determine un orden correcto de evaluación de la anterior definición dirigida por la sintáxis para la entrada* `b a b c`.

C hereda el atributo sintetizado de A. El problema es que, en la pila del analizador el atributo `$A{s}` puede estar en la posición 0 o -1 dependiendo de si la regla por la que se derivó fué S → a A C o bien S → b A B C. La solución a este tipo de problemas consiste en insertar acciones intermedias de copia del atributo de manera que se garantize que el atributo de interés está siempre a una distancia fija. Esto es, se inserta una variable sintáctica intermedia auxiliar M la cual deriva a vacío y que tiene como acción asociada una regla de copia:

S → a A C	`$C{i} = $A{s}`
S → b A B M C	`$M{i} = $A{s}; $C{i} = $M{s}`
C → c	`$C{s} = $C{i}`
A → a	`$A{s} = "a"`
B → b	`$B{s} = "b"`
M → ϵ	`$M{s} = $M{i}`

El nuevo esquema de traducción puede ser implantado mediante un programa `yapp`:

```
$ cat -n Inherited2.yp
     1  %%
     2  S : 'a' A C
     3    | 'b' A B  { $_[2]; } C
     4    ;
     5
     6  C : 'c' { print "Valor: ",$_[0]->YYSemval(0),"\n"; $_[0]->YYSemval(0) }
     7    ;
     8
     9  A : 'a' { 'a' }
    10    ;
    11
    12  B : 'b' { 'b' }
    13    ;
    14
    15  %%
```

La ejecución muestra como se ha propagado el valor del atributo:

```
$ ./useinherited2.pl '0x04'
Entrada (En Unix, presione CTRL-D para terminar):
b a b c
Shift 2.  Shift 6.
Reduce using rule 5 (A,1): Back to state 2, then state 5.
Shift 8.
Reduce 6 (B,1): Back to state 5, then state 9.
Reduce 2 (@1-3,0): Back to state 9, then state 12.
```

En este momento se esta ejecutando la acción intermedia. Lo podemos comprobar revisando el fichero `Inherited2.output` que fué generado usando la opción `-v` al llamar a `yapp`. La regla 2 por la que se reduce es la asociada con la acción intermedia:

```
$ cat -n Inherited2.output
     1  Rules:
     2  ------
     3  0:      $start -> S $end
     4  1:      S -> 'a' A C
     5  2:      @1-3 -> /* empty */
     6  3:      S -> 'b' A B @1-3 C
     7  4:      C -> 'c'
     8  5:      A -> 'a'
     9  6:      B -> 'b'
        ...
```

Obsérvese la notación usada por `yapp` para la *acción en medio de la regla*: `@1-3`. Continuamos con la antiderivación:

```
Shift 10.
Reduce 4 (C,1):
Valor: a
Back to state 12, then 13.
Reduce using rule 3 (S,5): Back to state 0, then state 1.
Shift 4.
Accept.
```

El método puede ser generalizado a casos en los que el atributo de interés este a diferentes distancias en diferentes reglas sin mas que introducir las correspondientes acciones intermedias de copia.

35.15. Recuperación de Errores

Las entradas de un traductor pueden contener errores. El lenguaje `yapp` proporciona un *token* especial, `error`, que puede ser utilizado en el programa fuente para extender el traductor con "producciones de error" que lo doten de cierta capacidad para recuperase de una entrada errónea y poder continuar analizando el resto de la entrada.

Consideremos lo que ocurre al ejecutar nuestra calculadora `yapp` con una entrada errónea. Recordemos la gramática:

```
 9  %right  '='
10  %left   '-' '+'
11  %left   '*' '/'
12  %left   NEG
13  %right  '^'
14
```

```
15  %%
16  input:  # empty
17          |   input line  { push(@{$_[1]},$_[2]); $_[1] }
18  ;
19
20  line:       '\n'       { $_[1] }
21          |   exp '\n'   { print "$_[1]\n" }
22          |   error '\n' { $_[0]->YYErrok }
23  ;
```

La regla `line` → `error '\n'` es una producción de error. La idea general de uso es que, a traves de la misma, el programador le indica a `yapp` que, cuando se produce un error dentro de una expresión, descarte todos los *tokens* hasta llegar al retorno del carro y prosiga con el análisis. Además, mediante la llamada al método `YYErrok` el programador anuncia que, si se alcanza este punto, la recuperación puede considerarse "completa" y que `yapp` puede emitir a partir de ese momento mensajes de error con la seguridad de que no son consecuencia de un comportamiento inestable provocado por el primer error.

El resto de la gramática de la calculadora era como sigue:

```
24
25  exp:        NUM
26          |   VAR                 { $_[0]->YYData->{VARS}{$_[1]} }
27          |   VAR '=' exp         { $_[0]->YYData->{VARS}{$_[1]}=$_[3] }
28          |   exp '+' exp         { $_[1] + $_[3] }
29          |   exp '-' exp         { $_[1] - $_[3] }
30          |   exp '*' exp         { $_[1] * $_[3] }
31          |   exp '/' exp         {
32                                    $_[3]
33                                    and return($_[1] / $_[3]);
34                                    $_[0]->YYData->{ERRMSG}
35                                      =   "Illegal division by zero.\n";
36                                    $_[0]->YYError;
37                                    undef
38                                  }
39          |   '-' exp %prec NEG   { -$_[2] }
40          |   exp '^' exp         { $_[1] ** $_[3] }
41          |   '(' exp ')'         { $_[2] }
42  ;
```

en la ejecución activamos el flag `yydebug` a `0x10` para obtener información sobre el tratamiento de errores:

```
$self->YYParse( yylex => \&_Lexer, yyerror => \&_Error, yydebug => 0x10 );
```

Pasemos a darle una primera entrada errónea:

```
$ ./usecalc.pl
3-+2
Syntax error.
**Entering Error recovery.
**Pop state 12.
**Pop state 3.
**Shift $error token and go to state 9.
**Dicard invalid token >+<.
**Pop state 9.
```

```
**Shift $error token and go to state 9.
**Dicard invalid token >NUM<.
**Pop state 9.
**Shift $error token and go to state 9.
**End of Error recovery.
```

El esquema general del algoritmo de recuperación de errores usado por la versión actual de `yapp` es el siguiente:

1. Cuando se encuentra ante una acción de error, el analizador genera un *token* `error`.

2. A continuación pasa a retirar estados de la pila hasta que descubre un estado capaz de transitar ante el *token* `error`. En el ejemplo anterior el analizador estaba en el estado 12 y lo retira de la pila. Los contenidos del estado 12 son:

   ```
   exp -> exp '-' . exp (Rule 10)
   '('shift 7 '-'shift 2 NUM shift 6 VAR shift 8
   exp go to state 21
   ```

   Obviamente no esperabamos ver un `'+'` aqui. El siguiente estado en la cima de la pila es el 3, el cual tampoco tiene ninguna transición ante el terminal `error`:

   ```
   line -> exp . '\n'(Rule 4)
   exp -> exp . '+' exp (Rule 9)
   exp -> exp . '-' exp (Rule 10)
   exp -> exp . '*' exp (Rule 11)
   exp -> exp . '/' exp (Rule 12)
   exp -> exp . '^' exp (Rule 14)

   '*'shift 17 '+'shift 13 '-'shift 12 '/'shift 15
   '\n'shift 14 '^'shift 16
   ```

   El pop sobre el estado 3 deja expuesto en la superficie al estado 1, el cuál "sabe" como manejar el error:

   ```
   $start -> input . $end (Rule 0)
   input -> input . line (Rule 2)

   $end shift 4 '('shift 7 '-'shift 2 '\n'shift 5
   NUM shift 6 VAR shift 8
   error shift 9

   exp go to state 3
   line go to state 10
   ```

3. En este punto transita al estado correspondiente a desplazar el *token* `error`.

   En consecuencia, con lo dicho, en el ejemplo se va al estado 9:

   ```
   line -> error . '\n'(Rule 5)
   '\n'shift, and go to state 20
   ```

4. Entonces el algoritmo de recuperación va leyendo *tokens* y descartandolos hasta encontrar uno que sea aceptable. En este caso hemos especificado que el terminal que nos da cierta confianza de recuperación es el retorno de carro:

Construcción	EBNF	yapp

secuencia opcional	x:{y}	`x : /* null */`
		`  | x y  { $_[0]->YYErrok; }`
		`  | x error`

secuencia	x:y{y}	`x : y`
		`  | xy  { $_[0]->YYErrok; }`
		`  | error`
		`  | x error`

lista	x:y{Ty}	`x : y`
		`  | x T y { $_[0]->YYErrok; }`
		`  | error`
		`  | x error`
		`  | x error y { $_[0]->YYErrok; }`
		`  | x  T error`

Cuadro 35.2: Recuperación de errores en listas

```
**Dicard invalid token >+<.
**Pop state 9.
**Shift $error token and go to state 9.
**Dicard invalid token >NUM<.
**Pop state 9.
**Shift $error token and go to state 9.
**End of Error recovery.
```

5. Sólo se envían nuevos mensajes de error una vez asimilados (desplazados) algunos símbolos terminales. De este modos se intenta evitar la aparición masiva de mensajes de error.

35.16. Recuperación de Errores en Listas

Aunque no existe un método exacto para decidir como ubicar las reglas de recuperación de errores, en general, los símbolos de error deben ubicarse intentado satisfacer las siguientes reglas:

- Tan cerca como sea posible del símbolo de arranque.
- Tan cerca como sea posible de los símbolos terminales.
- Sin introducir nuevos conflictos.

En el caso particular de las listas, se recomienda seguir el siguiente esquema:

Ejercicio 35.16.1. *Compruebe el funcionamiento de la metodología para la recuperación de errores en listas presentada en la tabla 35.2 estudie el siguiente programa* `yapp` *siguiendo la traza de estados, generando entradas con todos los tipos de error posibles. ¿Cómo se recupera el analizador en caso de existencia de un segundo error? ¿Que ocurre si dos errores consecutivos están muy próximos? El programa corresponde al tercer caso de la tabla 35.2, el caso x:y{ Ty} con* $x =$ `list`*,* $T =$ `','` *e* $y =$ `NUMBER`*:*

```
%token NUMBER
%%
command
  :
```

```
  | command list '\n' { $_[0]->YYErrok; }
  ;

list
    : NUMBER            { put($1); }
    | list ',' NUMBER   { put($3); $_[0]->YYErrok; }
    | error             { err(1); }
    | list error        { err(2); }
    | list error NUMBER { err(3); put($3); $_[0]->YYErrok; }
    | list ',' error    { err(4); }
    ;

%%
sub put { my $x = shift; printf("%2.1lf\n",$x); }
sub err { my $code = shift; printf("err %d\n",$code); }
 ...
```

35.17. Consejos a seguir al escribir un programa `yapp`

Cuando escriba un programa `yapp` asegurese de seguir los siguientes consejos:

1. Coloque el punto y coma de separación de reglas en una línea aparte. Un punto y coma "pegado" al final de una regla puede confundirse con un terminal de la regla.
2. Si hay una regla que produce vacío, coloquela en primer lugar y acompáñela de un comentario resaltando ese hecho.
3. Nunca escriba dos reglas de producción en la misma línea.
4. Sangre convenientemente todas las partes derechas de las reglas de producción de una variable, de modo que queden alineadas.
5. Ponga nombres representativos a sus variables sintácticas. No llame `Z` a una variable que representa el concepto "lista de parámetros", llámela `ListaDeParametros`.
6. Es conveniente que declare los terminales simbólicos, esto es, aquellos que llevan un identificador asociado. Si no llevan prioridad asociada o no es necesaria, use una declaración `%token`. De esta manera el lector de su programa se dará cuenta rápidamente que dichos identificadores no se corresponden con variables sintácticas. Por la misma razón, si se trata de terminales asociados con caracteres o cadenas no es tan necesario que los declare, a menos que, como en el ejemplo de la calculadora para `'+'` y `'*'`, sea necesario asociarles una precedencia.
7. Es importante que use la opción `-v` para producir el fichero `.output` conteniendo información detallada sobre los conflictos y el autómata. Cuando haya un conflicto shift-reduce no resuelto busque en el fichero el estado implicado y vea que LR(0) items $A \to \alpha_{\uparrow}$ y $B \to \beta_{\uparrow}\gamma$ entran en conflicto.
8. Si según el informe de `yapp` el conflicto se produce ante un terminal a, es porque $a \in FOLLOW(A)$ y $a \in FIRST(\gamma)$. Busque las causas por las que esto ocurre y modifique su gramática con vistas a eliminar la presencia del terminal a en uno de los dos conjuntos implicados o bien establezca reglas de prioridad entre los terminales implicados que resuelvan el conflicto.
9. Nótese que cuando existe un conflicto de desplazamiento reducción entre $A \to \alpha_{\uparrow}$ y $B \to \beta_{\uparrow}\gamma$, el programa `yapp` contabiliza un error por cada terminal $a \in FOLLOW(A) \cap FIRST(\gamma)$. Por esta razón, si hay 16 elementos en $FOLLOW(A) \cap FIRST(\gamma)$, el analizador `yapp` informará de la existencia de 16 conflictos *shift-reduce*, cuando en realidad se trata de uno sólo. No desespere, los conflictos "auténticos" suelen ser menos de los que `yapp` anuncia.

10. Si necesita declarar variables globales, inicializaciones, etc. que afectan la conducta global del analizador, escriba el código correspondiente en la cabecera del analizador, protegido por los delimitadores `%{` y `%}`. Estos delimitadores deberán aparecer en una línea aparte. Por ejemplo:

```
%{
our contador = 0;
%}

%token NUM
...
%%
```

11. Si tiene problemas en tiempo de ejecución con el comportamiento del análizador sintáctico use la opción `yydebug => 0x1F` en la llamada al analizador.

12. Si trabaja en windows y pasa los ficheros a unix tenga cuidado con la posible introducción de caractéres espúreos en el fichero. Debido a la presencia de caractéres de control invisibles, el analizador `yapp` pasará a rechazar una gramatica aparentemente correcta.

13. Sea consciente de que los analizadores sintáctico y léxico mantienen una relación de corutinas en `yapp`: Cada vez que el analizador sintáctico necesita un nuevo terminal para decidir que regla de producción se aplica, llama al analizador léxico, el cuál deberá retornar el siguiente terminal. La estrategia es diferente de la utilizada en el ejemplo usado para el lenguaje Tutu en el capítulo 33. Allí generabamos en una primera fase la lista de terminales. Aquí los terminales se generan de uno en uno y cada vez que se encuentra uno nuevo se retorna al analizador sintáctico. La ventaja que tiene este método es que permite colaborar al analizador sintáctico y al analizador léxico para "dinámicamente" modificar la conducta del análisis léxico. Por ejemplo en los compiladores del lenguaje C es común hacer que el analizador léxico cuando descubre un identificador que previamente ha sido declarado como identificador de tipo (mediante el uso de `typedef`) retorne un terminal `TYPENAME` diferente del terminal `ID` que caracteriza a los identificadores. Para ello, el analizador sintáctico, cuando detecta una tal declaración, "avisa" al analizador léxico para que modifique su conducta. El analizador sintáctico volverá a avisarlo cuando la declaración del identificador como identificador de tipo salga de ámbito y pierda su especial condición.

14. En `yapp` el analizador sintáctico espera que el analizador léxico devuelva de cada vez una pareja formada por dos escalares. El primer escalar es la cadena que designa el terminal. A diferencia de la habitual costumbre `yacc` de codificar los terminales como enteros, en `yapp` se suelen codificar como cadenas. La segunda componente de la pareja es el atributo asociado con el terminal. Si el atributo es un atributo complejo que necesitas representar mediante un hash o un vector, lo mejor es hacer que esta componente sea una referencia al objeto describiendo el atributo. El analizador léxico le indica al sintáctico la finalización de la entrada enviándole la pareja `('',undef)` formada por la palabra vacía con atributo `undef`.

15. Hay fundamentalmente dos formas de hacer el analizador léxico: hacerlo destructivo o no destructivo. En los destructivos se usa el operador de sustitución `s` (véase el ejemplo de la sección 35.1), en cuyo caso la entrada procesada es retirada de la cadena leída. En los no destructivos utilizamos el operador de emparejamiento `m`. Véase el ejemplo de analizador léxico en la sección 35.13 (concretamente la subrutina `scanner` en la línea 20 del fichero `useinherited.pl`)

Ejemplo 35.17.1. *Consideremos de nuevo el programa* `yapp` *para producir árboles para las expresiones en infijo. Supongamos que olvidamos introducir una prioridad explícita al terminal* `'='`*:*

```
$ cat -n Infixtree_conflict.yp
 1  #
 2  # Infixtree.yp
```

```
 3  #
 4
 5  %{
 6  use Data::Dumper;
 7  %}
 8  %left   '-' '+'
 9  %left   '*' '/'
10  %left   NEG
11
12  %%
13  input:  #empty
14          |   input line
15  ;
16
17  line:     '\n'         { $_[1] }
18          | exp '\n'     { print Dumper($_[1]); }
19          | error '\n'   { $_[0]->YYErrok }
20  ;
21
22  exp:        NUM
23          |   VAR                { $_[1] }
24          |   VAR '=' exp        { bless [$_[1], $_[3]], 'ASSIGN' }
25          |   exp '+' exp        { bless [$_[1], $_[3] ], 'PLUS'}
26          |   exp '-' exp        { bless [$_[1], $_[3] ], 'MINUS'}
27          |   exp '*' exp        { bless [$_[1], $_[3]], 'TIMES' }
28          |   exp '/' exp        { bless [$_[1], $_[3]], 'DIVIDE' }
   ....
```

en este caso al compilar encontraremos conflictos:

```
$ yapp -v -m Infixtree Infixtree_conflict.yp
4 shift/reduce conflicts
```

En tal caso lo que debemos hacer es editar el fichero `.output`*. El comienzo del fichero es como sigue:*

```
$ cat -n Infixtree_conflict.output
1  Warnings:
2  ---------
3  4 shift/reduce conflicts
4
5  Conflicts:
6  ----------
7  Conflict in state 11 between rule 13 and token '-' resolved as reduce.
8  Conflict in state 11 between rule 13 and token '*' resolved as reduce.
  ...
```

Tal y como indica la expresión ...`resolved as` *..., las líneas como la 7, la 8 y siguientes se refieren a conflictos resueltos. Mas abajo encontraremos información sobre la causa de nuestros conflictos no resueltos:*

```
    ...
26  Conflict in state 23 between rule 11 and token '/' resolved as reduce.
27  State 25 contains 4 shift/reduce conflicts
```

Lo que nos informa que los conflictos ocurren en el estado 25 ante 4 terminales distintos. Nos vamos a la parte del fichero en la que aparece la información relativa al estado 25. Para ello, como el fichero es grande, buscamos por la cadena adecuada. En `vi` *buscaríamos por* `/^State 25`*. Las líneas correspondientes contienen:*

```
291 State 25:
292
293   exp -> VAR '=' exp .  (Rule 8)
294   exp -> exp . '+' exp  (Rule 9)
295   exp -> exp . '-' exp  (Rule 10)
296   exp -> exp . '*' exp  (Rule 11)
297   exp -> exp . '/' exp  (Rule 12)
298
299   '*' shift, and go to state 16
300   '+' shift, and go to state 13
301   '-' shift, and go to state 12
302   '/' shift, and go to state 15
303
304   '*' [reduce using rule 8 (exp)]
305   '+' [reduce using rule 8 (exp)]
306   '-' [reduce using rule 8 (exp)]
307   '/' [reduce using rule 8 (exp)]
308   $default  reduce using rule 8 (exp)
```

El comentario en la línea 308 (`$default` *. . .) indica que por defecto, ante cualquier otro terminal que no sea uno de los explícitamente listados, la acción a tomar por el analizador será reducir por la regla 8.*

Una revisión a la numeración de la gramática, al comienzo del fichero `.output` *nos permite ver cuál es la regla 8:*

```
29 Rules:
30 ------
31 0:  $start -> input $end
32 1:  input -> /* empty */
33 2:  input -> input line
34 3:  line -> '\n'
35 4:  line -> exp '\n'
36 5:  line -> error '\n'
37 6:  exp -> NUM
38 7:  exp -> VAR
39 8:  exp -> VAR '=' exp
40 9:  exp -> exp '+' exp
41 10: exp -> exp '-' exp
42 11: exp -> exp '*' exp
43 12: exp -> exp '/' exp
44 13: exp -> '-' exp
45 14: exp -> '(' exp ')'
```

Efectivamente, es la regla de asignación `exp -> VAR '=' exp`*. El conflicto aparece por que los terminales* `* + - /` *están en el conjunto FOLLOW(*`exp`*) y también cabe esperarlos respectivamente en las reglas 9, 10, 11 y 12 ya que el estado 25 contiene:*

```
294   exp -> exp . '+' exp  (Rule 9)
295   exp -> exp . '-' exp  (Rule 10)
296   exp -> exp . '*' exp  (Rule 11)
297   exp -> exp . '/' exp  (Rule 12)
```

Estamos ante un caso en el que se aplica el consejo número 8. Los items de la forma $B \rightarrow \beta_{\uparrow}\gamma$*, son los de la forma* `exp -> exp . '+' exp`*, etc. El item de la forma* $A \rightarrow \alpha_{\uparrow}$ *es en este caso* `exp -> VAR '=' exp`*.*

En efecto, en una expresión como `a = 4 + 3` *se produce una ambiguedad. ¿Debe interpretarse como* `(a = 4) + 3`*? ¿O bien como* `a = (4 + 3)`*?. La primera interpretación corresponde a reducir por la regla 8. La segunda a desplazar al estado 13. En este ejemplo, el conflicto se resuelve haciendo que tenga prioridad el desplazamiento, dando menor prioridad al terminal* `=` *que a los terminales* `* + - /`*.*

Ejercicio 35.17.1. *¿Que ocurre en el ejemplo anterior si dejamos que* `yapp` *aplique las reglas por defecto?*

35.18. Práctica: Un C simplificado

Escriba un analizador sintáctico usando `Parse::Yapp` para el siguiente lenguaje. La descripción utiliza una notación tipo BNF: las llaves indican 0 o mas repeticiones y los corchetes opcionalidad.

program	→	definitions { definitions }
definitions	→	datadefinition | functiondefinition
datadefinition	→	basictype declarator { ',' declarator } ';'
declarator	→	ID { '[' constantexp ']' }
functiondefinition	→	[basictype] functionheader functionbody
basictype	→	INT | CHAR
functionheader	→	ID '(' [parameters] ')'
parameters	→	basictype declarator { ',' basictype declarator }
functionbody	→	'{' { datadefinition } { statement } '}'
statement	→	[exp] ';'
		| '{' { datadefinition } { statement } '}'
		| IF '(' exp ')' statement [ELSE statement]
		| WHILE '(' exp ')' statement
		| RETURN [exp] ';'
constantexp	→	exp
exp	→	lvalue '=' exp | lvalue '+=' exp
		| exp '&&' exp | exp '||' exp |
		| exp '==' exp | exp '!=' exp |
		| exp '<' exp | exp '>' exp | exp '<=' exp | exp '>=' exp |
		| exp '+' exp | exp '-' exp |
		| exp '*' exp | exp '/' exp |
		| unary
unary	→	'++' lvalue | '−−' lvalue | primary
primary	→	'(' exp ')' | ID '(' [argumentlist] ')' | lvalue | NUM | CHARACTER
lvalue	→	ID { '[' exp ']' }
argumentlist	→	exp { ',' exp }

Su analizador, además de seguir los consejos explícitados en la sección 35.17, deberá cumplir las siguientes especificaciones:

1. **Método de Trabajo**

   Parta de la definición BNF y proceda a introducir las reglas poco a poco:

```
1 %token declarator basictype functionheader functionbody
2 %%
3 program: definitionslist
4        ;
5
6 definitionslist: definitions definitionslist
7                | definitions
```

```
8               ;
9
10 definitions: datadefinition
11            | functiondefinition
12            ;
13 datadefinition: basictype declaratorlist ';'
14              ;
15
16 declaratorlist: declarator ',' declaratorlist
17              | declarator
18              ;
19 functiondefinition: basictype functionheader functionbody
20                   | functionheader functionbody
21                   ;
22
23 %%
```

a) Comienze trabajando en el cuerpo de la gramática.

b) Olvídese del analizador léxico por ahora. Su objetivo es tener una gramática limpia de conflictos y que reconozca el lenguaje dado.

c) Sustituya las repeticiones BNF por listas. Si una variable describe una lista de `cosas` llámela `cosaslist`.

d) Si tiene un elemento opcional en la BNF, por ejemplo, en la regla:
functiondefinition → [basictype] functionheader functionbody
sustitúyala por dos reglas una en la que aparece el elemento y otra en la que no.

```
19 functiondefinition: basictype functionheader functionbody
20                   | functionheader functionbody
21                   ;
```

e) Cada par de reglas que introduzca vuelva a recompilar con `yapp` la gramática para ver si se han producido conflictos. Cuando estoy editando la gramática suelo escribir a menudo la orden
`:!yapp %`
para recompilar:

```
     15
     16 declaratorlist: declarator declaratorlist
     17                | declarator
     18                ;
     19 functiondefinition: basictype functionheader functionbody
     20                    | functionheader functionbody
     21                    ;
     22
     23 %%
~
~
~
~
~
~
~
~
~
~
:!yapp %
```

Esto llama a `yapp` con el fichero bajo edición. Si hay errores los detectaré enseguida.

f) Insisto, procure detectar la aparición de un conflicto lo antes posible. Es terrible tener que limpiar una gramática llena de conflictos.

g) Ponga nombres significativos a las variables y terminales. Por favor, no los llame `d1`, `d2`, etc.

h) Cuando esté en el proceso de construcción de la gramática y aún le queden por rellenar variables sintácticas, declárelas como terminales mediante `%token` como en el código que aparece encima. De esta manera evitará las quejas de `yapp`.

2. **Resolución de Conflictos**

Las operaciones de asignación tienen la prioridad mas baja, seguidas de las lógicas, los test de igualdad y después de los de comparación, a continuación las aditivas, multiplicativas y por último los `unary` y `primary`. Exprese la asociatividad natural y la prioridad especificada usando los mecanismos que `yapp` provee al efecto.

La gramática es ambigua, ya que para una sentencia como

$$\text{if } E_1 \text{ then if } E_2 \text{ then } S_1 \text{ else } S_2$$

existen dos árboles posibles: uno que asocia el "else" con el primer "if" y otra que lo asocia con el segundo. Los dos árboles corresponden a las dos posibles parentizaciones:

$$\text{if } E_1 \text{ then (if } E_2 \text{ then } S_1 \text{ else } S_2)$$

Esta es la regla de prioridad usada en la mayor parte de los lenguajes: un "else" casa con el "if" mas cercano. La otra posible parentización es:

$$\text{if } E_1 \text{ then (if } E_2 \text{ then } S_1) \text{ else } S_2$$

Utilice los mecanismos de priorización proporcionados por `yapp` para resolver el conflicto shift-reduce generado. ¿Es correcta en este caso particular la conducta a la que da lugar la acción `yapp` por defecto?

3. **Analizador Léxico**

   Además del tipo de terminal y su valor el analizador léxico deberá devolver el número de línea. El analizador léxico deberá aceptar comentarios C. En la gramática, el terminal `CHARACTER` se refiere a caracteres entre comillas simples (por ejemplo `'a'`).

   Se aconseja que las palabras reservadas del lenguaje no se traten con expresiones regulares específicas sino que se capturen en el patrón de identificador `[a-z_]\w+`. Se mantiene para ello un hash con las palabras reservadas que es inicializado al comienzo del programa. Cuando el analizador léxico encuentra un identificador mira en primer lugar en dicho hash para ver si es una palabra reservada y, si lo es, devuelve el terminal correspondiente. En caso contrario se trata de un identificador.

4. **Recuperación de Errores**

   Extienda la práctica con reglas para la recuperación de errores. Para las listas, siga los consejos dados en la sección 35.2. En aquellos casos en los que la introducción de las reglas de recuperación produzca ambiguedad, resuelva los conflictos.

5. **Árbol de Análisis Abstracto**

   La semántica del lenguaje es similar a la del lenguaje C (por ejemplo, las expresiones lógicas se tratan como expresiones enteras). El analizador deberá producir un árbol sintáctico abstracto. Como se hizo para el lenguaje Tutu introducido en el capítulo 33, cada clase de nodo deberá corresponderse con una clase Perl. Por ejemplo, para una regla como

   ```
   exp '*' exp
   ```

   la acción asociada sería algo parecido a

   ```
   { bless [ $_[1], $_[3]], 'MULT' }
   ```

   donde usamos un array anónimo. Mejor aún es usar un hash anónimo:

   ```
   { bless { LEFT => $_[1], RIGHT => $_[3]}, 'MULT' }
   ```

   Defina formalmente el arbol especificando la gramática árbol correspondiente a su diseño (repase la sección 33.9.1). Introduzca en esta parte la tabla de símbolos. La tabla de símbolos es, como en el compilador de Tutu, una lista de referencias a hashes conteniendo las tablas de símbolos locales a cada bloque. En cada momento, la lista refleja el anidamiento de bloques actual. Es posible que, en la declaración de funciones, le interese crear un nuevo bloque en el que guardar los parámetros, de manera que las variables globales queden a nivel 0, los parámetros de una función a nivel 1 y las variables locales de la función a nivel 2 o superior.

35.19. La Gramática de `yapp` / `yacc`

En esta sección veremos con mas detalle, la sintaxis de `Parse::Yapp`, usando la propia notación `yapp` para describir el lenguaje. Un programa `yapp` consta de tres partes: la cabeza, el cuerpo y la cola. Cada una de las partes va separada de las otras por el símbolo `%%` en una línea aparte.

```
yapp:   head body tail
head:   headsec '%%'
headsec: #empty
    |   decls
decls:  decls decl  | decl
body:   rulesec '%%'
rulesec:  rulesec rules | rules
rules:  IDENT ':' rhss ';'
tail:   /*empty*/
    |   TAILCODE
```

35.19.1. La Cabecera

En la cabecera se colocan las declaraciones de variables, terminales, etc.

```
decl:  '\n'
    |   TOKEN typedecl symlist '\n'
    |   ASSOC typedecl symlist '\n'
    |   START ident '\n'
    |   HEADCODE '\n'
    |   UNION CODE '\n'
    |   TYPE typedecl identlist '\n'
    |   EXPECT NUMBER '\n'

typedecl: # empty
    |   '<' IDENT '>'
```

El terminal `START` se corresponde con una declaración `%start` indicando cual es el símbolo de arranque de la gramática. Por defecto, el símbolo de arranque es el primero de la gramática.

El terminal `ASSOC` está por los terminales que indican precedencia y asociatividad. Esto se ve claro si se analiza el contenido del fichero `YappParse.yp` (**??**) en el que se puede encontrar el código del analizador léxico del módulo `Parse::Yapp`. El código dice:

```
    ...
    if($lexlevel == 0) {# In head section
            $$input=~/\G%(left|right|nonassoc)/gc
        and return('ASSOC',[ uc($1), $lineno[0] ]);
            $$input=~/\G%(start)/gc
        and return('START',[ undef, $lineno[0] ]);
            $$input=~/\G%(expect)/gc
        and return('EXPECT',[ undef, $lineno[0] ]);
            $$input=~/\G%{/gc
    ...
```

La variable `$lexlevel` indica en que sección nos encontramos: cabecera, cuerpo o cola. El terminal `EXPECT` indica la presencia de una declaración `%expect` en el fuente, la cual cuando es seguida de un número indica el numero de conflictos shift-reduce que cabe esperar. Use `EXPECT` si quiere silenciar las advertencias de `yapp` sobre la presencia de conflictos cuya resolución automática considere correcta.

35.19.2. La Cabecera: Diferencias entre `yacc` y `yapp`

Las declaraciones de tipo correspondientes a `%union` y a las especificaciones de tipo entre símbolos menor mayor (`<tipo>`) en declaraciones `token` y `%type` no son usadas por `yapp`. Estas declaraciones son necesarias cuando el código de las acciones semánticas se escribe en `C` como es el caso de `yacc` y `bison`. Sigue un ejemplo de programa `yacc`/`bison` que usa declaraciones `%union` y de tipo para los atributos:

```
 1 %{
 2 #include <stdio.h>
 3
 4 #define CLASE(x) ((x == 1)?"global":"local")
 5 #define TIPO(x) ((x == 1)?"float":"integer")
 6 %}
 7
 8 %union {
 9   int n;   /* enumerado */
10   char *s; /* cadena */
11 }
12
13 %token <n> FLOAT INTEGER
14 %token <n> GLOBAL
15 %token <n> LOCAL
16 %token <s> NAME
17 %type <n> class type
18
19 %%
```

La declaración `%union` de la línea 8 indica que los atributos son de dos tipos: enteros y punteros a caracteres. El nombre del campo es posteriormente usado en las declaraciones de las líneas 13-17 para indicar el tipo del atributo asociado con la variable o con el terminal. Así, la declaración de la línea 13 indica que los terminales `FLOAT` e `INTEGER` son de tipo entero, mientras que la declaración de la línea 16 nos dice que el terminal `NAME` es de tipo cadena.

```
29 class
30   : GLOBAL { $$ = 1; }
31   | LOCAL  { $$ = 2; }
32   ;
33
34 type
35   : FLOAT   { $$ = 1; }
36   | INTEGER { $$ = 2; }
37   ;
```

La información proveída sobre los tipos permite a `yacc` introducir automáticamente en el código `C` producido los *typecasting* o ahormados para las asignaciones de las líneas 30-31 y 35-36. Obsérve que en `yacc` el atributo de la variable en la parte izquierda se denota por `$$`.

Otra diferencia entre `yacc` y `yapp` es que en `yacc` los atributos de la parte derecha no constituyen un vector, denotándose por `$1`, `$2`, `$3` ...

En ocasiones `yacc` no puede determinar el tipo de un atributo. En particular cuando se habla del atributo asociado con una acción intermedia, ya que esta no tiene variable sintáctica asociada explícitamente o bien cuando se habla de los atributos de símbolos que están a la izquierda de la reducción actual (véase la sección 35.13). Los atributos de símbolos a la izquierda de la producción actual se denotan en `yacc` por números no positivos `$0`, `$-1`, `$-2` ....

En estos casos el programador deberá especificar explícitamente el tipo del atributo usando la notación `$<tipo>#`. Donde `tipo` es uno de los campos de la `union` y `#` es el numeral del símbolo correspondiente:

```
39 namelist
40   : NAME   { printf("%s de clase %s, tipo %s\n",$1,CLASE($<n>-1),TIPO($<n>0)); }
41   | namelist ',' NAME
42       { printf("%s de clase %s, tipo %s\n",$3,CLASE($<n>-1),TIPO($<n>0)); }
43   ;
44 %%
```

35.19.3. El Cuerpo

El cuerpo de un programa `yapp` contiene la gramática y las acciones

```
rhss:   rhss '|' rule | rule
rule:   rhs prec epscode | rhs
rhs:    #empty
    |   rhselts
rhselts: rhselts rhselt | rhselt
rhselt: symbol | code
prec: PREC symbol
epscode: # vacio
    |   code
code:   CODE
```

Las acciones semánticas (variable sintáctica `code` y terminal `CODE`) se ejecutan siempre que ocurre una reducción por una regla y, en general, devuelven un valor semántico. El código de la acción se copia verbatim en el analizador. La estrategia usada por el analizador léxico es contar las llaves abrir y cerrar en el texto. Véase el correspondiente fragmento del analizador léxico:

```
    ....
    $lineno[0]=$lineno[1];
    ....
    $$input=~/\G{/gc
    and do {
        my($level,$from,$code);

        $from=pos($$input);
        $level=1;
        while($$input=~/([{}])/gc) {
            substr($$input,pos($$input)-1,1) eq '\\' #Quoted
            and next;
            $level += ($1 eq '{' ? 1 : -1) or last;
        }
        $level and  _SyntaxError(2,"Unmatched { opened line $lineno[0]",-1);
        $code = substr($$input,$from,pos($$input)-$from-1);
        $lineno[1]+= $code=~tr/\n//;
        return('CODE',[ $code, $lineno[0] ]);
    };
```

Las llaves dentro de cadenas y comentarios no son significativas en la cuenta. El problema es que el reconocimiento de cadenas en Perl es mas difícil que en otros lenguajes: existe toda una variedad de formas de denotar una cadena. Por tanto, si el programador usuario de `yapp` necesita escribir una llave dentro de una cadena de doble comilla, deberá escaparla. Si la cadena es de simple comilla escaparla no es solución, pues aparecería el símbolo de escape en la cadena. En ese caso se deberá añadir un comentario con la correspondiente falsa llave. Siguen algunos ejemplos tomadados de la documentación de `Parse::Yapp`

```
"{ My string block }"
"\{ My other string block \}"
qq/ My unmatched brace \} /

# Casamos con el siguiente: {
q/ for my closing brace } / #

q/ My opening brace { /
# debe cerrarse: }
```

Ejercicio 35.19.1. *Genere programas de prueba* `yapp` *con cadenas que produzcan confusión en el analizador y observe el comportamiento. Pruébelas en las diferentes secciones en las que puede ocurrir código: en la cabecera, en el cuerpo y en la cola.*

35.19.4. La Cola: Diferencias entre `yacc` y `yapp`

La cola de un program `yapp` contiene las rutinas de soporte.

```
tail:   /*empty*/
    |   TAILCODE
```

el terminal `TAILCODE` al igual que los terminales `CODE` y `HEADCODE` indican que en ese punto se puede encontrar código Perl. La detección de `TAILCODE` y `HEADCODE` son mas sencillas que las de `CODE`.

La cola de un programa `yacc` es similar. Para el programa `yacc` cuya cabecera y cuerpo se mostraron en la sección 35.19.2 la cola es:

```
 1 %%
 2
 3 extern FILE * yyin;
 4
 5 main(int argc, char **argv) {
 6   if (argc > 1) yyin = fopen(argv[1],"r");
 7   /* yydebug = 1;
 8   */
 9   yyparse();
10 }
11
12 yyerror(char *s) {
13   printf("%s\n",s);
14 }
```

La declaración del manejador de fichero `yyin` en la línea 14 referencia el archivo de entrada para el analizador. La variable (comentada, línea 7) `yydebug` controla la información para la depuración de la gramática. Para que sea realmente efectiva, el programa deberá además compilarse definiendo la macro `YYDEBUG`. Sigue un ejemplo de `Makefile`:

```
 1   inherited: y.tab.c lex.yy.c
 2      gcc -DYYDEBUG=1 -g -o inherited1 y.tab.c lex.yy.c
 3   y.tab.c y.tab.h: inherited1.y
 4      yacc -d -v inherited1.y
 5   lex.yy.c: inherited1.l y.tab.h
 6      flex -l inherited1.l
 7   clean:
 8      - rm -f y.tab.c lex.yy.c *.o core inherited1
```

Al compilar tenemos:

```
pl@nereida:~/src/inherited$ make
yacc -d -v inherited1.y
flex -l inherited1.l
gcc -DYYDEBUG=1 -g -o inherited1 y.tab.c lex.yy.c
pl@nereida:~/src/inherited$ ls -ltr
total 232
-rw-r-----  1 pl users   242 Dec 10  2003 Makefile
-rw-r-----  1 pl users   404 Dec 10  2003 inherited1.l
-rw-r-----  1 pl users   878 Dec 10  2003 inherited1.y
```

```
-rw-rw----  1 pl users  1891 Jan 26 15:41 y.tab.h
-rw-rw----  1 pl users 30930 Jan 26 15:41 y.tab.c
-rw-rw----  1 pl users  2365 Jan 26 15:41 y.output
-rw-rw----  1 pl users 44909 Jan 26 15:41 lex.yy.c
-rwxrwx--x  1 pl users 56336 Jan 26 15:41 inherited1
```

35.19.5. El Análisis Léxico en `yacc`: `flex`

El analizador léxico para `yacc` desarrollado en las secciones anteriores ha sido escrito usando la variante `flex` del lenguaje `LEX`. Un programa `flex` tiene una estructura similar a la de un program `yacc` con tres partes: cabeza, cuerpo y cola separados por `%%`. Veamos como ejemplo de manejo de `flex`, los contenidos del fichero flex `inherited1.l` utilizado en las secciones anteriores:

```
 1   %{
 2   #include <string.h>
 3   #include "y.tab.h"
 4   %}
 5   id [A-Za-z_][A-Za-z_0-9]*
 6   white [ \t\n]+
 7   %%
 8   global    {  return GLOBAL; }
 9   local     {  return LOCAL; }
10   float     {  return FLOAT; }
11   int       {  return INTEGER; }
12   {id}      {  yylval.s = strdup(yytext); return NAME; }
13   {white}   { ; }
14   ,         { return yytext[0]; }
15   .         { fprintf(stderr,"Error. carácter inesperado.\n"); }
16   %%
17   int yywrap() { return 1; }
```

La cabeza contiene declaraciones `C` asi como definiciones regulares. El fichero `y.tab.h` que es incluído en la línea 3, fué generado por `yacc` y contiene, entre otras cosas, la información recolectada por `yacc` sobre los tipos de los atributos (declaración `%union`) y la enumeración de los terminales. Es, por tanto, necesario que la compilación con `yacc` preceda a la compilación con `flex`. La información en `y.tab.h` es usada por el analizador léxico para "sincronizarse" con el analizador sintáctico. Se definen en las líneas 5 y 6 las macros para el reconocimiento de identificadores (`id`) y blancos (`white`). Estas macros son llamadas en el cuerpo en las líneas 12 y 13. La estructura del cuerpo consiste en parejas formadas por una definición regular seguidas de una acción. La variable `yylval` contiene el atributo asociado con el terminal actual. Puesto que el token `NAME` fué declarado del tipo cadena (véase 35.19.2), se usa el correspondiente nombre de campo `yylval.s`. La cadena que acaba de casar queda guardada en la variable `yytext`, y su longitud queda en la variable entera global `yyleng`.

Una vez compilado con `flex` el fuente, obtenemos un fichero denominado `lex.yy.c`. Este fichero contiene la rutina `yylex()` que realiza el análisis léxico del lenguaje descrito.

La función `yylex()` analiza las entradas, buscando la secuencia mas larga que casa con alguna de las expresiones regulares y ejecuta la correspondiente acción. Si no se encuentra ningun emparejamiento se ejecuta la regla "por defecto", que es:

```
(.|\n)  { printf("%s",yytext); }
```

Si encuentran dos expresiones regulares con las que la cadena mas larga casa, elige la que figura primera en el programa `flex`.

Una vez que se ha ejecutado la correspondiente acción, `yylex()` continúa con el resto de la entrada, buscando por subsiguientes emparejamientos. Asi continúa hasta encontrar un final de fichero en cuyo

caso termina, retornando un cero o bien hasta que una de las acciones explicitamente ejecuta una sentencia `return`.

Cuando el analizador léxico alcanza el final del fichero, el comportamiento en las subsiguientes llamadas a `yylex` resulta indefinido. En el momento en que `yylex` alcanza el final del fichero llama a la función `yywrap`, la cual retorna un valor de 0 o 1 según haya mas entrada o no. Si el valor es 0, la función `yylex` asume que la propia `yywrap` se ha encargado de abrir el nuevo fichero y asignarselo a `yyin`.

35.19.6. Práctica: Uso de `Yacc` y `Lex`

Use `yacc` y `flex` para completar los analizadores sintáctico y léxico descritos en las secciones 35.19.2, 35.19.4 y 35.19.5. La gramática en cuestión es similar a la descrita en la sección 35.13. Usando la variable `yydebug` y la macro `YYDEBUG` analize el comportamiento para la entrada `global float x,y`.

35.20. El Analizador Ascendente `Parse::Yapp`

El program `yapp` es un traductor y, por tanto, constituye un ejemplo de como escribir un traductor. El lenguaje fuente es el lenguaje `yacc` y el lenguaje objeto es `Perl`. Como es habitual en muchos lenguajes, el lenguaje objeto se ve .[ex]pandidoçon un conjunto de funciones de soporte. En el caso de `yapp` estas funciones de soporte, son en realidad métodos y están en el módulo `Parse::Yapp::Driver`. Cualquier módulo generado por `yapp` hereda de dicho módulo (véase por ejemplo, el módulo generado para nuestro ejemplo de la calculadora, en la sección 35.4).

Como se ve en la figura 35.3, los módulos generados por `yapp` heredan y usan la clase `Parse::Yapp::Driver` la cual contiene el analizador sintáctico LR genérico. Este módulo contiene los métodos de soporte visibles al usuario `YYParse`, `YYData`, `YYErrok`, `YYSemval`, etc.

La figura 35.3 muestra además el resto de los módulos que conforman el "compilador" `Parse::Yapp`. La herencia se ha representado mediante flechas contínuas. Las flechas punteadas indican una relación de uso entre los módulos. El guión `yapp` es un programa aparte que es usado para producir el correspondiente módulo desde el fichero conteniendo la gramática.

Figura 35.3: Esquema de herencia de `Parse::Yapp`. Las flechas contínuas indican herencia, las punteadas uso. La clase `Calc` es implementada en el módulo generado por `yapp`

(Para ver el contenido de los módulos, descarge `yapp` desde CPAN:

`\http://search.cpan.org/~fdesar/Parse-Yapp-1.05/lib/Parse/Yapp.p`

o bien desde uno de nuestros servidores locales; en el mismo directorio en que se guarda la versión HTML de estos apuntes encontrará una copia de Parse-Yapp-1.05.tar.gz). La versión a la que se refiere este capítulo es la 1.05.

El módulo `Parse/Yapp/Yapp.pm` se limita a contener la documentación y descansa toda la tarea de análisis en los otros módulos. El módulo `Parse/Yapp/Output.pm` contiene los métodos `_CopyDriver` y `Output` los cuales se encargan de escribir el analizador: partiendo de un esqueleto genérico rellenan las partes específicas a partir de la información computada por los otros módulos.

El módulo `Parse/Yapp/Options.pm` analiza las opciones de entrada. El módulo `Parse/Yapp/Lalr.pm` calcula las tablas de análisis LALR. Por último el módulo `Parse/Yapp/Grammar` contiene varios métodos de soporte para el tratamiento de la gramática.

El modulo `Parse::Yapp::Driver` contiene el método `YYparse` encargado del análisis. En realidad, el método `YYparse` delega en el método privado `_Parse` la tarea de análisis. Esta es la estructura del analizador genérico usado por `yapp`. Léalo con cuidado y compare con la estructura explicada en la sección 35.5.

```
1 sub _Parse {
2   my($self)=shift;
```

```
 3
 4   my($rules,$states,$lex,$error)
 5      = @$self{ 'RULES', 'STATES', 'LEX', 'ERROR' };
 6   my($errstatus,$nberror,$token,$value,$stack,$check,$dotpos)
 7      = @$self{ 'ERRST', 'NBERR', 'TOKEN', 'VALUE', 'STACK', 'CHECK', 'DOTPOS' };
 8
 9   $$errstatus=0;
10   $$nberror=0;
11   ($$token,$$value)=(undef,undef);
12   @$stack=( [ 0, undef ] ); # push estado 0
13   $$check='';
```

La componente 0 de `@$stack` es el estado, la componente 1 es el atributo.

```
14
15   while(1) {
16     my($actions,$act,$stateno);
17
18     $stateno=$$stack[-1][0];     # sacar el estado en el top de
19     $actions=$$states[$stateno]; # la pila
```

`$states` es una referencia a un vector. Cada entrada `$$states[$stateno]` es una referencia a un hash que contiene dos claves. La clave `ACTIONS` contiene las acciones para ese estado. La clave `GOTOS` contiene los saltos correspondientes a ese estado.

```
20
21     if  (exists($$actions{ACTIONS})) {
22       defined($$token) or do {
23         ($$token,$$value)=&$lex($self); # leer siguiente token
24       };
25
26       # guardar en $act la acción asociada con el estado y el token
27       $act = exists($$actions{ACTIONS}{$$token})?
28          $$actions{ACTIONS}{$$token} :
29          exists($$actions{DEFAULT})? $$actions{DEFAULT} : undef;
30     }
31     else { $act=$$actions{DEFAULT}; }
```

La entrada `DEFAULT` de una acción contiene la acción a ejecutar por defecto.

```
32
33     defined($act) and do {
34       $act > 0 and do { # $act >0 indica shift
35          $$errstatus and do { --$$errstatus; };
```

La línea 35 esta relacionada con la recuperación de errores. Cuando `yapp` ha podido desplazar varios terminales sin que se produzca error considerará que se ha recuperado con éxito del último error.

```
36          # Transitar: guardar (estado, valor)
37          push(@$stack,[ $act, $$value ]);
38          $$token ne ''   #Don't eat the eof
39          and $$token=$$value=undef;
40          next; # siguiente iteración
41       };
```

A menos que se trate del final de fichero, se reinicializa la pareja `($$token, $$value)` y se repite el bucle de análisis. Si `$act` es negativo se trata de una reducción y la entrada `$$rules[-$act]` es una referencia a un vector con tres elementos: la variable sintáctica, la longitud de la parte derecha y el código asociado:

```
43        # $act < 0, indica reduce
44        my($lhs,$len,$code,@sempar,$semval);
45
46        #obtenemos una referencia a la variable,
47        #longitud de la parte derecha, referencia
48        #a la acción
49        ($lhs,$len,$code)=@{$$rules[-$act]};
50        $act or $self->YYAccept();
```

Si `$act` es cero indica una acción de aceptación. El método `YYAccept` se encuentra en `Driver.pm`. Simplemente contiene:

```
sub YYAccept {
  my($self)=shift;

  ${$$self{CHECK}}='ACCEPT';
    undef;
}
```

Esta entrada será comprobada al final de la iteración para comprobar la condición de aceptación (a través de la variable `$check`, la cuál es una referencia).

```
51        $$dotpos=$len; # dotpos es la longitud de la regla
52        unpack('A1',$lhs) eq '@'    #In line rule
53        and do {
54          $lhs =~ /^\@[0-9]+\-([0-9]+)$/
55            or  die "In line rule name '$lhs' ill formed: ".
56                    "report it as a BUG.\n";
57          $$dotpos = $1;
58        };
```

En la línea 52 obtenemos el primer carácter en el nombre de la variable. Las acciones intermedias en `yapp` producen una variable auxiliar que comienza por `@` y casa con el patrón especificado en la línea 54. Obsérvese que el número después del guión contiene la posición relativa en la regla de la acción intermedia.

```
60        @sempar = $$dotpos ?
61           map { $$_[1] } @$stack[ -$$dotpos .. -1 ] : ();
```

El array `@sempar` se inicia a la lista vacía si `$len` es nulo. En caso contrario contiene la lista de los atributos de los últimos `$$dotpos` elementos referenciados en la pila. Si la regla es intermedia estamos haciendo referencia a los atributos de los símbolos a su izquierda.

```
62        $semval = $code ? &$code( $self, @sempar ) :
63                          @sempar ? $sempar[0] : undef;
```

Es en este punto que ocurre la ejecución de la acción. La subrutina referenciada por `$code` es llamada con primer argumento la referencia al objeto analizador `$self` y como argumentos los atributos que se han computado previamente en `@sempar`. Si no existe tal código se devuelve el atributo del primer elemento, si es que existe un tal primer elemento.

El valor retornado por la subrutina/acción asociada es guardado en `$semval`.

```
65        splice(@$stack,-$len,$len);
```

La función `splice` toma en general cuatro argumentos: el array a modificar, el índice en el cual es modificado, el número de elementos a suprimir y la lista de elementos extra a insertar. Aquí, la llamada a `splice` cambia los elementos de `@$stack` a partir del índice `-$len`. El número de elementos a suprimir es `$len`. A continuación se comprueba si hay que terminar, bien porque se ha llegado al estado de aceptación (`$$check eq 'ACCEPT'`) o porque ha habido un error fatal:

```
      $$check eq 'ACCEPT' and do { return($semval); };
      $$check eq 'ABORT' and  do { return(undef); };
```

Si las cosas van bien, se empuja en la cima de la pila el estado resultante de transitar desde el estado en la cima con la variable sintáctica en el lado izquierdo:

```
      $$check eq 'ERROR' or  do {
          push(@$stack, [ $$states[$$stack[-1][0]]{GOTOS}{$lhs}, $semval ]);
          $$check='';
          next;
      };
```

La expresión `$$states[$$stack[-1][0]]` es una referencia a un hash cuya clave `GOTOS` contiene una referencia a un hash conteniendo la tabla de transiciones del éstado en la cima de la pila (`$stack[-1][0]`). La entrada de clave `$lhs` contiene el estado al que se transita al ver la variable sintáctica de la izquierda de la regla de producción. El atributo asociado es el devuelto por la acción: `$semval`.

```
      $$check='';

    }; # fin de defined($act)

    # Manejo de errores: código suprimido
    ...

  }
}#_Parse
```

... y el bucle `while(1)` de la línea 15 continúa. Compare este código con el seudo-código introducido en la sección 35.5.

35.21. La Estructura de Datos Generada por `YappParse.yp`

El fichero `YappParse.yp`. contiene la gramática `yapp` del lenguaje `yacc`[1] . Además de las dos rutinas de soporte típicas, la de tratamiento de errores `_Error` y la de análisis léxico `_Lexer`, el fichero contiene una subrutina para el manejo de las reglas `_AddRules` y otra rutina `Parse` la cuál actúa como *wrapper* o filtro sobre el analizador `YYParse`.

Durante el análisis sintáctico de un programa `yapp` se construye una estructura de datos para la posterior manipulación y tratamiento de la gramática. Como ejemplo usaremosla gramática:

```
pl@nereida:~/src/perl/Parse-AutoTree/trunk/scripts$ cat -n int.yp
 1  %right '+'
 2  %left 'a'
 3  %nonassoc 'b'
 4  %%
 5  S:   /* empty rule */                        { print "S -> epsilon\n" }
 6   |   'a' { print "Intermediate\n"; } S 'b'  { print "S -> a S b\n" }
 7   |   '+' S '+' %prec 'a'                    { print "S -> + S + prec a\n" }
 8  ;
 9  %%
10
11  sub _Error {
12          exists $_[0]->YYData->{ERRMSG}
13      and do {
```

[1] La versión a la que se refiere esta sección es la 1.05 (Parse-Yapp-1.05.tar.gz)

```
14          print $_[0]->YYData->{ERRMSG};
15          delete $_[0]->YYData->{ERRMSG};
16          return;
17      };
18      print "Syntax error.\n";
19  }
20
21  sub _Lexer {
22      my($parser)=shift;
23
24          defined($parser->YYData->{INPUT})
25      or  $parser->YYData->{INPUT} = <STDIN>
26      or  return('',undef);
27
28      $parser->YYData->{INPUT}=~s/^[ \t\n]//;
29
30      for ($parser->YYData->{INPUT}) {
31          s/^(.)//s and return($1,$1);
32      }
33  }
34
35  sub Run {
36      my($self)=shift;
37      $self->YYParse( yylex => \&_Lexer, yyerror => \&_Error, yydebug => 0x1F );
38  }
```

Para construir la estructura podemos usar la siguiente subrutina:

```
sub Parse {
  my $grammar = shift;

  my $x = new Parse::Yapp::Parse;
  my $r = $x->Parse($grammar);

  return $r;
}
```

La llamada a `Parse` produce la siguiente estructura de datos:

```
nereida:~/src/perl/Parse-AutoTree/trunk/scripts> grammar.pl int.yp
$VAR1 = {
  'START' => 'S', # Símbolo de arranque
  'SYMS' => { 'S' => 5, 'b' => 3, 'a' => 2, '+' => 1 }, # Símbolo => línea
  'TERM' => {
    'b' => [ 'NONASSOC', 2 ],  # terminal => [ Asociatividad, precedencia ]
    'a' => [ 'LEFT', 1 ],      # terminal => [ Asociatividad, precedencia ]
    '+' => [ 'RIGHT', 0 ] },   # terminal => [ Asociatividad, precedencia ]
    # Si el terminal no tiene precedencia toma la forma  terminal => []
  'NTERM' => { 'S' => [ '1', '3', '4' ] }, # variable => [ indice en RULES de las reglas de S
  'PREC' => { 'a' => 1 }, # Terminales que son usandos en una directiva %prec
  'NULL' => { 'S' => 1 }, # Variables que producen vacío
  'EXPECT' => 0, # Número de conflictos esperados
  'RULES' => [
     [ '$start', [ 'S', '' ], undef, undef ], # Regla de superarranque
     [
```

```
      'S', [], # producción
      undef,   # precedencia explícita de la regla
      [ ' print "S -> epsilon\n" ', 5 ] # [ acción asociada, línea ]
    ],
    [
      '@1-1', [], # Regla intermedia: Variable temporal
      undef,
      [ ' print "Intermediate\n"; ', 6 ]
    ],
    [
      'S', [ 'a', '@1-1', 'S', 'b' ],
      undef,
      [ ' print "S -> a S b\n" ', 6 ]
    ],
    [
      'S', [ '+', 'S', '+' ],
      1, # precedencia explícita de la regla
      [ ' print "S -> + S + prec a\n" ', 7 ]
    ]
  ],
  'HEAD' => undef, # Código de cabecera
  'TAIL' => [ '... código de cola ...', 9 ], # Línea en la que comienza la sección de cola
};
```

Las componentes del hash que aparece arriba se corresponden con diversas variables usadas por `YYParse` durante el análisis. La correspondencia se establece dentro del método `Parse` cuando se hace la asignación:

```
  @$parsed{ 'HEAD', 'TAIL', 'RULES', 'NTERM', 'TERM',
            'NULL', 'PREC', 'SYMS',  'START', 'EXPECT' }
   =       (  $head,  $tail,  $rules,  $nterm,  $term,
              $nullable, $precterm, $syms, $start, $expect);
```

esta asignación es la que crea el hash. Las variables con identificadores en minúsculas son usadas en el analizador. Son visibles en todo el fichero ya que, aunque declaradas léxicas, su declaración se encuentra en la cabecera del analizador:

```
%{
require 5.004;

use Carp;

my($input,$lexlevel,@lineno,$nberr,$prec,$labelno);
my($syms,$head,$tail,$token,$term,$nterm,$rules,$precterm,$start,$nullable);
my($expect);

%}
```

35.22. Práctica: El Análisis de las Acciones

Modifique el código de `YappParse.yp` para que el análisis léxico de las secciones de código (`HEADCODE`, `CODE` y `TAILCODE`) se haga a través de las correspondientes rutinas proveida como parámetros para el análisis por el usuario. La idea es ofrecer un primer paso que facilite la generación de analizadores en diferentes lenguajes `Perl`, `C`, etc.

Estudie el módulo `Text::Balanced`. Basándose en las funciones
`extract_codeblock` y `extract_quotelike`
del módulo `Text::Balanced`, resuelva el problema del reconocimiento de código Perl dentro del analizador léxico de `Parse::Yapp`, evitando forzar al usuario en la escritura de "llaves fantasma". Compare el rendimiento de esta solución con la que provee `Yapp`. Para analizar el *rendimiento* use el módulo `Benchmark`.

¿Cuáles son sus conclusiones? ¿Qué es mejor?

35.23. Práctica: Autoacciones

Extienda `Parse::Yapp` con una directiva `%autoaction CODE` la cuál cambia la acción por defecto. Cuando una regla de producción no tenga una acción asociada, en vez de ejecutarse la acción `yapp` por defecto se ejecutará el código especificado en `CODE`. La directiva podrá aparecer en la parte de cabecera o en el cuerpo del programa `yapp` en una sóla línea aparte. Si aparece en el cuerpo no debe hacerlo en medio de una regla.

Sigue un ejemplo de uso:

```
%{
use Data::Dumper;
my %tree_name = ('=' => 'eq', '+' => 'plus', '-' => 'minus',
                 '*' => 'times', '/' => 'divide');
%}
%right  '='
%left   '-' '+'
%left   '*' '/'
%left   NEG
%autoaction { [$tree_name{$_[2]}, $_[1], $_[3]] }

%%
input:               { undef }
        | input line { undef }
;

line:     '\n'         { undef }
        | exp '\n'     { print Dumper($_[1]); }
        | error '\n'   { $_[0]->YYErrok }
;

exp:        NUM                { $_[1] }
        |   VAR                { $_[1] }
        |   VAR '=' exp | exp '+' exp | exp '-' exp | exp '*' exp | exp '/' exp
        |   '-' exp %prec NEG   { ['neg', $_[2]] }
        |   '(' exp ')'         { $_[2] }
;

%%
```

y un ejemplo de ejecución:

```
$ ./useautoaction1.pl
2+3*4
^D
$VAR1 = [
          'plus',
```

```
          '2',
          [ 'times', '3', '4' ]
        ];
```

Analice la adecuación de los mensajes de error emitidos por el compilador de Perl cuando el código en la auto-acción contiene errores. ¿Son apropiados los números de línea?

Tenga en cuenta los siguientes consejos:

- Cuando compile con `yapp` su módulo use una orden como: `yapp -m Parse::Yapp::Parse Parse.yp`. Este es un caso en que el nombre del fichero de salida (`Parse.pm`) y el nómbre del package `Parse::Yapp::Parse` no coinciden. Este es un caso en que el nombre del fichero de salida (`Parse.pm`) y el nómbre del package `Parse::Yapp::Parse` no coinciden.
- Ahora tiene dos versiones de `Parse::Yapp` en su ordenador. El compilador de Perl va a intentar cargar la instalada. Para ello en su versión del script `yapp` puede incluir una línea que le indique al compilador que debe buscar primero en el lugar en el que se encuentra nuestra librería:

  ```
  BEGIN { unshift @INC, '/home/lhp/Lperl/src/yapp/Parse-Yapp-Auto/lib/' }
  ```

- ¿Qué estrategia a seguir? Una posibilidad es "hacerle creer" al resto de los módulos en `Yapp` que el usuario ha escrito el código de la autoacción en aquellas reglas en las que no existe código explícito asociado. Es posible realizar esta práctica modificando sólo el fichero `YappParse.yp`. El código original `Yapp` usa `undef` para indicar, en el campo adecuado, que una acción no fué definida. La idea es sustituir ese `undef` por el código asociado con la autoacción:

  ```
  my($code)= $autoaction? $autoaction:undef;
  ```

35.24. Práctica: Nuevos Métodos

Continuemos extendiendo `Yapp`.

- Introduzca en el módulo `Driver.pm` de `Yapp` un método `YYLhs` que devuelva el identificador de la variable sintáctica en el lado izquierdo de la regla de producción por la que se está reduciendo.
- Para tener disponible el lado izquierdo deberá modificar la conducta del analizador LALR (subrutina `_Parse`) para que guarde como un atributo el identificador de dicho noterminal.
- ¿Que identificador se devuelve asociado con las acciones intermedias?

Sigue un ejemplo de como programar haciendo uso de esta y la anterior extensión:

```
%right  '='
%left   '-' '+'
%left   '*' '/'
%left   NEG
%autoaction { my $n = $#_;  bless [@_[1..$n]], $_[0]->YYLhs }
%%
input:
        |   input line
;
line:     '\n'       { }
        | exp '\n'   { [ $_[1] ] }
        | error '\n' { }
;
exp:        NUM         |   VAR         |   VAR '=' exp
        |   exp '+' exp |   exp '-' exp |   exp '*' exp |   exp '/' exp
```

```
        |   '-' exp %prec NEG
        |   '(' exp ')' { [ $_[2] ] }
;
%%
...
```

Veamos la ejecución correspondiente al ejemplo anterior:

```
$ ./uselhs2.pl
2+3*4
$VAR1 = bless(
[
  bless( [], 'input' ),
  [
    bless( [
      bless( [ '2' ], 'exp' ),
      '+',
      bless( [
        bless( [ '3' ], 'exp' ), '*', bless( [ '4' ], 'exp' ) ], 'exp' )
    ], 'exp' )
  ]
], 'input' );
```

35.25. Práctica: Generación Automática de Árboles

Partiendo de la práctica anterior, introduzca una directiva `%autotree` que de lugar a la construcción del árbol de análisis concreto. La acción de construcción del árbol:

```
{ my $n = $#_;  bless [@_[1..$n]], $_[0]->YYLhs }
```

se ejecutará para cualquier regla que no tenga una acción explícita asociada.

35.26. Recuperacion de Errores: Visión Detallada

La subrutina `_Parse` contiene el algoritmo de análisis LR genérico. En esta sección nos concentraremos en la forma en la que se ha implantado en `yapp` la recuperación de errores.

```
 1 sub _Parse {
 2   my($self)=shift;
 3   ...
 4   $$errstatus=0; $$nberror=0;
```

La variable `$$errstatus` nos indica la situación con respecto a la recuperación de errores. La variable `$$nberror` contiene el número total de errores.

```
 5   ($$token,$$value)=(undef,undef);
 6   @$stack=( [ 0, undef ] ); $$check='';
 7   while(1) {
 8     my($actions,$act,$stateno);
 9     $stateno=$$stack[-1][0];
10     $actions=$$states[$stateno];
11
12     if  (exists($$actions{ACTIONS})) {
13       defined($$token) or  do { ($$token,$$value)=&$lex($self); };
```

```
14        ...
15      }
16      else { $act=$$actions{DEFAULT}; }
```

Si `$act` no esta definida es que ha ocurrido un error. En tal caso no se entra a estudiar si la acción es de desplazamiento o reducción.

```
17      defined($act) and do {
18        $act > 0 and do { #shift
19            $$errstatus and  do { --$$errstatus; };
20            ...
21            next;
22        };
23        #reduce
24        ....
25        $$check eq 'ERROR' or  do {
26          push(@$stack, [ $$states[$$stack[-1][0]]{GOTOS}{$lhs}, $semval ]);
27          $$check='';
28          next;
29        };
30        $$check='';
31      };
```

Si `$$errstatus` es cero es que estamos ante un nuevo error:

```
32      #Error
33      $$errstatus or   do {
34        $$errstatus = 1;
35        &$error($self);
36        $$errstatus # if 0, then YYErrok has been called
37        or  next;   # so continue parsing
38        ++$$nberror;
39      };
```

Como el error es "nuevo" se llama a la subrutina de tratamiento de errores `&$error`. Obsérvese que no se volverá a llamar a la rutina de manejo de errores hasta que `$$errstatus` vuelva a alcanzar el valor cero. Puesto que `&$error` ha sido escrita por el usuario, es posible que este haya llamado al método `YYErrok`. Si ese es el caso, es que el programador prefiere que el análisis continúe como si la recuperación de errores se hubiera completado.

Ahora se pone `$$errstatus` a 3:

```
47      $$errstatus=3;
```

Cada vez que se logre un desplazamiento con éxito `$$errstatus` será decrementado (línea 19).

A continuación se retiran estados de la pila hasta que se encuentre alguno que pueda transitar ante el terminale especial `error`:

```
48      while(@$stack
49          and (not exists($$states[$$stack[-1][0]]{ACTIONS})
50          or  not exists($$states[$$stack[-1][0]]{ACTIONS}{error})
51          or  $$states[$$stack[-1][0]]{ACTIONS}{error} <= 0)) {
52        pop(@$stack);
53      }
54      @$stack or do {
55        return(undef);
56      };
```

Si la pila quedó vacía se devuelve `undef`. En caso contrario es que el programador escribió alguna regla para la recuperación de errores. En ese caso, se transita al estado correspondiente:

```
57     #shift the error token
58     push(@$stack, [ $$states[$$stack[-1][0]]{ACTIONS}{error}, undef ]);
59   }
60   #never reached
61   croak("Error in driver logic. Please, report it as a BUG");
62 }#_Parse
```

Un poco antes tenemos el siguiente código:

```
41     $$errstatus == 3  #The next token is not valid: discard it
42     and  do {
43       $$token eq ''  # End of input: no hope
44       and  do { return(undef); };
45       $$token=$$value=undef;
46     };
```

Si hemos alcanzado el final de la entrada en una situación de error se abandona devolviendo `undef`.

Ejercicio 35.26.1. *Explique la razón para el comentario de la línea 41. Si* `$$errstatus` *es 3, el último terminal no ha producido un desplazamiento correcto. ¿Porqué?*

A continuación aparecen los códigos de los métodos implicados en la recuperación de errores:

```
sub YYErrok {
  my($self)=shift;

  ${$$self{ERRST}}=0;
    undef;
}
```

El método `YYErrok` cambia el valor referenciado por `$errstatus`. De esta forma se le da al programador `yapp` la oportunidad de anunciar que es muy probable que la fase de recuperación de errores se haya completado.

Los dos siguientes métodos devuelven el número de errores hasta el momento (`YYNberr`) y si nos encontramos o no en fase de recuperación de errores (`YYRecovering`):

```
sub YYNberr {
  my($self)=shift;

  ${$$self{NBERR}};
}

sub YYRecovering {
  my($self)=shift;

  ${$$self{ERRST}} != 0;
}
```

35.27. Descripción Eyapp del Lenguaje SimpleC

En este capítulo usaremos `Parse::Eyapp` para desarrollar un compilador para el siguiente lenguaje, al que denominaremos `Simple C` :

```
program: definition+

definition: funcDef | basictype funcDef | declaration

basictype: INT | CHAR

funcDef: ID '('  params  ')' block

params: ( basictype ID arraySpec)<* ','>

block: '{' declaration* statement* '}'

declaration: basictype declList ';'

declList: (ID arraySpec) <+ ','>

arraySpec: ( '[' INUM ']')*

statement:
    expression ';'
  | ';'
  | BREAK ';'
  | CONTINUE ';'
  | RETURN ';'
  | RETURN expression ';'
  | block
  | ifPrefix statement %prec '+'
  | ifPrefix statement 'ELSE' statement
  | loopPrefix statement

ifPrefix: IF '(' expression ')'

loopPrefix: WHILE '(' expression ')'

expression: binary <+ ','>

Variable: ID ('[' binary ']') *

Primary:
    INUM
  | CHARCONSTANT
  | Variable
  | '(' expression ')'
  | function_call

function_call:  ID  '(' binary <* ',> ')'

Unary: '++' Variable | '--' Variable | Primary

binary:
    Unary
  | binary '+' binary
  | binary '-' binary
```

```
| binary '*' binary
| binary '/' binary
| binary '%' binary
| binary '<' binary
| binary '>' binary
| binary '>=' binary
| binary '<=' binary
| binary '==' binary
| binary '!=' binary
| binary '&' binary
| binary '**' binary
| binary '|' binary
| Variable '=' binary
| Variable '+=' binary
| Variable '-=' binary
| Variable '*=' binary
| Variable '/=' binary
| Variable '%=' binary
| etc. etc.
```

35.28. Diseño de Analizadores con `Parse::Eyapp`

A la hora de construir un analizador sintáctico tenga en cuenta las siguientes normas de buena programación:

1. Comienze trabajando en el cuerpo de la gramática.
2. Olvídese al principio del analizador léxico. Su primer objetivo es tener una gramática limpia de conflictos y que reconozca el lenguaje dado.
3. Sustituya las repeticiones BNF por listas usando los operadores `eyapp` `+`, `*` y sus variantes con separadores. Si una variable describe una lista de `cosas` pongale un adjetivo adecuado como `cosaslist`. Ponga nombres significativos a las variables y terminales. No los llame `d1`, `d2`, etc.
4. Si tiene un elemento opcional en la BNF, por ejemplo, en la regla:

   functiondefinition → [basictype] functionheader functionbody

   use el operador `?`.
5. Cada par de reglas que introduzca vuelva a recompilar con `eyapp` la gramática para ver si se introducido ambiguedad. Cuando estoy editando la gramática suelo escribir a menudo la orden

   `:!eyapp %`

   para recompilar:

```
     15
     16 declaratorlist: declarator +
     17 ;
     18 functiondefinition:
     19    basictype functionheader functionbody
     20  | functionheader functionbody
     21 ;
     22
     23 %%
~
~
~
~
~
~
~
~
~
~
:!eyapp %
```

Esto llama a `eyapp` con el fichero bajo edición. Si hay errores o conflictos (esto es, hemos introducido ambiguedad) los detectarémos enseguida. Procure detectar la aparición de un conflicto lo antes posible. Observe el sangrado del ejemplo. Es el que le recomiendo.

6. Cuando esté en el proceso de construcción de la gramática y aún le queden por rellenar variables sintácticas, declárelas como terminales mediante `%token`. De esta manera evitará las quejas de `eyapp`.

7. **Resolución de Ambiguedades y Conflictos**

   Las operaciones de asignación tienen la prioridad mas baja, seguidas de las lógicas, los test de igualdad, los de comparación, a continuación las aditivas, multiplicativas y por último las operaciones de tipo `unary` y `primary`. Exprese la asociatividad natural y la prioridad especificada usando los mecanismos que `eyapp` provee al efecto: `%left`, `%right`, `%nonassoc` y `%prec`.

8. La gramática de SimpleC es ambigua, ya que para una sentencia como

   $$\text{if } E_1 \text{ then if } E_2 \text{ then } S_1 \text{ else } S_2$$

   existen dos árboles posibles: uno que asocia el "else" con el primer "if" y otra que lo asocia con el segundo. Los dos árboles corresponden a las dos posibles parentizaciones:

   $$\text{if } E_1 \text{ then (if } E_2 \text{ then } S_1 \text{ else } S_2)$$

   Esta es la regla de prioridad usada en la mayor parte de los lenguajes: un "else" casa con el "if" mas cercano. La otra posible parentización es:

   $$\text{if } E_1 \text{ then (if } E_2 \text{ then } S_1) \text{ else } S_2$$

   La conducta por defecto de `eyapp` *es parentizar a derechas.* El generador `eyapp` nos informará del conflicto pero si no se le indica como resolverlo parentizará a derechas. Resuelva este conflicto.

9. *¿Que clase de árbol debe producir el analizador?* La respuesta es que sea lo mas abstracto posible. Debe

   - Contener toda la información necesaria para el manejo eficiente de las fases subsiguientes: Análisis de ámbito, Comprobación de tipos, Optimización independiente de la máquina, etc.
   - Ser uniforme
   - Legible (human-friendly)
   - No contener nodos que no portan información.

   El siguiente ejemplo muestra una versión aceptable de árbol abstracto. Cuando se le proporciona el programa de entrada:

```
nereida:~/doc/casiano/PLBOOK/PLBOOK/code> cat -n prueba5.c
     1  int f(int a)
     2  {
     3    if (a>0)
     4      a = f(a-1);
     5  }
```

   El siguiente árbol ha sido producido por un analizador usando la directiva `%tree` y añadiendo las correspondientes acciones de `bypass`. Puede considerarse un ejemplo aceptable de AST:

```
nereida:~/doc/casiano/PLBOOK/PLBOOK/code> eyapp Simple2 ;\
                                    usesimple2.pl prueba5.c
PROGRAM(
  TYPEDFUNC(
    INT(TERMINAL[INT:1]),
    FUNCTION(
      TERMINAL[f:1],
      PARAMS(
        PARAM(
          INT(TERMINAL[INT:1]),
          TERMINAL[a:1],
          ARRAYSPEC
        )
      ),
      BLOCK(
        DECLARATIONS,
        STATEMENTS(
          IF(
            GT(
              VAR(TERMINAL[a:3]),
              INUM(TERMINAL[0:3])
            ),
            ASSIGN(
              VAR(TERMINAL[a:4]),
              FUNCTIONCALL(
                TERMINAL[f:4],
                ARGLIST(
                  MINUS(
                    VAR(TERMINAL[a:4]),
                    INUM(TERMINAL[1:4])
```

```
                )
              ) # ARGLIST
            ) # FUNCTIONCALL
          ) # ASSIGN
        ) # IF
      ) # STATEMENTS
    ) # BLOCK
  ) # FUNCTION
) # TYPEDFUNC
) # PROGRAM
```

Es deseable darle una estructura uniforme al árbol. Por ejemplo, como consecuencia de que la gramática admite funciones con declaración implícita del tipo retornado cuando este es entero

```
1  definition:
2      funcDef { $_[1]->type("INTFUNC"); $_[1] }
3    | %name TYPEDFUNC
4      basictype funcDef
5    | declaration { $_[1] }
6  ;
```

se producen dos tipos de árboles. Es conveniente convertir las definiciones de función con declaración implícita en el mismo árbol que se obtiene con declaración explícita.

35.29. Práctica: Construcción del AST para el Lenguaje Simple C

Utilice Parse-Eyapp. para construir un árbol de análisis sintáctico abstracto para la gramática descrita en la sección 35.27. Su analizador deberá seguir los consejos explícitados en la sección 35.28.

Analizador Léxico Además del tipo de terminal y su valor el analizador léxico deberá devolver el número de línea. El analizador léxico deberá aceptar comentarios C. En la gramática, el terminal `CHARACTER` se refiere a caracteres entre comillas simples (por ejemplo `'a'`). El terminal `STRING` se refiere a caracteres entre comillas dobles (por ejemplo `"hola"`).

Se aconseja que las palabras reservadas del lenguaje no se traten con expresiones regulares específicas sino que se capturen en el patrón de identificador `[a-z_]\w+`. Se mantiene para ello un hash con las palabras reservadas que es inicializado al comienzo del programa. Cuando el analizador léxico encuentra un identificador mira en primer lugar en dicho hash para ver si es una palabra reservada y, si lo es, devuelve el terminal correspondiente. En caso contrario se trata de un identificador.

35.30. El Generador de Analizadores `byacc`

Existe una version del `yacc` de Berkeley que permite producir código para Perl:

```
> byacc -V
byacc: Berkeley yacc version 1.8.2 (C or perl)
```

Se trata por tanto de un generador de analizadores LALR. Es bastante compatible con `AT&T yacc`. Puedes encontrar una versión en formato `tar.gz` en nuestro servidor http://nereida.deioc.ull.es/~pl/pyacc-pack.tgz o también desde http://www.perl.com/CPAN/src/misc/.

El formato de llamada es:

```
byacc  [ -CPcdlrtv ] [ -b file_prefix ] [ -p symbol_prefix ] filename
```

Las opciones `C` o `c` permiten generar código C. Usando `-P` se genera código Perl. Las opciones `d` y `v` funcionan como es usual en `yacc`. Con `t` se incorpora código para la depuración de la gramática. Si se especifica `l` el código del usuario no es insertado. La opción `r` permite generar ficheros separados para el código y las tablas. No la use con Perl.

Fichero conteniendo la gramática:

```
%{
%}

%token INT EOL
%token LEFT_PAR RIGHT_PAR
%left PLUS MINUS
%left MULT DIV

%%
start: |
start input
;

input: expr EOL { print $1 . "\n"; }
| EOL
;

expr: INT { $p->mydebug("INT -> Expr!"); $$ = $1; }
| expr PLUS expr { $p->mydebug("PLUS -> Expr!"); $$ = $1 + $3; }
| expr MINUS expr { $p->mydebug("MINUS -> Expr!"); $$ = $1 - $3; }
| expr MULT expr { $p->mydebug("MULT -> Expr!"); $$ = $1 * $3; }
| expr DIV expr { $p->mydebug("DIV -> Expr!"); $$ = $1 / $3; }
| LEFT_PAR expr RIGHT_PAR { $p->mydebug("PARENS -> Expr!"); $$ = $2; }
;
%%

sub yyerror {
    my ($msg, $s) = @_;
    my ($package, $filename, $line) = caller;

    die "$msg at <DATA> \n$package\n$filename\n$line\n";
}

sub mydebug {
    my $p = shift;
    my $msg = shift;
    if ($p->{'yydebug'})
    {
        print "$msg\n";
    }
}
```

La compilación con `byacc` del fichero `calc.y` conteniendo la descripción de la gramática produce el módulo Perl conteniendo el analizador.

```
> ls -l
total 12
-rw-r-----    1 pl       casiano        47 Dec 29  2002 Makefile
-rw-r-----    1 pl       casiano       823 Dec 29  2002 calc.y
```

```
-rwxr-x--x    1 pl       casiano       627 Nov 10 15:37 tokenizer.pl
> cat Makefile
MyParser.pm: calc.y
        byacc -d -P MyParser $<

> make
byacc -d -P MyParser calc.y
> ls -ltr
total 28
-rw-r-----    1 pl       casiano       823 Dec 29  2002 calc.y
-rw-r-----    1 pl       casiano        47 Dec 29  2002 Makefile
-rwxr-x--x    1 pl       casiano       627 Nov 10 15:37 tokenizer.pl
-rw-rw----    1 pl       users          95 Nov 16 12:49 y.tab.ph
-rw-rw----    1 pl       users        9790 Nov 16 12:49 MyParser.pm
```

Observe que la opción `-P` es la que permite producir código Perl. Anteriormente se usaba la opción `-p`. Esto se hizo para mantener la compatibilidad con otras versiones de `yacc` en las que la opción `-p` se usa para cambiar el prefijo por defecto (`yy`). Ese es el significado actual de la opción `-p` en `perl-byacc`.

El fichero `y.tab.ph` generado contiene las definiciones de los *tokens*:

```
 cat y.tab.ph
$INT=257;
$EOL=258;
$LEFT_PAR=259;
$RIGHT_PAR=260;
$PLUS=261;
$MINUS=262;
$MULT=263;
$DIV=264;
```

El programa `tokenizer.pl` contiene la llamada al analizador y la definición del analizador léxico:

```
> cat tokenizer.pl
#!/usr/local/bin/perl5.8.0

require 5.004;
use strict;
use Parse::YYLex;
use MyParser;

print STDERR "Version $Parse::ALex::VERSION\n";

my (@tokens) = ((LEFT_PAR => '\(',
                 RIGHT_PAR => '\)',
                 MINUS => '-',
                 PLUS => '\+',
                 MULT => '\*',
                 DIV => '/',
                 INT => '[1-9][0-9]*',
                 EOL => '\n',
                 ERROR => '.*'),
                sub { die "!can\'t analyze: \"$_[1]\"\n!"; });

my $lexer = Parse::YYLex->new(@tokens);
```

```
sub yyerror
{
    die "There was an error:" . join("\n", @_). "\n";
}

my $debug = 0;
my $parser = new MyParser($lexer->getyylex(), \&MyParser::yyerror , $debug);
$lexer->from(\*STDIN);
$parser->yyparse(\*STDIN);
```

El módulo `Parse::YYLex` contiene una versión de `Parse::Lex` que ha sido adaptada para funcionar con `byacc`. Todas las versiones de `yacc` esperan que el analizador léxico devuelva un *token* numérico, mientras que `Parse::Lex` devuelbe un objeto de la clase *token*. Veamos un ejemplo de ejecución:

```
> tokenizer.pl
Version 2.15
yydebug: state 0, reducing by rule 1 (start :)
yydebug: after reduction, shifting from state 0 to state 1
3*(5-9)
yydebug: state 1, reading 257 (INT)
yydebug: state 1, shifting to state 2
yydebug: state 2, reducing by rule 5 (expr : INT)
INT -> Expr!
yydebug: after reduction, shifting from state 1 to state 6
yydebug: state 6, reading 263 (MULT)
yydebug: state 6, shifting to state 11
yydebug: state 11, reading 259 (LEFT_PAR)
yydebug: state 11, shifting to state 4
yydebug: state 4, reading 257 (INT)
yydebug: state 4, shifting to state 2
yydebug: state 2, reducing by rule 5 (expr : INT)
INT -> Expr!
yydebug: after reduction, shifting from state 4 to state 7
yydebug: state 7, reading 262 (MINUS)
yydebug: state 7, shifting to state 10
yydebug: state 10, reading 257 (INT)
yydebug: state 10, shifting to state 2
yydebug: state 2, reducing by rule 5 (expr : INT)
INT -> Expr!
yydebug: after reduction, shifting from state 10 to state 15
yydebug: state 15, reading 260 (RIGHT_PAR)
yydebug: state 15, reducing by rule 7 (expr : expr MINUS expr)
MINUS -> Expr!
yydebug: after reduction, shifting from state 4 to state 7
yydebug: state 7, shifting to state 13
yydebug: state 13, reducing by rule 10 (expr : LEFT_PAR expr RIGHT_PAR)
PARENS -> Expr!
yydebug: after reduction, shifting from state 11 to state 16
yydebug: state 16, reducing by rule 8 (expr : expr MULT expr)
MULT -> Expr!
yydebug: after reduction, shifting from state 1 to state 6
yydebug: state 6, reading 258 (EOL)
yydebug: state 6, shifting to state 8
```

```
yydebug: state 8, reducing by rule 3 (input : expr EOL)
-12
yydebug: after reduction, shifting from state 1 to state 5
yydebug: state 5, reducing by rule 2 (start : start input)
yydebug: after reduction, shifting from state 0 to state 1
yydebug: state 1, reading 0 (end-of-file)
```

Capítulo 36

Análisis Sintáctico Ascendente en C con yacc y bison

36.1. Introducción a yacc

Los fuentes de esta sección pueden encontrarse en https://github.com/crguezl/yacc-examples.

Veamos un ejemplo sencillo de analizador sintáctico escrito en yacc . La gramática se especifica entre las dos líneas de %%. Por defecto, el símbolo de arranque es el primero que aparece, en este caso list. En bison es posible hacer que otro variable lo sea utilizando la declaración %start:

Ejemplo: La Calculadora en yacc

Programa 36.1.1. *Calculadora elemental. Análizador sintáctico.*

```
nereida:~/src/precedencia/hoc1> cat -n hoc1.y
 1  %{
 2  /* File: /home/pl/src/precedencia/hoc1/hoc1.y */
 3  #define YYSTYPE double
 4  #include <stdio.h>
 5  %}
 6  %token NUMBER
 7  %left '+' '-'
 8  %left '*' '/'
 9  %%
10  list
11      :
12      | list '\n'
13      | list expr   { printf("%.8g\n",$2);}
14      ;
15
16  expr
17      : NUMBER { $$ = $1;}
18      | expr '+' expr {$$ = $1 + $3;}
19      | expr '-' expr {$$ = $1 - $3;}
20      | expr '*' expr {$$ = $1 * $3;}
21      | expr '/' expr {$$ = $1 / $3;}
22      ;
23
24  %%
25
26  extern FILE * yyin;
27
```

```
28  main(int argc, char **argv) {
29    if (argc > 1) yyin = fopen(argv[1],"r");
30    yydebug = 1;
31    yyparse();
32  }
33
34  yyerror(char *s) {
35    printf("%s\n",s);
36  }
```

La macro `YYSTYPE` (línea 3) contiene el tipo del valor semántico. Si no se declara se asume `int`.

El fichero `yyin` (líneas 26 y 29) es definido en el fichero conteniendo el analizador léxico `lex.yy.c` . Refiere al fichero de entrada conteniendo el texto a analizar.

Al poner la variable `yydebug` a 1 activamos el modo depuración. Para que la depuración se haga efectiva es necesario definir además la macro `YYDEBUG` .

El analizador sintáctico proveido por `yacc` se llama `yyparse` (línea 31). Por defecto su declaración es `int yyparse ()`

El Analizador Léxico

Programa 36.1.2. *Calculadora elemental. Analizador léxico:*

```
nereida:~/src/precedencia/hoc1> cat -n hoc1.l
 1  %{
 2  #include "y.tab.h"
 3  extern YYSTYPE yylval;
 4  %}
 5  number [0-9]+(\.[0-9]+)?([eE][+-]?[0-9]+)?
 6  %%
 7  {number} { yylval = atof(yytext); return NUMBER; }
 8  .|\n     { return yytext[0];}
 9  %%
10  int yywrap() { return 1; }
```

Compilación

Al compilar el program `yacc` con la opción `-d` se produce además del fichero `y.tab.c` conteniendo el analizador sintáctico un fichero adicional de cabecera `y.tab.h` conteniendo las definiciones de los terminales:

```
nereida:~/src/precedencia/hoc1> yacc -d -v hoc1.y
nereida:~/src/precedencia/hoc1> ls -lt | head -4
total 200
-rw-rw----  1 pl users    2857 2007-01-18 10:26 y.output
-rw-rw----  1 pl users   35936 2007-01-18 10:26 y.tab.c
-rw-rw----  1 pl users    1638 2007-01-18 10:26 y.tab.h
nereida:~/src/precedencia/hoc1> sed -ne '27,48p' y.tab.h | cat -n
 1  #ifndef YYTOKENTYPE
 2  # define YYTOKENTYPE
 3     /* Put the tokens into the symbol table, so that GDB and other debuggers
 4        know about them.  */
 5     enum yytokentype {
 6       NUMBER = 258
 7     };
 8  #endif
 9  /* Tokens.  */
```

```
10  #define NUMBER 258
..  ....................................................
15  #if ! defined (YYSTYPE) && ! defined (YYSTYPE_IS_DECLARED)
16  typedef int YYSTYPE;
17  # define yystype YYSTYPE /* obsolescent; will be withdrawn */
18  # define YYSTYPE_IS_DECLARED 1
19  # define YYSTYPE_IS_TRIVIAL 1
20  #endif
21
22  extern YYSTYPE yylval;
```

La variable `yylval` (líneas 3 y 7 del listado 36.1.2) es declarada por el analizador sintáctico y usada por el analizador léxico. El analizador léxico deja en la misma el valor semántico asociado con el token actual.

Makefile

Para compilar todo el proyecto usaremos el siguiente fichero Makefile:

Programa 36.1.3. *Calculadora elemental. Makefile:*

```
> cat Makefile
hoc1: y.tab.c lex.yy.c
        gcc -DYYDEBUG=1 -g -o hoc1 y.tab.c lex.yy.c
y.tab.c y.tab.h: hoc1.y
        yacc -d -v hoc1.y
lex.yy.c: hoc1.l y.tab.h
        flex -l hoc1.l
clean:
        - rm -f y.tab.c lex.yy.c *.o core hoc1
```

Ejecución

Ejecución 36.1.1. *Para saber que esta haciendo el analizador, insertamos una asignación:* `yydebug = 1;` *justo antes de la llamada a* `yyparse()` *y ejecutamos el programa resultante:*

```
$ hoc1
yydebug: state 0, reducing by rule 1 (list :)
yydebug: after reduction, shifting from state 0 to state 1
2.5+3.5+1
```

Introducimos la expresión `2.5+3.5+1`*. Antes que incluso ocurra la entrada, el algoritmo LR reduce por la regla* $List \rightarrow \epsilon$.

```
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 2
yydebug: state 2, reducing by rule 4 (expr : NUMBER)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 43 ('+')
yydebug: state 4, shifting to state 5
yydebug: state 5, reading 257 (NUMBER)
yydebug: state 5, shifting to state 2
yydebug: state 2, reducing by rule 4 (expr : NUMBER)
yydebug: after reduction, shifting from state 5 to state 6
yydebug: state 6, reducing by rule 5 (expr : expr '+' expr)
```

Observe como la declaración de la asociatividad a izquierdas `%left '+'` *se traduce en la reducción por la regla 5.*

```
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 43 ('+')
yydebug: state 4, shifting to state 5
yydebug: state 5, reading 257 (NUMBER)
yydebug: state 5, shifting to state 2
yydebug: state 2, reducing by rule 4 (expr : NUMBER)
yydebug: after reduction, shifting from state 5 to state 6
yydebug: state 6, reducing by rule 5 (expr : expr '+' expr)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 10 ('\n')
yydebug: state 4, reducing by rule 3 (list : list expr)
7
```

La reducción por la regla list → list expr produce la ejecución del `printf("%.8g\n",$2);` *asociado con la regla y la salida del valor* `7` *que constituye el atributo de expr.*

```
yydebug: after reduction, shifting from state 0 to state 1
yydebug: state 1, shifting to state 3
yydebug: state 3, reducing by rule 2 (list : list '\n')
yydebug: after reduction, shifting from state 0 to state 1
yydebug: state 1, reading 0 (end-of-file)
$
```

En Unix la combinación de teclas CTRL-D nos permite generar el final de fichero.

36.2. Precedencia y Asociatividad

En caso de que no existan indicaciones específicas *yacc* resuelve los conflictos que aparecen en la construcción de la tabla utilizando las siguientes reglas:

1. Un conflicto *reduce-reduce* se resuelve eligiendo la producción que se listó primero en la especificación de la gramática.

2. Un conflicto *shift-reduce* se resuelve siempre en favor del *shift*

La precedencia se utiliza para modificar estos criterios. Para ello se define:

1. La precedencia de los *tokens* es determinada según el orden de declaración. La declaración de *tokens* mediante la palabra reservada `token` no modifica la precedencia. Si lo hacen las declaraciones realizadas usando las palabras `left`, `right` y `nonassoc`. Los *tokens* declarados en la misma línea tienen igual precedencia. La precedencia es mayor cuanto mas abajo en el texto. Así, en el ejemplo que sigue, el *token* `*` tiene mayor precedencia que `+` pero la misma que `/`.

2. La precedencia de una regla $A \rightarrow \alpha$ se define como la del terminal mas a la derecha que aparece en α. En el ejemplo, la producción

   ```
   expr : expr '+' expr
   ```

   tiene la precedencia del *token* `+`.

3. Para decidir en un conflicto *shift-reduce* se comparan la precedencia de la regla con la del terminal que va a ser desplazado. Si la de la regla es mayor se reduce si la del *token* es mayor, se desplaza.

4. Si en un conflicto *shift-reduce* ambos la regla y el terminal que va a ser desplazado tiene la misma precedencia *yacc* considera la asociatividad, si es asociativa a izquierdas, reduce y si es asociativa a derechas desplaza. Si no es asociativa, genera un mensaje de error.

Obsérvese que, en esta situación, la asociatividad de la regla y la del *token* han de ser por fuerza, las mismas. Ello es así, porque en *yacc* los *tokens* con la misma precedencia se declaran en la misma línea.

Por tanto es imposible declarar dos tokens *con diferente asociatividad y la misma precedencia.*

5. Es posible modificar la precedencia "natural" de una regla, calificándola con un *token* específico. para ello se escribe a la derecha de la regla `prec token`, donde `token` es un *token* con la precedencia que deseamos. Vea el uso del *token* `dummy` en el siguiente ejercicio.

Programa 36.2.1. *Este programa muestra el manejo de las reglas de precedencia.*

```
%{
#define YYSTYPE double
#include <stdio.h>
%}
%token NUMBER
%left '@'
%right '&'  dummy
%%
list
    :
    | list '\n'
    | list e
    ;

e : NUMBER
  | e '&' e
  | e '@' e %prec dummy
    ;

%%
extern FILE * yyin;

main(int argc, char **argv) {
  if (argc > 1) yyin = fopen(argv[1],"r");
  yydebug = 1;
  yyparse();
}

yyerror(char *s) {
  printf("%s\n",s);
}
```

Ejercicio 36.2.1. *Dado el programa* `yacc` *36.2.1 Responda a las siguientes cuestiones:*

1. *Construya las tablas SLR de acciones y* gotos.
2. *Determine el árbol construido para las frases:* `4@3@5`, `4&3&5`, `4@3&5`, `4&3@5`.
3. *¿Cuál es la asociatividad final de la regla* `e : e '@' e`*?*

Listado 36.2.1. *Fichero* y.output*:*

```
   0  $accept : list $end
```

```
1  list :
2       | list '\n'
3       | list e

4  e : NUMBER
5    | e '&' e
6    | e '@' e
```

```
state 0
$accept : . list $end  (0)
list : .  (1)

.  reduce 1

list  goto 1


state 1
$accept : list . $end  (0)
list : list . '\n'  (2)
list : list . e  (3)

$end  accept
NUMBER  shift 2
'\n'  shift 3
.  error

e  goto 4


state 2
e : NUMBER .  (4)

.  reduce 4


state 3
list : list '\n' .  (2)

.  reduce 2


state 4
list : list e .  (3)
e : e . '&' e  (5)
e : e . '@' e  (6)

'@'  shift 5
'&'  shift 6
$end  reduce 3
NUMBER  reduce 3
'\n'  reduce 3
```

```
state 5
e : e '@' . e  (6)

NUMBER  shift 2
.  error

e  goto 7


state 6
e : e '&' . e  (5)

NUMBER  shift 2
.  error

e  goto 8


state 7
e : e . '&' e  (5)
e : e . '@' e  (6)
e : e '@' e .  (6)

'&'  shift 6
$end  reduce 6
NUMBER  reduce 6
'@'  reduce 6
'\n'  reduce 6

state 8
e : e . '&' e  (5)
e : e '&' e .  (5)
e : e . '@' e  (6)

'&'  shift 6
$end  reduce 5
NUMBER  reduce 5
'@'  reduce 5
'\n'  reduce 5
```

```
7 terminals, 3 nonterminals
7 grammar rules, 9 states
```

Ejemplo 36.2.1. *Contrasta tu respuesta con la traza seguida por el programa anterior ante la entrada* `1@2&3`*, al establecer la variable* `yydebug = 1` *y definir la macro* `YYDEBUG`*:*

Ejecución 36.2.1. `$ hocprec`

```
yydebug: state 0, reducing by rule 1 (list :)
yydebug: after reduction, shifting from state 0 to state 1
1@2&3
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 2
yydebug: state 2, reducing by rule 4 (e : NUMBER)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 64 ('@')
yydebug: state 4, shifting to state 5
yydebug: state 5, reading 257 (NUMBER)
yydebug: state 5, shifting to state 2
yydebug: state 2, reducing by rule 4 (e : NUMBER)
yydebug: after reduction, shifting from state 5 to state 7
yydebug: state 7, reading 38 ('&')
yydebug: state 7, shifting to state 6
```

¿Por que se desplaza? ¿No va eso en contra de la declaración `%left '@'`*?. ¿O quizá es porque la precedencia de* `@` *es menor que la de* `&`*? La respuesta es que la precedencia asignada por la declaración*

```
e : e '@' e %prec dummy
```

cambio la asociatividad de la regla. Ahora la regla se "enfrenta" a un token*,* `&` *con su misma precedencia. Al pasar a ser asociativa a derechas (debido a que* `dummy` *lo es), se debe desplazar y no reducir.*

Ejemplo 36.2.2. *Otra ejecución, esta vez con entrada* `1&2@3`*. Compara tus predicciones con los resultados.*

Ejecución 36.2.2. `$ hocprec`

```
yydebug: state 0, reducing by rule 1 (list :)
yydebug: after reduction, shifting from state 0 to state 1
1&2@3
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 2
yydebug: state 2, reducing by rule 4 (e : NUMBER)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 38 ('&')
yydebug: state 4, shifting to state 6
yydebug: state 6, reading 257 (NUMBER)
yydebug: state 6, shifting to state 2
yydebug: state 2, reducing by rule 4 (e : NUMBER)
yydebug: after reduction, shifting from state 6 to state 8
yydebug: state 8, reading 64 ('@')
yydebug: state 8, reducing by rule 5 (e : e '&' e)
```

En este caso se comparan la producción :

$$\texttt{e} \rightarrow \texttt{e \& e}$$

con el token `@`*. La regla tiene mayor precedencia que el* token*,dado que la precedencia de la regla es la de* `&`*.*

36.3. Uso de union y type

En general, los atributos asociados con las diferentes variables sintácticas y terminales tendrán tipos de datos distintos. Para ello, `yacc` provee la declaración `%union`.

La declaración `%union` especifica la colección de posibles tipos de datos de `yylval` y de los atributos `$1`, `$2`, ...

He aqui un ejemplo:

```
%union {
  double val;
  int index;
}
```

Esto dice que los dos tipos de alternativas son `double` y `int` . Se les han dado los nombres `val` y `index`.

```
%token <val>   NUMBER
%token <index> VAR
%type  <val>   expr
%right '='
%left '+' '-'
%left '*' '/'
```

Estos nombres `<val>` e `<index>` se utilizan en las declaraciones de `%token` y `%type` para definir el tipo del correspondiente atributo asociado.

Dentro de las acciones, se puede especificar el tipo de un símbolo insertando `<tipo>` despues del `$` que referencia al atributo. En el ejemplo anterior podríamos escribir `$<val>1` para indicar que manipulamos el atributo del primer símbolo de la parte derecha de la regla como si fuera un `double`.

La información que provee la declaración `%union` es utilizada por `yacc` para realizar la sustitución de las referencias textuales/formales (`$$`, `$1`, ... `$`$|\alpha|$) a los atributos de los símbolos que conforman la regla (que pueden verse como parámetros formales de las acciones) por las referencias a las zonas de memoria en las que se guardan (que están asociadas con los correspondientes estados de la pila) cuando tiene lugar la reducción en el algoritmo de análisis LR:

```
case "reduce A → α" :
  execute("reduce A → α", top(|α|-1), ··· , top(0));
  pop(|α|);
  push(goto[top(0)][A];
  break;
```

Asi, `yacc` es capaz de insertar el código de ahormado de tipos correcto. Ello puede hacerse porque se conocen los tipos asociados con los símbolos en la parte derecha de una regla, ya que han sido proveídos en las declaraciones `%union`, `%token` y `%type`.

36.4. Acciones en medio de una regla

A veces necesitamos insertar una acción en medio de una regla. Una acción en medio de una regla puede hacer referencia a los atributos de los símbolos que la preceden (usando `$n`), pero no a los que le siguen.

Cuando se inserta una acción $\{action_1\}$ para su ejecución en medio de una regla $A \to \alpha\beta$:

$$A \to \alpha \{action_1\} \beta \{action_2\}$$

`yacc` crea una variable sintáctica temporal T e introduce una nueva regla:

1. $A \to \alpha T \beta \{action_2\}$

2. $T \rightarrow \epsilon \{action_1\}$

Las acciones en mitad de una regla cuentan como un símbolo mas en la parte derecha de la regla. Asi pues, en una acción posterior en la regla, se deberán referenciar los atributos de los símbolos, teniendo en cuenta este hecho.

Las acciones en mitad de la regla pueden tener un atributo. La acción en cuestión puede hacer referencia a ese atributo mediante `$$`, y las acciones posteriores en la regla se referirán a él como `$n`, siendo `n` su número de orden en la parte derecha. Dado que no existe un símbolo explícito que identifique a la acción, no hay manera de que el programador declare su tipo. Sin embargo, es posible utilizar la construcción `$<valtipo>#` para especificar la forma en la que queremos manipular su atributo.

Na hay forma de darle, en una acción a media regla, un valor al atributo asociado con la variable en la izquierda de la regla de producción (ya que `$$` se refiere al atributo de la variable temporal utilizada para introducir la acción a media regla).

Programa 36.4.1. *El siguiente programa ilustra el uso de* `%union` *y de las acciones en medio de una regla.*

```
%{
#include <string.h>
char buffer[256];
#define YYDEBUG 1
%}
%union {
  char tA;
  char *tx;
}
%token <tA> A
%type <tx> x
%%
s : x { *$1 = '\0'; printf("%s\n",buffer); } '\n' s
  |
  ;

x : A { $$ = buffer + sprintf(buffer,"%c",$1); }
  | A { $<tx>$ = strdup("**"); } x
    { $$ = $3 + sprintf($3,"%s%c",$<tx>2,$1); free($2); }
  ;

%%

main() {
  yydebug=1;
  yyparse();
}

yyerror(char *s) {
  printf("%s\n",s);
}
```

Programa 36.4.2. *El analizador léxico utilizado es el siguiente:*

```
%{
#include "y.tab.h"
%}
%%
```

```
[\t ]+
[a-zA-Z0-9]   { yylval.tA = yytext[0]; return A; }
(.|\n)        { return yytext[0]; }
%%
yywrap() { return 1; }
```

Ejemplo 36.4.1. *Considere el programa* `yacc` *36.4.1. ¿Cuál es la salida para la entrada ABC?*

La gramática inicial se ve aumentada con dos nuevas variables sintácticas temporales y dos reglas $t_1 \to \epsilon$ *y* $t_2 \to \epsilon$*. Además las reglas correspondientes pasan a ser:* $s \to xt_1 s$ *y* $x \to At_2 x$*. El análisis de la entrada ABC nos produce el siguiente árbol anotado:*

Ejecución 36.4.1. *Observe la salida de la ejecución del programa 36.4.1. La variable "temporal" creada por* `yacc` *para la acción en medio de la regla*

```
s →  x { *$1 = '\0'; printf("%s\n",buffer); } '\n' s
```

se denota por `$$1`*. La asociada con la acción en medio de la regla*

```
x  →  A { $<tx>$ = strdup("**"); } x
```

se denota `$$2`*.*

```
$ yacc -d -v media4.y ; flex -l medial.l ; gcc -g y.tab.c lex.yy.c
$ a.out
ABC
yydebug: state 0, reading 257 (A)
yydebug: state 0, shifting to state 1
yydebug: state 1, reading 257 (A)
yydebug: state 1, reducing by rule 5 ($$2 :)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, shifting to state 1
yydebug: state 1, reading 257 (A)
yydebug: state 1, reducing by rule 5 ($$2 :)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, shifting to state 1
yydebug: state 1, reading 10 ('\n')
yydebug: state 1, reducing by rule 4 (x : A)
yydebug: after reduction, shifting from state 4 to state 6
yydebug: state 6, reducing by rule 6 (x : A $$2 x)
yydebug: after reduction, shifting from state 4 to state 6
yydebug: state 6, reducing by rule 6 (x : A $$2 x)
yydebug: after reduction, shifting from state 0 to state 3
yydebug: state 3, reducing by rule 1 ($$1 :)
C**B**A
yydebug: after reduction, shifting from state 3 to state 5
yydebug: state 5, shifting to state 7
yydebug: state 7, reading 0 (end-of-file)
yydebug: state 7, reducing by rule 3 (s :)
yydebug: after reduction, shifting from state 7 to state 8
yydebug: state 8, reducing by rule 2 (s : x $$1 '\n' s)
yydebug: after reduction, shifting from state 0 to state 2
```

Ejemplo 36.4.2. *¿Que ocurre si en el programa 36.4.1 adelantamos la acción intermedia en la regla*

```
x →  A { $<tx>$ = strdup("**"); } x
```

y la reescribimos

```
x →   { $<tx>$ = strdup("**"); } A x?
```

Ejecución 36.4.2. *En tal caso obtendremos:*

```
$ yacc -d -v media3.y
yacc: 1 rule never reduced
yacc: 3 shift/reduce conflicts.
```

¿Cuáles son esos 3 conflictos?

Listado 36.4.1. *El fichero* `y.output` *comienza enumerando las reglas de la gramática extendida:*

```
 1    0  $accept : s $end
 2
 3    1  $$1 :
 4
 5    2  s : x $$1 '\n' s
 6    3    |
 7
 8    4  x : A
 9
10    5  $$2 :
11
12    6  x : $$2 A x
13 ^L
```

A continuación nos informa de un conflicto en el estado 0. Ante el token `A` *no se sabe si se debe desplazar al estado 1 o reducir por la regla 5:* `$$2 : .`

```
14 0: shift/reduce conflict (shift 1, reduce 5) on A
15 state 0
16         $accept : . s $end  (0)
17         s : .  (3)
18         $$2 : .  (5)
19
20         A  shift 1
21         $end  reduce 3
22
23         s  goto 2
24         x  goto 3
25         $$2  goto 4
```

Observe que, efectivamente, `$$2 : .` *esta en la clausura del estado de arranque del NFA (*`$accept : . s $end`*) Esto es asi, ya que al estar el marcador junto a* `x`*, estará el item* `s : . x $$1 '\n' s` *y de aqui que también este* `x : . $$2 A x`*.*

Además el token `A` *está en el conjunto* $FOLLOW$(`$$2`) *(Basta con mirar la regla 6 para confirmarlo). Por razones análogas también está en la clausura del estado de arranque el item* `x : . A` *que es el que motiva el desplazamiento al estado 1.*

La dificultad para `yacc` *se resolvería si dispusiera de información acerca de cual es el* token *que viene después de la* `A` *que causa el conflicto.*

Ejercicio 36.4.1. *¿Que acción debe tomarse en el conflicto del ejemplo 36.4.2 si el* token *que viene después de* `A` *es* `\n`*? ¿Y si el* token *es* `A`*? ¿Se debe reducir o desplazar?*

36.5. Recuperación de Errores

Las entradas de un traductor pueden contener errores. El lenguaje `yacc` proporciona un *token* especial, `error`, que puede ser utilizado en el programa fuente para extender el traductor con producciones de error que lo doten de cierta capacidad para recuperase de una entrada errónea y poder continuar analizando el resto de la entrada.

Ejecución 36.5.1. *Consideremos lo que ocurre al ejecutar el programa* `yacc` *36.1.1 con una entrada errónea:*

```
$ hoc1
yydebug: state 0, reducing by rule 1 (list :)
yydebug: after reduction, shifting from state 0 to state 1
3--2
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 2
yydebug: state 2, reducing by rule 4 (expr : NUMBER)
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 45 (illegal-symbol)
syntax error
yydebug: error recovery discarding state 4
yydebug: error recovery discarding state 1
yydebug: error recovery discarding state 0
```

Después de detectar el mensaje `yacc` emite el mensaje `syntax error` y comienza a sacar estados de la pìla hasta que esta se vacía.

Programa 36.5.1. *La conducta anterior puede modificarse si se introducen "reglas de recuperación de errores" como en la siguiente modificación del programa 36.1.1:*

```
%{
#define YYSTYPE double
#define YYDEBUG 1
#include <stdio.h>
%}
%token NUMBER
%left '-'
%%
list
    :
    | list '\n'
    | list error '\n'  { yyerrok; }
    | list expr   { printf("%.8g\n",$2);}
    ;

expr
    : NUMBER { $$ = $1;}
    | expr '-' expr {$$ = $1 - $3;}
    ;

%%
```

La regla `list` → `list error '\n'` es una producción de error. La idea general de uso es que, a traves de la misma, el programador le indica a `yacc` que, cuando se produce un error dentro de una expresión, descarte todos los *tokens* hasta llegar al retorno del carro y prosiga con el análisis. Además, mediante la llamada a la macro `yyerrok` el programador anuncia que, si se alcanza este punto, la

recuperación puede considerarse "completa" y que `yacc` puede emitir a partir de ese moemnto mensajes de error con la seguridad de que no son consecuencia de un comportamiento inestable provocado por el primer error.

Algoritmo 36.5.1. *El esquema general del algoritmo de recuperación de errores usado por la versión actual de* `yacc` *es el siguiente:*

1. *Cuando se encuentra ante una acción de error, el analizador genera un* token `error`*.*
2. *A continuación pasa a retirar estados de la pila hasta que descubre un estado capaz de transitar ante el* token `error`*.*
3. *En este punto transita al estado correspondiente a desplazar el* token `error`*.*
4. *Entonces lee* tokens *y los descarta hasta encontrar uno que sea aceptable.*
5. *Sólo se envían nuevos mensajes de error una vez asimilados (desplazados) tres símbolos terminales. De este modos se intenta evitar la aparición masiva de mensajes de error.*

Algoritmo 36.5.2. *El cuerpo principal del analizador LR permanece sin demasiados cambios:*

```
goodtoken = 3; b = yylex();
for( ; ; ;) {
  s = top(); a = b;
  switch (action[s][a])) {
    case "shift t" : push(t); b = yylex(); goodtoken++; break;
    case "reduce A -> alpha" :
            pop(strlen(alpha));
            push(goto[top()][A];
          break;
    case "accept" : return (1);
    default : if (errorrecovery("syntax error")) return (ERROR);
  }
}
```

Algoritmo 36.5.3. *El siguiente seudocódigo es una reescritura mas detallada del algoritmo 36.5.1. Asumimos que las funciones* `pop()` *y* `popstate()` *comprueban que hay suficientes elementos en la pila para retirar. En caso contrario se emitirá un mensaje de error y se terminará el análisis.*

```
errorrecovery(char * s) {
 if (goodtoken > 2) {
   yyerror(s); goodtoken = 0;
 }
 while (action[s][error] != shift)
   popstate(s);
 push(goto[s][error]);
 s = top();
 while (action[s][a] == reduce A -> alpha) {
   pop(strlen(|alpha|));
   push(goto[top()][A];
   s = top();
 }
 switch (action[s][a])) {
   case "shift t" :
     push(t);
     b = yylex();
```

```
      goodtoken++;
      RETURN RECOVERING;
    case "accept" : return (ERROR);
    default :
      do b = yylex();
      while ((b != EOF)&&(action[s][b] == error);
      if (b == EOF)
        return (ERROR);
      else
        RETURN RECOVERING
}
```

Parecen existir diferencias en la forma en la que *bison* y *yacc* se recuperan de los errores.

Ejecución 36.5.2. *Ejecutemos el programa 36.5.1 con* `yacc`.

```
$ yacc -d hoc1.y; flex -l hoc1.l; gcc y.tab.c lex.yy.c; a.out
yydebug: state 0, reducing by rule 1 (list :)
yydebug: after reduction, shifting from state 0 to state 1
2--3-1
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 3
yydebug: state 3, reducing by rule 5 (expr : NUMBER)
yydebug: after reduction, shifting from state 1 to state 5
yydebug: state 5, reading 45 ('-')
yydebug: state 5, shifting to state 7
yydebug: state 7, reading 45 ('-')
syntax error
```

Puesto que es el primer error, se cumple que `(goodtoken > 2)`*, emitiéndose el mensaje de error. Ahora comienzan a ejecutarse las líneas:*

```
while (!(action[s][error] != shift)) popstate(s);
```

que descartan los estados, hasta encontrar el estado que contiene el item
`list` $\rightarrow$ `list` $\uparrow$ `error '\n'`

```
yydebug: error recovery discarding state 7
yydebug: error recovery discarding state 5
yydebug: state 1, error recovery shifting to state 2
```

Una vez en ese estado, transitamos con el token `error`*,*

```
yydebug: state 2, error recovery discards token 45 ('-')
yydebug: state 2, reading 257 (NUMBER)
yydebug: state 2, error recovery discards token 257 (NUMBER)
yydebug: state 2, reading 45 ('-')
yydebug: state 2, error recovery discards token 45 ('-')
yydebug: state 2, reading 257 (NUMBER)
yydebug: state 2, error recovery discards token 257 (NUMBER)
```

Se ha procedido a descartar tokens *hasta encontrar el retorno de carro, ejecutando las líneas:*

```
b = yylex(); while ((b != EOF)&&(action[s][b] == error);
```

```
yydebug: state 2, reading 10 ('\n')
yydebug: state 2, shifting to state 6
yydebug: state 6, reducing by rule 3 (list : list error '\n')
yydebug: after reduction, shifting from state 0 to state 1
```

Al reducir por la regla de error, se ejecuta `yyerrok` *y* `yacc` *reestablece el valor de* `goodtoken`*. Si se producen nuevos errores serán señalados.*

36.6. Recuperación de Errores en Listas

Aunque no existe un método exacto para decidir como ubicar las reglas de recuperación de errores, en general, los símbolos de error deben ubicarse intentado satisfacer las siguientes reglas:

- Tan cerca como sea posible del símbolo de arranque.
- Tan cerca como sea posible de los símbolos terminales.
- Sin introducir nuevos conflictos.

Esquema 36.6.1. *En el caso particular de las listas, se recomienda seguir el siguiente esquema:*

Construct	*EBNF*	*yacc input*
optional sequence	$x : \{y\}$	`x : /* null */` $xyyyerrok;\|$ $xerror\|$
sequence	$x : y\{y\}$	`x : y` $xyyyerrok;\|$ $error\|$ $xerror\|$
list	$x : y\{Ty\}$	`x : y` $xTyyyerrok;\|$ $error\|$ $xerror\|$ $xerroryyyerrok;\|$ $xTerror\|$

Programa 36.6.1. *Para comprobar el funcionamiento y la validez de la metodología esquematizada en el esquema 36.6.1, consideremos los contenidos del fichero* `error.y`*. En el se muestra el tercer caso* $x{:}y\{Ty\}$ *con* $x =$ `list`*,* $T =$ `,` *e* $y =$ `NUMBER`*:*

```
%{
#include <stdio.h>
void put(double x);
void err(int code);
%}

%union {
  double val;
}
%token <val>NUMBER
%%
command
  :
  | command list '\n' { yyerrok; }
  ;

list
    : NUMBER            { put($1); }
    | list ',' NUMBER   { put($3); yyerrok; }
    | error             { err(1); }
    | list error        { err(2); }
    | list error NUMBER { err(3); put($3); yyerrok; }
    | list ',' error    { err(4); }
    ;
```

```
%%
void put(double x) {
  printf("%2.1lf\n",x);
}

void err(int code) {
  printf("err %d\n",code);
}

main() {
  yydebug = 1;
  yyparse();
}

yyerror(char *s) {
  printf("%s\n",s);
}
```

Listado 36.6.1. *La compilación con* `yacc` *da lugar a una tabla ligeramente diferente de la producida por* `bison`*. El fichero* `y.output` *contiene la tabla:*

```
   0  $accept : command $end

   1  command :
   2          | command list '\n'

   3  list : NUMBER
   4       | list ',' NUMBER
   5       | error
   6       | list error
   7       | list error NUMBER
   8       | list ',' error

state 0
$accept : . command $end  (0)
command : .  (1)

.  reduce 1

command  goto 1


state 1
$accept : command . $end  (0)
command : command . list '\n'  (2)

$end  accept
error  shift 2
NUMBER  shift 3
.  error

list  goto 4
```

```
state 2
list : error .  (5)

.  reduce 5


state 3
list : NUMBER .  (3)

.  reduce 3


state 4
command : command list . '\n'  (2)
list : list . ',' NUMBER  (4)
list : list . error  (6)
list : list . error NUMBER  (7)
list : list . ',' error  (8)

error  shift 5
'\n'  shift 6
','  shift 7
.  error


state 5
list : list error .  (6)
list : list error . NUMBER  (7)

NUMBER  shift 8
error  reduce 6
'\n'  reduce 6
','  reduce 6


state 6
command : command list '\n' .  (2)

.  reduce 2


state 7
list : list ',' . NUMBER  (4)
list : list ',' . error  (8)

error  shift 9
NUMBER  shift 10
.  error


state 8
list : list error NUMBER .  (7)
```

```
.  reduce 7


state 9
list : list ',' error .  (8)

.  reduce 8


state 10
list : list ',' NUMBER .  (4)

.  reduce 4


5 terminals, 3 nonterminals
9 grammar rules, 11 states
```

Ejecución 36.6.1. *La ejecución del programa generado por* `yacc` *es como sigue:*

```
> error
yydebug: state 0, reducing by rule 1 (command :)
yydebug: after reduction, shifting from state 0 to state 1
10 20
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 3
yydebug: state 3, reducing by rule 3 (list : NUMBER)
10.0
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 257 (NUMBER)
syntax error
yydebug: state 4, error recovery shifting to state 5
yydebug: state 5, shifting to state 8
yydebug: state 8, reducing by rule 7 (list : list error NUMBER)
err 3
20.0
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 10 ('\n')
yydebug: state 4, shifting to state 6
yydebug: state 6, reducing by rule 2 (command : command list '\n')
yydebug: after reduction, shifting from state 0 to state 1
10;20 30
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 3
yydebug: state 3, reducing by rule 3 (list : NUMBER)
10.0
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 59 (illegal-symbol)
syntax error
yydebug: state 4, error recovery shifting to state 5
yydebug: state 5, error recovery discards token 59 (illegal-symbol)
yydebug: state 5, reading 257 (NUMBER)
yydebug: state 5, shifting to state 8
yydebug: state 8, reducing by rule 7 (list : list error NUMBER)
```

```
err 3
20.0
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 257 (NUMBER)
syntax error
yydebug: state 4, error recovery shifting to state 5
yydebug: state 5, shifting to state 8
yydebug: state 8, reducing by rule 7 (list : list error NUMBER)
err 3
30.0
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 10 ('\n')
yydebug: state 4, shifting to state 6
yydebug: state 6, reducing by rule 2 (command : command list '\n')
yydebug: after reduction, shifting from state 0 to state 1
3,
yydebug: state 1, reading 257 (NUMBER)
yydebug: state 1, shifting to state 3
yydebug: state 3, reducing by rule 3 (list : NUMBER)
3.0
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, reading 44 (',')
yydebug: state 4, shifting to state 7
yydebug: state 7, reading 10 ('\n')
syntax error
yydebug: state 7, error recovery shifting to state 9
yydebug: state 9, reducing by rule 8 (list : list ',' error)
err 4
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, shifting to state 6
yydebug: state 6, reducing by rule 2 (command : command list '\n')
yydebug: after reduction, shifting from state 0 to state 1
#
yydebug: state 1, reading 35 (illegal-symbol)
syntax error
yydebug: state 1, error recovery shifting to state 2
yydebug: state 2, reducing by rule 5 (list : error)
err 1
yydebug: after reduction, shifting from state 1 to state 4
yydebug: state 4, error recovery discards token 35 (illegal-symbol)
yydebug: state 4, reading 10 ('\n')
yydebug: state 4, shifting to state 6
yydebug: state 6, reducing by rule 2 (command : command list '\n')
yydebug: after reduction, shifting from state 0 to state 1
yydebug: state 1, reading 0 (end-of-file)
```

Capítulo 37

Análisis Semántico

Capítulo 38

Generación de Código

Capítulo 39

Optimización de Código

Parte III

TERCERA PARTE: BITÁCORA DEL CURSO

Capítulo 40

2012

40.1. 01

40.1.1. 31/01/13

- Presentación de la Asignatura
- Ejercicio: Darse de alta en la comunidad de google plus PL Grado ULL 12/13¿
- Ejercicio: Indicar un mail en gmail para compartir recursos
- Ejercicio: Indicar página en GitHub

40.2. 04

40.2.1. Proyecto: Diseña e Implementa un Lenguaje de Dominio Específico

Se trata de realizar un proyecto relacionado con el procesamiento de lenguajes. El objetivo puede ser:

1. Diseñar un lenguaje de dominio específico para simplificar cualquier tarea en la que estés interesado:
    - Para escribir exámenes,
    - Para dibujar árboles,
    - Para calcular fechas,
    - Para generar emails
    - Para escribir música
    - Para escribir autómatas finitos
    - etc.
2. Estudiar un traductor existente en profundidad como:
    - ECMAscript 5.1: Creating a JavaScript Parser Una implementación de ECMAScript 5.1 usando Jison disponible en GitHub en https://github.com/cjihrig/jsparser. Puede probarse en: http://www.cjihrig.com/development/jsparser/
    - Roy
    - CoffeScript
    - Jison
    - Javascript 1.4
    - etc.

3. También puedes proponer tu propio tema relacionado al profesor

Se recomienda para ello organizar equipos de no menos de dos y no mas de cuatro.
Las presentaciones de los proyectos tendrán lugar el último día de clase Martes 21 de Mayo.

Índice general

Índice de figuras

Índice de cuadros

Índice alfabético

Bibliografía

[1] A. Osmani. *Learning JavaScript Design Patterns*. Oreilly and Associate Series. O'Reilly Media, Incorporated, 2012.

[2] S. Powers. *Learning Node*. Oreilly and Associate Series. O'Reilly Media, Incorporated, 2012.

[3] M. Haverbeke. *Eloquent JavaScript: A Modern Introduction to Programming*. No Starch Press Series. No Starch Press, 2011.

[4] S. Holzner. *Sams Teach Yourself Xml in 21 Days*. Sams Teach Yourself. Sams, 2003.

[5] Mark Pilgrim. *Dive into HTML5*. http://diveinto.html5doctor.com/index.html, 2013.

[6] Dale Dougherty. *Sed and AWK*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 1991.

[7] Jerry Peek, Tim O'Reilly, Dale Dougherty, Mike Loukides, Chris Torek, Bruce Barnett, Jonathan Kamens, Gene Spafford, and Simson Garfinkel. *UNIX power tools*. Bantam Books, Inc., New York, NY, USA, 1993.

[8] Jefrrey E.F. Friedl. *Mastering Regular Expressions*. O'Reilly, USA, 1997. ISBN 1-56592-257-3.

[9] G. Wilson and A. Oram. *Beautiful Code: Leading Programmers Explain How They Think*. O'Reilly Media, 2008.

[10] Casiano Rodriguez Leon. *Perl: Fundamentos*. `http://nereida.deioc.ull.es/~lhp/perlexamples/`, 2001.

[11] Ian Langworth and chromatic. *Perl Testing: A Developer's Notebook*. O'Reilly Media, Inc., 2005.