



DEPARTAMENTO  
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

# Teoría de lenguajes

## Recuperatorio Trabajo práctico

### Parser

#### *Resumen*

*Este trabajo consiste en parser estilo c++*

Integrante	LU	Correo electrónico
Acosta, Javier Sebastian	338/11	acostajavier.ajs@gmail.com
Mastropasqua Nicolas Ezequiel	828/13	mastropasqua.nicolas@gmail.com
Negri, Franco	893/13	franconegri2004@hotmail.com

Palabras claves:

TP

## Contents

<b>1</b>	<b>Introducción</b>	<b>3</b>
<b>2</b>	<b>Gramatica</b>	<b>4</b>
2.1	Descripción General . . . . .	4
2.1.1	Gramática inicial . . . . .	4
2.1.2	Gramática ambigua . . . . .	4
2.1.3	Gramática no ambigua . . . . .	4
2.2	Descripción de la gramática implementada: . . . . .	4
2.2.1	Valores . . . . .	4
2.2.2	Expresiones Matemáticas . . . . .	5
2.2.3	Expresiones booleanas . . . . .	6
2.2.4	Expresiones de string . . . . .	6
2.2.5	Problema del “Dangling else”: . . . . .	6
2.3	Gramática Final . . . . .	7
<b>3</b>	<b>Implementacion del Lenguaje</b>	<b>12</b>
3.1	Implementación Del Lexer . . . . .	12
3.2	Implementacion Del Parser . . . . .	13
3.2.1	Implementación Salida . . . . .	13
3.2.2	Implementacion Del Chequeo de Tipos . . . . .	14
3.2.3	Resumen atributos . . . . .	14
3.2.4	Asignación con variables, vectores y registros . . . . .	14
3.3	Para ejecución . . . . .	15
3.4	Valores de vector y asignaciones . . . . .	16
<b>4</b>	<b>Conclusiones</b>	<b>17</b>

## 1 Introducción

El objetivo del presente trabajo práctico es la realización de un analizador léxico y sintáctico para un lenguaje de scripting, mediante el uso de una gramática LALR y la herramienta para generar analizadores sintácticos **PLY**.

Esperamos que al final del trabajo el analizador sea capaz de reconocer cadenas de caracteres similares a las del lenguaje C++.

Para ello comenzaremos generando una gramática que cumpla con las reglas del lenguaje. Luego procederemos a implementar esto y a desarrollar un chequeo de tipos que permita filtrar programas que si bien son sintácticamente validos, posean una semántica invalida (por ejemplo, multiplicar una cadena de caracteres con otra).

Además, dado un texto valido querremos guardarlo en un archivo nuevo indentandolo de manera correcta automática, agregando o quitando espacios, tabs, saltos de linea donde sea necesario.

## 2 Gramatica

### 2.1 Descripción General

#### 2.1.1 Gramática inicial

Comenzamos el armado de la gramática definiendo los tipos básicos que debíamos aceptar en nuestro lenguaje.

Así definimos los tipos Bool, String, Float, Bool, Int, Vector, Registro. De allí el siguiente paso fue ver que operaciones se le podía realizar a cada uno de ellos.

Mientras realizábamos esto empezamos a notar que necesitaríamos determinar la precedencia de estas operaciones y la asociatividad. Para esto nos basamos en la precedencia de c++<sup>1</sup> dado lo parecido que es al lenguaje de entrada.

Una vez que todas las expresiones del lenguaje estaban completas, lo pasamos a código Python usando la herramienta *PLY*. Ya terminado el código, nos encontramos con que la gramática tenía muchos mas conflictos que los que se podía ver a simple vista y por lo tanto no era LALR.

#### 2.1.2 Gramática ambigua

Luego de resolver algunos conflictos, nos dimos cuenta de que era un proceso complicado pues cada conflicto implicaba añadir mas reglas a la gramática haciendo que sea mas compleja y difícil de entender.

Después de leer la documentación de ply<sup>2</sup> y observar que ply puede encargarse de los conflictos de una gramática ambigua usando reglas de asociatividad y precedencia, decidimos hacer una gramática ambigua sencilla y aprovechar esta característica de ply.

#### 2.1.3 Gramática no ambigua

Si bien esto solucionó todos los conflictos de una manera sencilla, por desgracia, nos comunicaron que lo que se pedía era una gramática **no** ambigua. Por lo tanto decidimos tirar la gramática y comenzar de nuevo.

Con el tiempo contado y sin ideas mejores nos dimos cuenta de que la gramática inicial no estaba tan mal, tenía toda la precedencia definida de las expresiones y que una solución rápida era hacer la gramática menos restrictiva, es decir, que acepte mas cadenas de lo que se pida y rechazar las que no pertenecen al lenguaje mediante atributos. La tarea de los atributos (además de armar el código indentado) es chequear los tipos de las expresiones para que, si una operación acepta unos tipos determinados y se le pasa una expresión de otro tipo, lance una excepción de error de tipos. Dicha gramática se encuentra definida en la sección 5

### 2.2 Descripción de la gramática implementada:

#### 2.2.1 Valores

Los valores son cualquier expresión que denote un valor (por ejemplo  $3 * 4 + 3$ ). La primer idea era definir valores como **cualquier** expresión (para poder usarla como valor en asignaciones, en funciones como parámetro, etc.):

---

<sup>1</sup>[http://en.cppreference.com/w/cpp/language/operator\\_precedence](http://en.cppreference.com/w/cpp/language/operator_precedence)

<sup>2</sup> <http://www.dabeaz.com/ply/ply.html>

$Valores \rightarrow ExpresionMatematica \mid ExpresionBooleana \mid ExpresionString \mid ExpresionVector \mid ExpresionRegistro \mid OperacionVariables$

Donde cada producción que empieza con *Expresion* denota una expresión de un tipo determinado que también puede generar los valores primitivos (bool, string, float, bool, int, vec, reg), variables (id) y además funciones que retornen valores a su tipo correspondiente (por ejemplo *capitalizar* en *Expresionstring*). *OperacionVariables* genera las expresiones que usan operaciones de variables (asignación, incremento, decremento, etc).

El problema con esto es que **id** se puede generar en todas las expresiones salvo en la última. Si se tiene una producción  $A \rightarrow Valores$ , entonces se puede llegar de 5 maneras distintas a **id** (es decir, hay mas de una posible derivación mas a la izquierda para la cadena *id*) y, por lo tanto, hay conflictos (la gramática es ambigua).

Por eso decidimos sacar a **id** de cada producción. El problema es que se tiene que seguir produciendo expresiones complejas con **id** (por ejemplo *id and id*).

Para solucionar esto, decidimos sacar todos los valores primitivos de cada expresión y colocarlo en *Valores*. Además hay que observar que una función de un determinado tipo puede devolver una variable de ese tipo, por lo que decidimos sacar también a esas funciones. Lo mismo pasa con los operadores ternarios. De esta forma, cada producción *Expresion* genera expresiones con al menos un operador de su tipo:

$Valores \rightarrow ExpresionMatematica \mid ExpresionBooleana \mid ExpresionString \mid ExpresionVector \mid$   
 $\mid ExpresionRegistro \mid OperacionVariables \mid bool \mid string \mid float \mid bool \mid int \mid vec \mid reg$   
 $\mid id \mid FuncReturn \mid Ternario$   
 $FuncReturn \rightarrow FuncInt \mid FuncString \mid FuncBool \mid FuncVector$   
 $Ternario \rightarrow TernarioMat \mid TernarioBool \mid TernarioString \mid TernarioVector \mid TernarioRegistro$

### 2.2.2 Expresiones Matemáticas

Para las expresiones matemáticas consideramos la siguiente tabla de precedencia:

Tipo	Operador	Asociatividad
Binario	+, -	izquierda
Binario	*, /, %	izquierda
Binario	^	izquierda
Unario	+, -	
Unario	()	

Table 1: Tabla de menor a mayor precedencia

La producción que se encarga de realizar estas expresiones es eMat.

La primera aproximación fue hacer la gramática para poder respetar la precedencia y asociatividad de los operadores.

Este proceso se logra teniendo en cuenta que se puede fijar la precedencia (por ejemplo entre + y \*), con las reglas de la gramática. Como bien se sabe, el operador + tiene menos precedencia que el \*. Una gramática que respeta esto es la siguiente:

$$\begin{aligned} A &\rightarrow A + B \\ B &\rightarrow B * C \\ C &\rightarrow int \mid float \end{aligned}$$

Esto se debe a que antes de calcular la suma entre  $A$  y  $B$ , hay que parsear la producción  $B$  (pues es un parser ascendente).

Esto también respeta la asociatividad de ambos operados (que es a izquierda) por la misma razón que antes: La única forma de colocar mas de un operador  $+$  es usando la producción  $A$  y, por lo tanto, se tiene que parsear  $A$  antes que calcular la suma.

Este proceso se realiza con todos los operadores, de manera que si 2 operadores tienen la misma precedencia, entonces parten de un mismo no terminal.

A esta gramática le falta la posibilidad de agregar paréntesis para poder cambiar la precedencia. Una forma de hacerlos es agregarlos a  $C$  para que estos tengan mas prioridad que todas las operaciones.

Además también falta poder agregar variables. Para esto agregamos un **no-terminal** a  $C$  llamado  $id$  que las representa. La gramática resultante es esta:

$$\begin{aligned} A &\rightarrow A + B \\ B &\rightarrow B * C \\ C &\rightarrow (A) | int | float | id \end{aligned}$$

Para evitar problemas de ambigüedad, decidimos hacer que eMat devuelva expresiones con al menos un operador matemático. Esto se hizo así para evitar ambigüedades en la generación de una variable(ID). Ver explicación mas detallada en 4.1

### 2.2.3 Expresiones booleanas

Las expresiones booleanas también se solucionaron de la misma manera que las expresiones matemáticas. Para mas información ver **sección 5**.

### 2.2.4 Expresiones de string

En principio consideramos la siguiente gramática:

$$ExpresionString \rightarrow ExpresionString + ExpresionString \mid string$$

Esta gramática nos planteó un problema en el parser debido a que comparte un operador con las expresiones matemáticas. El problema viene cuando se tiene una expresión  $a + b$  en donde no está claro que tipo le corresponde e introduce una ambigüedad.

Por lo tanto decidimos fusionar a las expresiones de string con las expresiones matemáticas y restringir (mediante chequeo de tipos) que un string no se le puede hacer otra operación que no sea la suma.

### 2.2.5 Problema del “Dangling else”:

Otro problema que encontramos ya avanzados en el proceso de desarrollo de la gramática fue el problema del “Dangling else”<sup>3</sup> que consiste en que los else opcionales en un if hacen que la gramática sea ambigua.

Por ejemplo:

```
if cond1 then if cond2 then sentencia1 else sentencia2
```

---

<sup>3</sup>[https://en.wikipedia.org/wiki/Dangling\\_else](https://en.wikipedia.org/wiki/Dangling_else)

La ambigüedad surge de no poder decidir si la sentencia2 se ejecuta como rama alternativa a cond1( es decir, cuando esta condición es falsa), o bien como la rama alternativa del *if* asociado a **cond2**. Esto último es posible ya que el *if* de **cond1** puede prescindir del bloque *else*. Para solucionar esto, construimos una solución basada en el concepto de líneas abiertas o cerradas. La primera, son todas aquellas que contienen al menos un *if* que no tiene un *else* correspondiente o bien un bucle con línea abierta. La segunda mencionada, son aquellas que resultan de tener sentencias simples, comentarios, o bien:

```
if (cond1) bloqueCerrado else bloqueCerrado
```

Donde bloqueCerrado puede ser cualquier conjunto de sentencias escritas entre llaves, sentencias simples o , razonando inductivamente, líneas cerradas. También se incluyen los bucles con líneas cerrados. Finalmente, la situación se complejizó un poco ya que hay que considerar las distintas combinaciones recién mencionadas, sumando la posibilidad de tener, o no, comentarios en cada uno de estos bloques.

## 2.3 Gramática Final

A continuación se define la gramática utilizada para construir el parser. La misma es **no ambigua** y **LALR**. La garantía de esto es que ply acepta gramáticas de este tipo, y nuestra implementación no arroja conflictos **shift/reduce** o **reduce/reduce**

$$G = \langle V_t, V_{nt}, g, P \rangle$$

$V_t$  es el conjunto de símbolos terminales dado por los símbolos en **mayúsculas** que aparecen en las producciones.

$V_{nt}$  es el conjunto de símbolos no-terminales dado los literales(operadores) y los símbolos en **minúsculas** que aparecen en las producciones.

$P$  es el conjunto de producciones dadas a continuación

### Sentencias y estructura general

$$g \rightarrow \text{línea } g \mid \text{COMMENT } g \mid \text{empty}$$

$$\text{línea} \rightarrow \text{lAbierta} \mid \text{lCerrada}$$

*Línea Abierta: Hay por lo menos un IF que no matchea con un else*

$$\begin{aligned} \text{lAbierta} \rightarrow & \text{IF (cosaBooleana) línea} \mid \\ & \mid \text{IF (cosasBooleana) COMMENT com línea} \\ & \mid \text{IF (cosasBooleana) COMMENT com lCerrada ELSE lAbierta} \\ & \mid \text{IF (cosasBooleana) COMMENT com lCerrada ELSE COMMENT com lAbierta} \\ & \mid \text{IF (cosasBooleana) lCerrada ELSE lAbierta} \\ & \mid \text{IF (cosasBooleana) lCerrada ELSE COMMENT com lAbierta} \\ & \mid \text{IF (cosasBooleana) \{ g \} ELSE lAbierta} \end{aligned}$$

| IF (cosasBooleana) { g }  
 | loop Labierta

*Las siguientes son las variantes de tener bloques cerrados else bloques cerrados.*

*Un bloque cerrado puede ser una sentencia única o un bloque entre llaves.*

*En cada uno de estos casos puede haber, o no, comentarios. De ahí todas estas combinaciones.*

lCerrada  $\rightarrow$  sentencia

| IF (cosaBooleana) { g } ELSE { g }  
 | IF ( cosaBooleana ) lCerrada ELSE { g }  
 | IF ( cosaBooleana ) COMMENT com lCerrada ELSE { g }  
 | IF ( cosaBooleana ) { g } ELSE lCerrada  
 | IF ( cosaBooleana ) { g } ELSE lCerrada  
 | IF ( cosaBooleana ) lCerrada ELSE lCerrada  
 | IF ( cosaBooleana ) COMMENT com lCerrada ELSE lCerrada  
 | IF ( cosaBooleana ) lCerrada ELSE COMMENT com lCerrada  
 | IF ( cosaBooleana ) COMMENT com lCerrada ELSE COMMENT com lCerrada  
 | loop { g }  
 | loop lCerrada  
 | loop COMMENT com lCerrada  
 | DO { g } WHILE ( valores ) ;  
 | DO lCerrada WHILE ( valores ) ;  
 | DO COMMENT com lCerrada WHILE ( valores ) ;  
 | DO lCerrada COMMENT com WHILE ( valores ) ;

com  $\rightarrow$  COMMENT com |  $\lambda$

#### **Sentencias básicas:**

sentencia  $\rightarrow$  varsOps | func ; | varAsig ; | RETURN; | ;

#### **Bucles:**

loop  $\rightarrow$  WHILE (valores) | FOR (primerParam ; valores ; tercerParam)

primerParam  $\rightarrow$  varAsig |  $\lambda$

tercarParam  $\rightarrow$  varsOps | varAsig | func |  $\lambda$

cosaBooleana  $\rightarrow$  expBool | valoresBool

#### **Funciones:**

func  $\rightarrow$  FuncReturn | FuncVoid



$\text{funcReturn} \rightarrow \text{FuncInt} \mid \text{FuncString} \mid \text{FuncBool}$   
 $\text{funcInt} \rightarrow \text{MULTIESCALAR}(\text{valores}, \text{valores param})$   
 $\text{funcInt} \rightarrow \text{LENGTH}(\text{valores})$   
 $\text{funcString} \rightarrow \text{CAPITALIZAR}(\text{valores})$   
 $\text{funcBool} \rightarrow \text{colineales}(\text{valores}, \text{valores})$   
 $\text{funcVoid} \rightarrow \text{print}(\text{Valores})$   
 $\text{param} \rightarrow \text{valores} \mid \lambda$

#### Vectores y variables:

$\text{vec} \rightarrow [\text{elem}]$   
 $\text{elem} \rightarrow \text{valores}, \text{elem} \mid \text{valores}$   
 $\text{vecval} \rightarrow \text{id} [\text{expresion}] \mid \text{vec} [\text{expresion}] \mid \text{vecVal} [\text{expresion}] \mid \text{atributos} [\text{expresion}]$   
 $\quad \mid \text{id} [\text{INT}] \mid \text{vec} [\text{INT}] \mid \text{vecVal} [\text{INT}] \mid \text{atributos} [\text{INT}]$   
 $\text{expresion} \rightarrow \text{eMat} \mid \text{expBool} \mid \text{funcReturn} \mid \text{reg} \mid \text{FLOAT} \mid \text{STRING} \mid \text{RES}$   
 $\quad \mid \text{BOOL} \mid \text{varYVals} \mid \text{varsOps} \mid \text{vec} \mid \text{atributos} \mid \text{ternario}$   
 $\text{valores} \rightarrow \text{varYVals} \mid \text{varsOps} \mid \text{eMat} \mid \text{expBool} \mid \text{funcReturn} \mid \text{reg} \mid \text{INT} \mid \text{FLOAT}$   
 $\quad \mid \text{STRING} \mid \text{BOOL} \mid \text{ternario} \mid \text{atributos} \mid \text{vec} \mid \text{RES}$   
 $\text{atributos} \rightarrow \text{ID.valoresCampos} \mid \text{reg.valoresCampos}$   
 $\text{valoresCampos} \rightarrow \text{ID} \mid \text{end} \mid \text{begin}$   
 $\text{end} \rightarrow \text{END}$   
 $\text{begin} \rightarrow \text{BEGIN}$

#### Operadores ternarios:

$\text{ternario} \rightarrow \text{ternarioMat} \mid \text{ternarioBool} \mid (\text{ternarioBool}) \mid (\text{ternarioMat})$   
 $\quad \mid \text{ternarioVars} \mid (\text{ternarioVars})$   
 $\text{ternarioVars} \rightarrow \text{valoresBool} ? \text{valoresTernarioVars} : \text{valoresTernarioVars}$   
 $\quad \mid \text{valoresBool} ? \text{valoresTernarioVars} : \text{valoresTernarioMat}$   
 $\quad \mid \text{valoresBool} ? \text{valoresTernarioMat} : \text{valoresTernarioVars}$   
 $\quad \mid \text{valoresBool} ? \text{valoresTernarioVars} : \text{valoresTernarioBool}$   
 $\quad \mid \text{valoresBool} ? \text{valoresTernarioBool} : \text{valoresTernarioVars}$   
 $\quad \mid \text{expBool} ? \text{valoresTernarioVars} : \text{valoresTernarioVars}$   
 $\quad \mid \text{expBool} ? \text{valoresTernarioVars} : \text{valoresTernarioMat}$   
 $\quad \mid \text{expBool} ? \text{valoresTernarioMat} : \text{valoresTernarioVars}$   
 $\quad \mid \text{expBool} ? \text{valoresTernarioVars} : \text{valoresTernarioBool}$   
 $\quad \mid \text{expBool} ? \text{valoresTernarioBool} : \text{valoresTernarioVars}$

valoresTernarioVars  $\rightarrow$  reg | vec | ternarioVars | (ternarioVars) | atributos  
 | varsOps | varYVals | RES

TernarioMat  $\rightarrow$  valoresBool ? valoresTernarioMat : valoresTernarioMat  
 | expBool ? valoresTernarioMat : valoresTernarioMat

valoresTernarioMat  $\rightarrow$  INT | FLOAT | funcInt | STRING | eMat  
 | ternarioMat | (ternarioMat)

ternarioBool  $\rightarrow$  valoresBool ? valoresTernarioBool : valoresTernarioBool  
 | expBool ? valoresTernarioBool : valoresTernarioBool

valoresTernarioBool  $\rightarrow$  BOOL | funcBool | ternarioBool | ( ternarioBool )| expBool

#### varYVals:

varYVals  $\rightarrow$  ID | vecVal | vecVal.varYVals

#### Registros:

reg  $\rightarrow$  { campos }

campos  $\rightarrow$  ID:valores, campos | ID:valores

#### Operadores de variables:

varsOps  $\rightarrow$  MENOSMENOS variable | MASMAS variable  
 | variable MASMAS | variable MENOSMENOS

variable  $\rightarrow$  ID | vecVal | vecVal.varYVALS

#### Asignaciones:

varAsig  $\rightarrow$  variable MULEQ valores | variable DIVEQ valores | variable MASEQ valores  
 | variable MENOSEQ valores | variable = valores | ID . ID = valores

#### Operaciones binarias enteras:

valoresMat  $\rightarrow$  INT | FLOAT | funcInt | atributos | funcString  
 | STRING | varYVals | varsOps | (ternarioMat)

eMat  $\rightarrow$  eMat + p | valoresMat + p | eMat + valoresMat | valoresMat + valoresMat  
 | eMat - p | valoresMat - p | eMat - valoresMat | valoresMat - valoresMat | p

p  $\rightarrow$  p \* exp | p / exp | p | valoresMat \* exp | valoresMat / exp | valoresMat  
 | p \* valoresMat | p / valoresMat | p % valoresMat | valoresMat \* valoresMat  
 | valoresMat valoresMat | valoresMat % valoresMat | exp

exp  $\rightarrow$  exp iSing | valoresMat iSing | exp ^valoresMat  
 | valoresMat ^valoresMat | iSing

iSing  $\rightarrow$  - paren | + paren | - valoresMat | + valoresMat | paren

paren  $\rightarrow$  ( eMat ) | ( valoresMat )

**Expresiones booleanas:**

valoresBool  $\rightarrow$  BOOL | funcBool l | varYVals | varsOps | ( ternarioBool )

expBool  $\rightarrow$  expBool OR and | valoresBool OR and |

| expBool OR valoresBool | valoresBool OR valoresBool | and

and  $\rightarrow$  and AND eq | valoresBool AND eq | and AND valoresBool

| valoresBool AND valoresBool | eq

eq  $\rightarrow$  eq EQEQ mayor | eq DISTINTO mayor | tCompareEQ EQEQ mayor

| tCompareEQ DISTINTO mayor | eq EQEQ tCompareEQ

| eq DISTINTO tCompareEQ tCompareEQ EQEQ tCompareEQ

| tCompareEQ DISTINTO tCompareEQ | mayor

tCompareEQ  $\rightarrow$  BOOL | funcBool | varYVals | varsOps | INT

| FLOAT | funcInt | eMat | ( ternarioBool ) | ( ternarioMat )

tCompare  $\rightarrow$  eMat | varsOps | varYVals | INT | funcInt | FLOAT | ( ternarioMat )

mayor  $\rightarrow$  tCompare > tCompare | menor

menor  $\rightarrow$  tCompare < tCompare | not

not  $\rightarrow$  NOT not | NOT valoresBool | parenBool

parenBool  $\rightarrow$  ( expBool )

## 3 Implementacion del Lenguaje

### 3.1 Implementación Del Lexer

Para la construcción del lexer se definió un conjunto de literales con operadores y otros símbolos propios del lenguaje.

**literals** = `[+, -, *, /, %, <, >, =, !, {, }, (, ), [, ], ?, :, ;, ,, .]`

A su vez, utilizamos otra funcionalidad de ply para definir las palabras reservadas del lenguaje:

**reserved** = `{ 'begin' : 'BEGIN', 'end' : 'END', 'while' : 'WHILE', 'for' : 'FOR',  
'if' : 'IF', 'else' : 'ELSE', 'do' : 'DO', 'res' : 'RES', 'return' : 'RETURN',  
'AND' : 'AND', 'OR' : 'OR', 'NOT' : 'NOT', 'print' : 'PRINT',  
'multiplicacionEscalar' : 'MULTIESCALAR', 'capitalizar' : 'CAPITALIZAR',  
'colineales' : 'COLINEALES', 'print' : 'PRINT', 'length' : 'LENGTH', }`

Esto permitió evitar tener que definir demasiadas reglas simples para este tipo de operadores o palabras claves.

Para el resto de los tokens definidos, fue necesario utilizar expresiones regulares, se muestra mas abajo como se definieron cada una de ellos: (Por claridad, se omite el resto del código para la regla de los tipos, ya que es análoga a la de **string**):

<code>t_EQEQ = r"=="</code>	<code>def t_STRING(token):</code>
<code>t_DISTINTO = r"!="</code>	<code>r'\'\'\' \. *? \\'\'\'</code>
<code>t_MENOSEQ = r"-="</code>	<code>atributos = {}</code>
<code>t_MASEQ = r"\+="</code>	<code>atributos["type"] = "string"</code>
<code>t_MULEQ = r"\*="</code>	<code>atributos["value"] = token.value</code>
<code>t_DIVEQ = r"/="</code>	<code>token.value = atributos</code>
<code>t_MASMAS = r"\+\+"</code>	<code>return token</code>
<code>t_MENOSMENOS = r"-"</code>	
<code>def t_BOOL(token) :</code>	<code>def t_ID(token):</code>
<code>r"true   false"</code>	<code>r"[a-zA-Z_][a-zA-Z_0-9]*"</code>
<code>def t_FLOAT(token):</code>	<code>t = token.value</code>
<code>r"[.]?[0-9]</code>	<code>tipo = reserved.get(token.value)</code>
<code>def t_INT(token) :</code>	<code>if t.lower() not in reserved and t.upper()</code>
<code>r"[1-9][0-9]* — 0"</code>	<code>not in reserved and t !=</code>
	<code>"multiplicacionEscalar":</code>
	<code>tipo = 'ID'</code>
	<code>token.type = tipo</code>

```

def t_NEWLINE(token):
    r"\n+"
    token.lexer.lineno += len(token.value)
def t_error(token):
    message = "Token desconocido:"
    message += "\n type:" + token.type
    message += "\n value:" + str(token.value)
    message += "\n line:" + str(token.lineno)
    message += "\n position:" + str(token.lexpos)
    raise Exception(message)

def t_COMMENT(token):
    r'#. *'
    t_ignore = ' \ t'

```

## 3.2 Implementacion Del Parser

La implementación del parser consistió en transcribir la gramática final del apartado 2 a la sintaxis de ply.

De esta manera, dada una producción:

$$Valores \rightarrow ExpresionMatematica$$

Fue necesario reescribirla como:

```

def p_valores(subexpressions) :
    """ valores : ExpresionMatematica """

```

Ademas contamos con la funcionalidad de ply que, una vez que se a utilizado una producción permite ejecutar código adicional. Siguiendo el ejemplo anterior, suponiendo que cada vez que el parser utiliza la producción antedicha deseo imprimirla, puedo escribir:

```

def p_valores(subexpressions) :
    """ valores : ExpresionMatematica """
    print subexpressions[1]

```

Esto lo utilizamos tanto para escribir el output formateado con la salida correcta como para realizar el chequeo de tipos.

### 3.2.1 Implementación Salida

Para la implementación de la salida cada producción de la gramática tendrá un atributo sintetizado "value" en el que se guardará el texto de salida basándose en el valor de cada uno de los terminales y no terminales que lo componen. Esto es relativamente sencillo para las sentencias que solo requieren escribir sus terminales y no terminales seguidos de un salto de linea, pero requerirá un cuidado especial para el caso de los condicionales y los loops, ya que estos necesitan saltos de linea en lugares intermedios y una indentación adecuada.

Para ejemplificar que debemos hacer utilizamos la producción:

$lAbierta \rightarrow IF ( cosaBooleana ) COMMENT com lCerrada ELSE \{ g \}$

como caso de estudio.

Lo que queremos guardar en el atributo value de lAbierta es primero un  $IF ( cosaBooleana )$ . Nótese aquí que cosaBooleana es un conjunto de símbolos que deben reducir como ultima instancia a una Expresión Booleana. Queda a cargo de cosaBooleana la responsabilidad de saber como imprimir cada uno de esos posibles términos que se encuentren presentes.

Esto estará seguido de un salto de linea, seguido de un comentario (con un tab) y un salto de linea, luego un no terminal que contendrá cero o uno o varios comentarios (todos conteniendo su indentación apropiada), otro salto de linea con una palabra reservada else, unas llaves que abren y una o varias sentencias indentadas un corchete que cierra.

En primera instancia, la salida que nos dará este if es la deseada, sin embargo puede que ocurra el caso donde tenemos dos ifs anidados. En este caso lo esperable es que el segundo if (el interno) este indentado y que las sentencias dentro de ese if tengan dos tabs en vez de uno. Luego, no es suficiente con agregar tabs en los lugares adecuados, también necesitamos señalar donde comienza una nueva linea y en el caso de que sea necesario agregar varios tabs, que se pueda recorrer la salida agregándolos donde se necesite.

Para el manejo de la indentación utilizaremos el comienzo de cada nueva linea.

### 3.2.2 Implementacion Del Chequeo de Tipos

#### 3.2.3 Resumen atributos

Para el chequeo de tipos utilizamos atributos sintetizados que se enumeran a continuación:

- var : Denota la variable de la expresión (si es que hay solo una variable).
- type : Denota el tipo de la expresión.
- elems : Si la expresión es de tipo vector, denota los tipos de los elementos del vector.
- varsVec : Si la expresión es de tipo vector, denota las variables de los elementos del vector (si tienen alguna) para la asignación.
- regs: Si la expresión es de tipo vector de registros, denota los registros para la asignación.
- campo: denota el valor de un campo de un registro.
- varAsig e indice : denota la variable y el indice en asignaciones como  $a[i] = 1$ , en donde la expresión a la derecha del "=" **no** es un valor.

#### 3.2.4 Asignación con variables, vectores y registros

Los tipos posibles en el atributo *type* son: string, float, int, bool, vec y reg. En algunas producciones se usan como simples atributos sintetizados, en el caso de que una expresión sea una variable se necesita saber que tipo tuvo esa variable cuando se inicializó.

Para esto se cuenta con el diccionario global *variables* que tiene de claves a las variables y de valor su tipo (con fines declarativos hay otro diccionario dentro del valor, que tiene una clave "type" y el valor es el tipo de la variable).

Algo similar ocurre con variables que referencian a vectores. Se agrega un diccionario global *vectores* en donde se guarda la variable como clave y una lista de tipos como valor, que denotan los tipos de las respectivas posiciones del vector.

Para los vectores de vectores hay un diccionario global *variablesVector* que guarda como clave a la variable correspondiente al vector de vectores y como valor una lista de variables que se corresponden con las variables de los vectores dentro del vector.

Como se pueden hacer operaciones tales como  $[1,2,3][0]$  en donde no se involucra ninguna variable, se agrego el atributo *elems* que se encarga de obtener esa lista de tipos del vector. Observar que, como no existe referencia a este vector, no es necesario guardarlo en el diccionario *vectores* ya que no se utilizará en otro lugar.

En caso de los registros se guarda un atributo *campos* en donde se guardan los campos del registro. En el se guarda una lista de tuplas, donde la primer coordenada corresponde al nombre del campo, la segunda al tipo del campo y la tercera a la variable del vector si tuviera un campo de tipo vector.

En caso de las asignaciones se guarda en un diccionario global *registros* la variable correspondiente al registro y los campos obtenidos a través del atributos *campos*.

### 3.3 Para ejecución

Cuando se establece en un atributo *type* o *var* el string "Para ejecucion" significa que el parser admite la cadena pero se tiene que chequear en tiempo de ejecución su tipo.

La siguiente tabla muestra las operaciones donde se delega el chequeo de tipos a tiempo de ejecución:

Descripción	Ejemplo
chequear que el índice del operador <code>[]</code> (acceso a vector) sea entero en expresiones complejas	$a[4+5*3]$ $a[b < 2 ? 1 : 2]$ $a[b]$
Acceso a vectores de mas de 2 dimensiones	$a[4][5][1]$
Cadenas de acceso de registros y vectores de mas de 3 subexpresiones	$a.campo1[5].campo2[3]$ $a[5].campo1[4].campo2[4]$
Decidir un tipo para una operación donde un operando es algunos de los casos anteriores	$a[6][6][6] + 2$ not $a[b]$

### 3.4 Valores de vector y asignaciones

VecVal es la regla que permite crear accesos a vectores (por ejemplo  $a[5]$ ). Esta regla chequea que el índice sea menor que el tamaño del vector si el índice viene dado por un entero, en caso de cualquier otra expresión (variable, suma, etc) no se chequea nada. También setea el tipo correspondiente a la posición del vector y su variable si es que existe.

Consideremos el siguiente ejemplo:

$$a[1] = b[2];$$

La expresión de la derecha viene de VecVal y también la de la izquierda. El problema es que estas dos expresiones tienen distinto significado. La de la derecha corresponde a un valor de un vector, no interesa el índice ni que tipos tiene el vector. La de la izquierda, en cambio, refiere a la posición del vector y por lo tanto se requiere saber el índice y la variable a la cual se refiere.

Para solucionar esto creamos un no terminal nuevo llamado **Variable** donde se encarga de establecer tanto la variable como el índice del no terminal **VecVal**. También en VecVal guardamos la información del índice y de la variable en los atributos *indice* y *varAsig* respectivamente. Por último, se chequea que cualquier asignación que se le haga a una variable que tenga el atributo índice definido (es decir, como la del ejemplo) tenga el tipo adecuado. Por ejemplo, si  $a[1]$  se corresponde con la primera posición de un vector de enteros,  $b[2]$  debe ser necesariamente un entero o se levantará una excepción.

Si VecVal no tiene un atributo "var" entonces tira una excepción "Solo se puede acceder a variables."

Además en operaciones como  $(a = 2) ? 5 : 2$ , se establece como tipo de la asignación el tipo de la derecha y como variable el atributo var de la expresión de la izquierda.



## 4 Conclusiones

Este trabajo nos permitió tener una idea de como generar una gramática para un lenguaje de programación, cuales son los mayores desafíos, el problema de las ambigüedades en el lenguaje, como en el problema del dangling else, y también sirvió como disparador para pensar diferentes soluciones a estas problemáticas. Además, sirvió para discutir temas tales como que chequeos de tipo pueden realizarse efectivamente en tiempo de compilación y cuales no queda mas alternativa que solucionarlos en tiempo de ejecución.