Diseño e Implementación de un Analizador Léxico y Analizador Semántico para el Lenguaje AVISMO

Aramis E. Matos

Lenier Gerena

Angel Berrios Pellot

Segundo Semestre, 2022-2023

Tabla de Contenido

1	Intr	oducción	2		
2	Analizador Léxico				
	2.1	Gramatica del Lenguaje AVISMO	4		
	2.2	Diseño del del Analizador Léxico	12		
		2.2.1 Autómatas Finitos Deterministas	12		
		2.2.2 Tabla de Símbolos	13		
	2.3	Implementación del Analizador Léxico	14		
3	Imp	lementación del Analizador Sintáctico	18		
4	Con	clusiones y Recomendaciones	19		
Re	eferen	icias Bibliográficas	22		

Introducción

La visualización molecular puede ser considerada como una de las áreas mas importante dentro de la bioinformática. Entre sus aplicaciones mas relevantes se destacan el diseño de nuevo fármacos... (Narciso Farias, Rios, Hidrobo, & Vicuña, 2012). Este enunciado fue escrito hace mas de una década. Sin embargo, hoy día en un mundo pospandemia, reconocemos que tan sabio fue. El desarollo de la vacuna contra el COVID-19 tan rápido fue gracias a herramientas de visualización como el Ambiente de visualización Molecular (AVISMO) (Narciso Farias et al., 2012) El propósito de este proyecto es definir el automata de estado finito del lenguaje AVISMO, los patrones el cual caracterizan los lexemas del lenguaje, los atributos de los lexemas y la implementación del analizador léxico y sintáctico del lenguaje AVISO en C++. La implementación léxica se desarolló utilizando GNU Flex (Flex - a Scanner Generator, n.d.) con la implementación de Calc++ por bwasti (Wasti, 2020) como base. La implementación sintáctica se desarolló

en GNU Bison (*Bison - GNU Project - Free Software Foundation*, n.d.) con la implementación de Calc++ por bwasti (Wasti, 2020) como base.

Analizador Léxico

2.1 Gramatica del Lenguaje AVISMO

- <SENTENCIAS> ::= <FIN_DE_LINEA> <SENTENCIAS> | <SENTEN-CIA> <FIN_DE_LINEA>
- <FIN_DE_LINEA> ::= ":" | ";"
- <SENTENCIA> ::= "defina" <ID> "como" <TIPO> | <ID> "=" <MOD-ELO_MOLECULAR> | <OPERACION> "(" <ID> ")"
- <ID> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" | "J" | "K" | "L" |

 "M" | "N" | "O" | "P" | "Q" | "R" | "S" | "T" | "U" | "V" | "W" | "X" | "Y" |

 "Z" | "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" | "j" | "k" | "l" | "m" | "n"

 | "o" | "p" | "q" | "r" | "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z" | <LETRA>

 <IDCONT>

- <IDCONT> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" | "J" | "K" | "L" | "M" | "N" | "O" | "P" | "Q" | "R" | "S" | "T" | "U" | "V" | "W" | "X" | "Y" | "Z" | "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" | "j" | "k" | "1" | "m" | "n" | "o" | "p" | "q" | "r" | "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z" | <LETRA> <IDCONT> | "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9" | <DIGITO> <IDCONT>
- <LETRA> ::= "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" | "J" | "K" | "L" | "M" | "N" | "O" | "P" | "Q" | "R" | "S" | "T" | "U" | "V" | "W" | "X" | "Y" | "Z" | "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" | "j" | "k" | "l" | "m" | "n" | "o" | "p" | "q" | "r" | "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z"
- <DIGITO> ::= "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"
- <TIPO> ::= "modelo"
- <OPERACION> ::= "graficar2d" | "graficar3d" | "pesomolecular"
- <MODELO_MOLECULAR> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn" | <ELEMENTO_QUIMICO> <VALENCIA> | <ELE-

MENTO> <GRUPO_FUNCIONAL> | <COMPUESTO> <ELEMENTO> <GRUPO_FUNCIONAL> | <COMPUESTO> <COMPUESTO>

- <COMPUESTO> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" |

 "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V"

 | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh"

 | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" |

 "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" |

 "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" |

 "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" |

 "Xe" | "Rn" | <ELEMENTO | QUIMICO > <VALENCIA > | <ELEMENTO >

 <GRUPO_FUNCIONAL > | <ELEMENTO > <GRUPO_FUNCIONAL > <EN
 LACE > | <ELEMENTO > <ENLACE >
- <COMPUESTOS> ::= <COMPUESTO> <COMPUESTO> | <COMPUESTOS>
- <ELEMENTO> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn" | <ELEMENTO_QUIMICO> <VALENCIA>
- <ELEMENTO_QUIMICO> ::= "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" |

"Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" |

"V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" |

"Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu"

| "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si"

| "Ge" | "Sn" | "Pb" | "N" | "P" | "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" |

"Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn"

- <VALENCIA> ::= "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"
- <GRUPO_FUNCIONAL> ::= <GRUPO_FUNCIONAL_INFERIOR>
 <GRUPO_FUNCIONAL_SUPERIOR> | <GRUPO_FUNCIONAL_SUPERIOR>
 <GRUPO_FUNCIONAL_INFERIOR> | "(" <MODELO_GRUPO_FUNCIONAL>
 ")" | "[" <MODELO_GRUPO_FUNCIONAL> "]"
- <GRUPO_FUNCIONAL_SUPERIOR> ::= "[" < MODELO_GRUPO_FUNCIONAL> "]"
- <GRUPO_FUNCIONAL_INFERIOR> ::= "(" < MODELO_GRUYPO_FUNCIONAL> ")
- <MODELO_GRUYPO_FUNCIONAL> ::= <ENLACE> <MODELO_MOLECULAR> | "H" | "Li" | "Na" | "K" | "Rb" | "Cs" | "Fr" | "Be" | "Mg" | "Ca" | "Sr" | "Ba" | "Ra" | "Sc" | "Y" | "Ti" | "Zr" | "Hf" | "Db" | "V" | "Nb" | "Ta" | "Ji" | "Cr" | "Mo" | "W" | "Rf" | "Mn" | "Tc" | "Re" | "Bh" | "Fe" | "Ru" | "Os" | "Hn" | "Co" | "Rh" | "Ir" | "Mt" | "Ni" | "Pd" | "Pt" | "Cu" | "Ag" | "Au" | "Zn" | "Cd" | "Hg" | "B" | "Al" | "Ga" | "In" | "Ti" | "C" | "Si" | "Ge" | "Sn" | "Pb" | "N" | "P"

| "As" | "Sb" | "Bi" | "O" | "S" | "Se" | "Te" | "Po" | "F" | "Cr" | "Br" | "I" | "At" | "He" | "Ne" | "Ar" | "Kr" | "Xe" | "Rn" | <ELEMENTO_QUIMICO> <VA-LENCIA> | <ELEMENTO> <GRUPO_FUNCIONAL> | <COMPUESTO> <ELEMENTO> | <COMPUESTO> <COMPUESTO> <COMPUESTO> <

En la tabla 2.1, en la columna de patrones, note que cuando dice {TOKEN} donde TOKEN se refiere a el patrón asociado a token. Por ejemplo, si un patrón dice {ELEMENTO_QUIMICO}, esto significa que inserta el patrón asociado al token ELEMENTO_QUIMICO. Esto no significa que el analizador léxico espera un token de por si, sencillamente se hizo con el propósito de evitar redundancias.

<fin_de_linea></fin_de_linea>	;1:	:	Simbolo reservado
<palabra< td=""><td>defina I como</td><td>defina</td><td>Palabra reservada</td></palabra<>	defina I como	defina	Palabra reservada
_RESERVADA>			
<id></id>	[A-Za-z][A-Za-z0-9]*	var1	Modelo molecular aso-
			ciado
<idcont></idcont>	[A-Za-z0-9]+	1ar	ID asociado
<letra></letra>	[A-Za-z]	a	ID asociado
<digito></digito>	[0-9]	7	Valor numérico, lexema
			asociado
<tipo></tipo>	modelo	modelo	ID asociado
<operacion></operacion>	graficar2d graficar3d pesomolecular	pesomolecular	ID asociado
<modelo< td=""><td>({ELEMENTO _QUIMICO} {ELEMENTO _QUIMICO} {VALEN-</td><td>СН3(СН3)СНН</td><td>ID asociado</td></modelo<>	({ELEMENTO _QUIMICO} {ELEMENTO _QUIMICO} {VALEN-	СН3(СН3)СНН	ID asociado
_MOLECULAR>	CIA} {ELEMENTO} {GRUPO _FUNCIONAL} {ELEMENTO}		
	{GRUPO_FUNCIONAL} {ENLACE} {ELEMENTO} {ENLACE})		

Atributos

Lexema

Token

Patrón

<compuesto></compuesto>	COMPUESTO ({ELEMENTO _QUIMICO} {ELEMENTO _QUIM-	CH3::	Modelo molecular aso-
	ICO} {VALENCIA} {ELEMENTO} {GRUPO_FUNCIONAL} {ELE-		ciado, enlaces, valencias
	MENTO} {GRUPO_FUNCIONAL} {ENLACE} {ELEMENTO} {EN-		
	LACE})		
<compuestos></compuestos>	{COMPUESTO}+	CH3::(OH)3	Modelo molecular aso-
			ciado, enlaces, valencias
<elemento></elemento>	{ELEMENTO _QUIMICO} {VALENCIA}?	Ag3	Elemento, valencia
<elemento< td=""><td>("H" "Li" "Na" "K" "Rb" "Cs" "Fr" "Be" "Mg" "Ca" "Sr" </td><td>I</td><td>Elemento</td></elemento<>	("H" "Li" "Na" "K" "Rb" "Cs" "Fr" "Be" "Mg" "Ca" "Sr"	I	Elemento
_QUIMICO>	"Ba" "Ra" "Sc" "Y" "Ti" "Zr" "Hf" "Db" "V" "Nb" "Ta" "Ji"		
	"Cr" "Mo" "W" "Rf" "Mn" "Tc" "Re" "Bh" "Fe" "Ru" "Os"		
	"Hn" "Co" "Rh" "Ir" "Mt" "Ni" "Pd" "Pt" "Cu" "Ag" "Au"		
	"Zn" "Cd" "Hg" "B" "Al" "Ga" "In" "Ti" "C" "Si" "Ge" "Sn"		
	"Pb" "N" "P" "As" "Sb" "Bi" "O" "S" "Se" "Te" "Po" "F"		
	"Cr" "Br" "I" "At" "He" "Ne" "Ar" "Kr" "Xe" "Rn")		
<valencia></valencia>	[1-9]	2	Valor

<grupo _fun-<="" th=""><th>({GRUPO _FUNCIONAL _INFERIOR} {GRUPO _FUNCIONAL</th><th>(CH3){Ag2}</th><th>Grupos funcionales,</th></grupo>	({GRUPO _FUNCIONAL _INFERIOR} {GRUPO _FUNCIONAL	(CH3){Ag2}	Grupos funcionales,					
CIONAL>	_SUPERIOR} {GRUPO _FUNCIONAL _SUPERIOR} {GRUPO		grupo funcional infe-					
	_FUNCIONAL_INFERIOR} "(" {MODELO _GRUPO _FUN-	rior, grupo funcional						
	CIONAL} ")" "[" MODELO _GRUPO _FUNCIONAL "]")		superior					
<grupo _fun-<="" td=""><td>"[" {MODELO _GRUPO _FUNCIONAL} "]"</td><td>[CVHe3]</td><td colspan="6">Elementos, valencias</td></grupo>	"[" {MODELO _GRUPO _FUNCIONAL} "]"	[CVHe3]	Elementos, valencias					
CIONAL _INFE-								
RIOR>								
<grupo _fun-<="" td=""><td>"(" {MODELO _GRUPO _FUNCIONAL} ")"</td><td>(CVHe3)</td><td>Elementos, valencias</td></grupo>	"(" {MODELO _GRUPO _FUNCIONAL} ")"	(CVHe3)	Elementos, valencias					
CIONAL _SUPE-								
RIOR>								
<modelo< td=""><td>({ELEMENTO _QUIMICO}+ {VALENCIA}?)+ ({ELEMENTO}+</td><td>FeH=C3Si4</td><td>Elementos, enlaces, va-</td></modelo<>	({ELEMENTO _QUIMICO}+ {VALENCIA}?)+ ({ELEMENTO}+	FeH=C3Si4	Elementos, enlaces, va-					
_GRUPO _FUN-	{ENLACE} {ELEMENTO}+)+		lencias					
CIONAL>								
<enlace></enlace>	("-" "=" ":" "::")	-	Valencia					

Tabla 2.1: Tabla de Componentes Léxicos de AVISMO

2.2 Diseño del del Analizador Léxico

2.2.1 Autómatas Finitos Deterministas

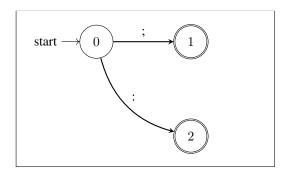


Figura 2.1: Automata del patrón para el token <FIN_DE_LINEA>

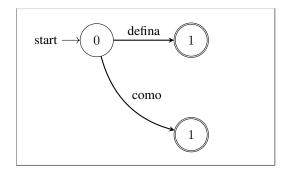


Figura 2.2: Automata del patrón para el token <PALABRAS_RESERVADA>

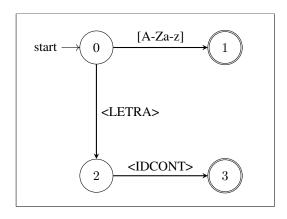


Figura 2.3: Automata del patrón para el token <ID>

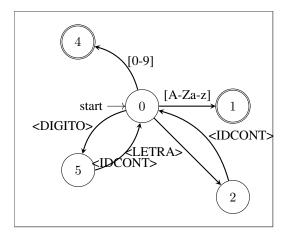


Figura 2.4: Automata del patrón para el token <IDCONT>





Figura 2.5: Automata del patrón para el token <ASIGNACION>

Figura 2.6: Automata del patrón para el token <LETRA>



Figura 2.7: Automata del patrón para el token <DIGITO>

2.2.2 Tabla de Símbolos

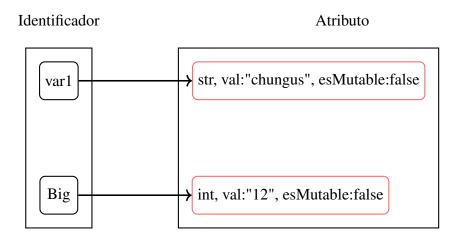


Figura 2.8: Tabla de símbolos implementada como un diccionario

2.3 Implementación del Analizador Léxico

Como se ha mencionado anteriormente, la implementación lexica del proyecto fue inspirada por la implementación de Calc++ por bwasti y adaptada para la gramática de AVISMO (Wasti, 2020).

Para compilar el programa, primero que todo se tiene que ejecutar *make clean* por la línea dentro del directorio *code*, Esto se hace con el propósito de evitar errores de compilación. Entonces, se ejecuta el comando *make*, estó compilará todas las dependencias necesarias, en particular, los archivos de contiene el léxico de Flex (con todos los archivos con extensión .cc, .hh y .ll). Además, el comando anterior compila todos los programas en código objeto (.o) que se crean para el programa *driver* (el cual está encargado de abrir archivos de entrada e instanciar el analizador), parser y scanner. Posterior a esto es que se puede ejecutar ./avismo fileName.txt el cual para nuestro caso seria el siguiente:

./avismo test_prog.txt

Al ejecutarse el comando anterior, el programa procede a leer **cada caracter** del programa e identificar si una serie de caracteres sigue un patrón que forma parte del lenguaje AVISMO. Al encontrar un patrón reconocido, tales como un identificador o modelo molecular, lo clasifica con un *token* correspondiente, lo emprime en el archivo de output.txt y lo devuelve al analizador sintáctico. Note, que el patrón de identificador reconoce a palabra reservadas también. Esto crea ambiguedad semántica debido a que la gramatica no tiene un mecanismo para diferenciar entre una palabra reservada y un identificador. Por esta razón, si

una serie de caracteres se identifica como un lexema de categoría identificador, se compara con los valores ya existente del diccionario *variables*, el cual es un miembro de la clase *driver*. Al inicializar un objeto *driver*, cuyo constructor está localizado en *driver.cc* este se encarga de abrir el archivo de palabras reservadas (*keywords.txt*) y añadir las palabras reservadas antes que cualquier variable se pueda inicializar. Mas aún, a las palabras reservada se le asigna el valor de la cadena vacía. Esto se hace con el propósito de poder diferenciar entre palabras reservadas e identificadores, ya que al nivél sintáctico, no es posible asignarle a una identificador una cadena vacía, como se puede apreciar a continuación:

```
{ID} {
    std::string text(yytext);
    if (drv.variables.find(text) != drv.variables.end() &&
        drv.variables[text] == "") {
        if (drv.variables[text] == "") {
            format_output("PALABRA_RESERVADA", yytext, loc);
            return yy::parser::make_PALABRA_RESERVADA (yytext, loc)
            ;
        }
        format_output("ID", yytext, loc);
        return yy::parser::make_ID(text, loc);
}
```

Con el propósito de visualizar los lexemas generados por el scanner, colocado en

el archivo scanner.ll, se utliza la siguiente función:

En el caso de un error léxico, se ejecuta el siguiente código:

```
. {
  file << "CARACTER INVALIDO " << std::string(yytext) <<
    "," << loc << std::endl;
}</pre>
```

Los patrones que se utilizan para este *scanner.ll* archivo (líneas 30-71) son una adaptación de la gramática el la tabla 2.1. Note las variables *yytext* y *loc* en el código anterior. *loc* contiene la referencia a la memoria de la variable *location* de la clase *driver*, cuyo propósito es retornar el valor de la linea en donde se encuentra un lexema dado. *yytext* contiene el lexema que fue aceptado por un patrón. *loc* y *yytext* se utilizan para imprimir la aceptación de una cadena de caracteres como un lexema de un *token* o para gestión de errores. Por ejemplo, el siguiente código genera una entrada en el archivo output.txt que declar que la cadena de caracteres se aceptó como una sentencia.

```
format_output("SENTENCIAS", yytext, loc);
```

Todas las definiciones de patrones, con la exepción del patrón que maneja la gestión de errores, contienen la función de *format_output* para imprimir la aceptación de una serie de caracteres como un lexema de un patrón en particular. Además, como se ha mencionado anteriormente, una vez se hace la aceptación, se retorna el *token* al analizador sintáctico.

```
{SENTENCIAS} {
   format_output("SENTENCIAS", yytext, loc); \\
   return yy::parser::make_SENTENCIAS(yytext, loc); \\
}
```

Implementación del Analizador Sintáctico

Conclusiones y Recomendaciones

List of Figures

2.1	Automata del patrón para el token <fin_de_linea></fin_de_linea>	12
2.2	Automata del patrón para el token <palabras_reservada></palabras_reservada>	12
2.3	Automata del patrón para el token <id></id>	12
2.4	Automata del patrón para el token <idcont></idcont>	12
2.5	Automata del patrón para el token < ASIGNACION>	13
2.6	Automata del patrón para el token <letra></letra>	13
2.7	Automata del patrón para el token <digito></digito>	13
2.8	Tabla de símbolos implementada como un diccionario	13

List of Tables

2 1	Tabla de Componente	s Léxicos de AVISMO						1 1
Z.1	Tabla de Componeme	S Lexicus de Avisivio	 					LI

Referencias Bibliográficas

- Bison GNU Project Free Software Foundation. (n.d.). https://www.gnu.org/software/bison/.
- Flex a scanner generator. (n.d.). https://ftp.gnu.org/old-gnu/Manuals/flex-2.5.4/html_mono/flex.html.
- Narciso Farias, F., Rios, A., Hidrobo, F., & Vicuña, O. (2012, May). Una gramática libre de contexto para el lenguaje del ambiente de visualización molecular AVISMO..
- Wasti, B. (2020, June). Bwasti/bison-example-calc-.