

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	Introducción	2
2	Analizador Léxico	4
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	Implementación del Analizador Sintáctico	18
4	Conclusiones y Recomendaciones	19
	Referencias Bibliográficas	22

Capítulo 1

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 desarrollo 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 desarrolló utilizando GNU Flex (*Flex - a Scanner Generator*, n.d.) con la implementación de Calc++ por bwasti (Wasti, 2020) como base. La implementación sintáctica se desarrolló

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

Capítulo 2

Analizador Léxico

2.1 Gramatica del Lenguaje AVISMO

- $\langle \text{SENTENCIAS} \rangle ::= \langle \text{FIN_DE_LINEA} \rangle \langle \text{SENTENCIAS} \rangle \mid \langle \text{SENTENCIA} \rangle \langle \text{FIN_DE_LINEA} \rangle$
- $\langle \text{FIN_DE_LINEA} \rangle ::= ":" \mid ";;"$
- $\langle \text{SENTENCIA} \rangle ::= \text{"defina"} \langle \text{ID} \rangle \text{"como"} \langle \text{TIPO} \rangle \mid \langle \text{ID} \rangle \text{"="} \langle \text{MODELO_MOLECULAR} \rangle \mid \langle \text{OPERACION} \rangle \text{"("} \langle \text{ID} \rangle \text{"}"}$
- $\langle \text{ID} \rangle ::= \text{"A"} \mid \text{"B"} \mid \text{"C"} \mid \text{"D"} \mid \text{"E"} \mid \text{"F"} \mid \text{"G"} \mid \text{"H"} \mid \text{"I"} \mid \text{"J"} \mid \text{"K"} \mid \text{"L"} \mid \text{"M"} \mid \text{"N"} \mid \text{"O"} \mid \text{"P"} \mid \text{"Q"} \mid \text{"R"} \mid \text{"S"} \mid \text{"T"} \mid \text{"U"} \mid \text{"V"} \mid \text{"W"} \mid \text{"X"} \mid \text{"Y"} \mid \text{"Z"} \mid \text{"a"} \mid \text{"b"} \mid \text{"c"} \mid \text{"d"} \mid \text{"e"} \mid \text{"f"} \mid \text{"g"} \mid \text{"h"} \mid \text{"i"} \mid \text{"j"} \mid \text{"k"} \mid \text{"l"} \mid \text{"m"} \mid \text{"n"} \mid \text{"o"} \mid \text{"p"} \mid \text{"q"} \mid \text{"r"} \mid \text{"s"} \mid \text{"t"} \mid \text{"u"} \mid \text{"v"} \mid \text{"w"} \mid \text{"x"} \mid \text{"y"} \mid \text{"z"} \mid \langle \text{LETRA} \rangle \langle \text{IDCONT} \rangle$

- <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" | "l" | "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> <ENLACE> | <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_GRUPO_FUNCIONAL>
 ")
- <MODELO_GRUPO_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> <COMPUESTOS>

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.

<i>Token</i>	Patrón	Lexema	Atributos
<FIN_DE_LINEA>	; :	:	Símbolo reservado
<PALABRA _RESERVADA>	defina como	defina	Palabra reservada
<ID>	[A-Za-z][A-Za-z0-9]*	var1	Modelo molecular asociado
<IDCONT>	[A-Za-z0-9]+	1ar	ID asociado
<LETRA>	[A-Za-z]	a	ID asociado
<DIGITO>	[0-9]	7	Valor numérico, lexema asociado
<TIPO>	modelo	modelo	ID asociado
<OPERACION>	graficar2d graficar3d pesomolecular	pesomolecular	ID asociado
<MODELO _MOLECULAR>	({ELEMENTO _QUIMICO} {ELEMENTO _QUIMICO} {VALENCIA} {ELEMENTO} {GRUPO _FUNCIONAL} {ELEMENTO} {GRUPO _FUNCIONAL} {ENLACE} {ELEMENTO} {ENLACE})	CH3(CH3)CHH	ID asociado

<COMPUESTO>	COMPUESTO ({ELEMENTO _QUIMICO} {ELEMENTO _QUIMICO} {VALENCIA}) {ELEMENTO} {GRUPO_FUNCIONAL} {ELEMENTO} {GRUPO_FUNCIONAL} {ENLACE} {ELEMENTO} {ENLACE})	CH3::	Modelo molecular asociado, enlaces, valencias
<COMPUESTOS>	{COMPUESTO}+	CH3:::(OH)3	Modelo molecular asociado, enlaces, valencias
<ELEMENTO>	{ELEMENTO _QUIMICO} {VALENCIA}?	Ag3	Elemento, 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" "Tl" "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")	I	Elemento
<VALENCIA>	[1-9]	2	Valor

<GRUPO _FUNCIONAL>	({GRUPO _FUNCIONAL _INFERIOR} {GRUPO _FUNCIONAL _SUPERIOR} {GRUPO _FUNCIONAL _SUPERIOR} {GRUPO _FUNCIONAL _INFERIOR} "(" {MODELO _GRUPO _FUNCIONAL} ")" "[" MODELO _GRUPO _FUNCIONAL "]")	(CH3){Ag2}	Grupos funcionales, grupo funcional inferior, grupo funcional superior
<GRUPO _FUNCIONAL _INFERIOR>	"[" {MODELO _GRUPO _FUNCIONAL} "]"	[CVHe3]	Elementos, valencias
<GRUPO _FUNCIONAL _SUPERIOR>	"(" {MODELO _GRUPO _FUNCIONAL} ")"	(CVHe3)	Elementos, valencias
<MODELO _GRUPO _FUNCIONAL>	((ELEMENTO _QUIMICO)+ {VALENCIA}?) + ((ELEMENTO)+ {ENLACE} {ELEMENTO})+)	FeH=C3Si4	Elementos, enlaces, valencias
<ENLACE>	("-" "=" ":" "::")	-	Valencia

Tabla 2.1: Tabla de Componentes Léxicos de AVISMO

2.2 Diseño del del Analizador Léxico

2.2.1 Autómatas Finitos Deterministas

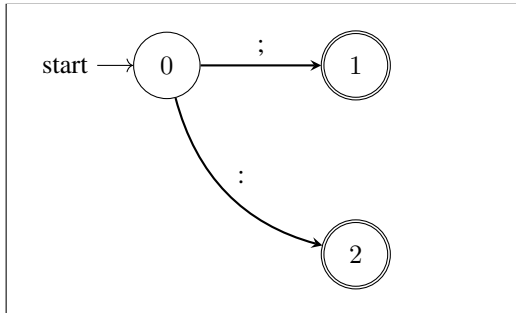


Figura 2.1: Automata del patrón para el token <FIN_DE_LINEA>

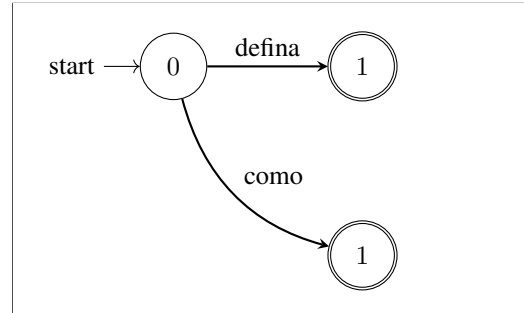


Figura 2.2: Automata del patrón para el token <PALABRAS_RESERVADA>

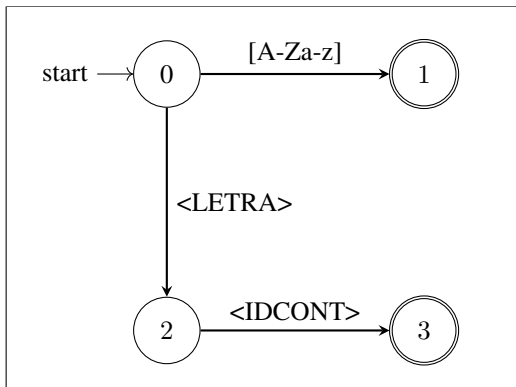


Figura 2.3: Automata del patrón para el token <ID>

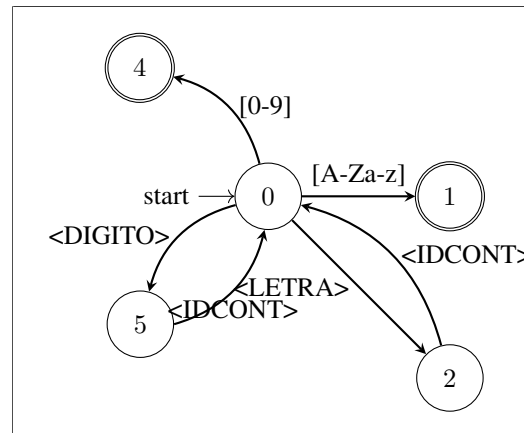


Figura 2.4: Automata del patrón para el token <IDCONT>

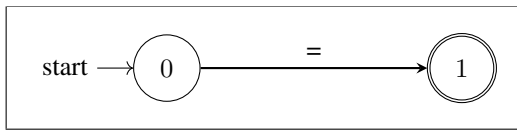


Figura 2.5: Automata del patrón para el token <ASIGNACION>

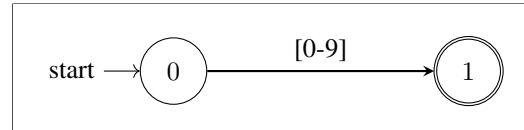


Figura 2.6: Automata del patrón para el token <LETRA>

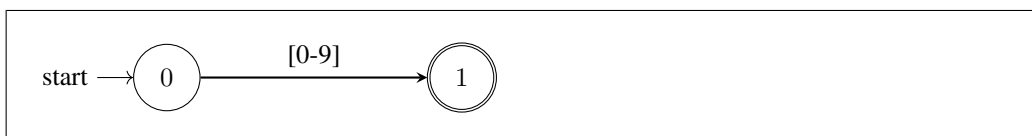


Figura 2.7: Automata del patrón para el token <DIGITO>

2.2.2 Tabla de Símbolos

Identificador

Atributo

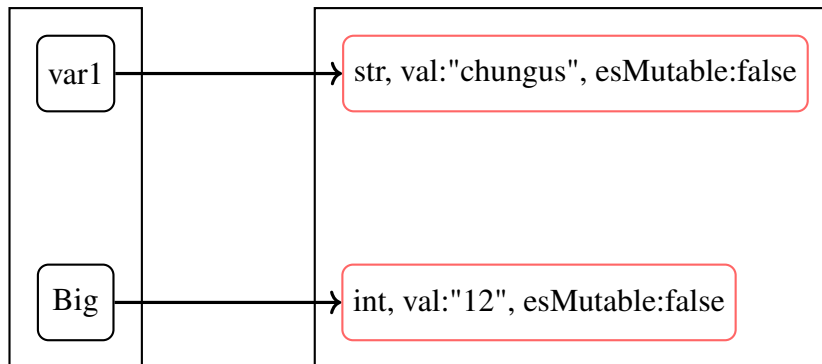


Figura 2.8: Tabla de símbolos implementada como un diccionario

2.3 Implementación del Analizador Léxico

Como se ha mencionado anteriormente, la implementación léxica 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*, esto 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 se encargara de 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 clasica 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 las palabra reservadas. Esto crea ambigüedad semántica debido a que la gramatica no tiene un mecanismo para diferenciar entre una palabra reservada y un identificador. Por esto, si una serie de car-

acteres si identifica como un lexema de categoría identificador, se compara con los valores dentro 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 nivel sintáctico, no es posible asignarle a una identificador una cadena vacía.

Con el propósito de visualizar los lexemas generados por el scanner, colocado en el archivo *scanner.ll*, se utiliza la siguiente función:

```
std::ofstream& file("output.txt");  
void format_output (std::string token,const char* yytext,  
    yy::location& loc) {  
    file << "(" << "<" << token << ">," << std::string(  
        yytext) << "," << loc << ")" << std::endl;  
}
```

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

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

Las definiciones que utilizamos para este archivo *scanner.ll* son las mismas que fueron presentadas en la sección anterior 2.1 Gramatica del Lenguaje avismo pero

es en este archivo que damos la definición de qué cada función ejecutará una vez tenga estos lexemas que siguen el patrón establecido (tokenizados). Se utilizara una variable llamada *loc* que contiene la referencia a memoria de la variable localización de la clase *driver*, el cual usaremos para retornar el valor de la linea en donde encontramos un lexema dado. Tambien esta función devuelve tanto el nombre del token que se categoriza, y su valor como se envia al parser como *yytext* (contiene el valor del lexema); todos estos valores como arguments para especificar el formato que queremos el archivo de salida (*output.txt*):

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

Todas las definiciones (excepto **ID**) contienen la función de *format_output* para la salida de nuestro programa y siempre devuelven una función que se utilizara en el parser (que Flex y Bison se encargan en crear por nosotros) para crear una instancia del token.

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

En el caso de **ID**, hay que hacer una diferenciación entre si es una palabra reservada o si es un identificador, pues si no se pone una restricción el programa no pudiera saber cual caso seria. Para eso se crea un archivo *keywords.txt* el cual contiene estas palabras reservadas los cuales se cargan con el programa *driver*. En el caso de que el valor del lexema que esta en "variables" en la clase de *drv* (*driver*, la cual contiene el lexema que se esta leyendo) resulta ser el mismo valor que

yytext (el cual contiene el lexema retornado del patrón reconocido) y cuyo texto se encuentra en la lista y su valor es vacío (se sabe por cómo se llenan las palabras al driver al momento de ejecutar el constructor), sabemos entonces que cumple con todos los requisitos de ser una palabra reservada en el contexto del programa. De no ser el caso, pues entonces no hay otra opción que ser un identificador.

```
{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);  
}
```

Capítulo 3

Implementación del Analizador Sintáctico

Capítulo 4

Conclusiones y Recomendaciones

List of Figures

2.1	Automata del patrón para el token <FIN_DE_LINEA>	12
2.2	Automata del patrón para el token <PALABRAS_RESERVADA>	12
2.3	Automata del patrón para el token <ID>	12
2.4	Automata del patrón para el token <IDCONT>	12
2.5	Automata del patrón para el token <ASIGNACION>	13
2.6	Automata del patrón para el token <LETRA>	13
2.7	Automata del patrón para el token <DIGITO>	13
2.8	Tabla de símbolos implementada como un diccionario	13

List of Tables

2.1	Tabla de Componentes Léxicos de AVISMO	11
-----	--	----

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-](https://ftp.gnu.org/old-gnu/Manuals/flex-2.5.4/html_mono/flex.html)

[2.5.4/html_mono/flex.html](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-*.