2013

**INFORME DE DISEÑO DE COMPILADORES**

**Primer parte**

**GRUPO N 13**

**INTEGRANTES:**

* **Fernández, Sebastián (**[**sebachk90@gmail.com**](mailto:sebachk90@gmail.com)**)**
* **Gómez Ortiz, Ibrian (**[**ib.gomezo@gmail.com**](mailto:ib.gomezo@gmail.com)**)**
* **Pagano, Diego (**[**diegoipagano@gmail.com**](mailto:diegoipagano@gmail.com)**)**

**TEMAS ASIGNADOS**

El siguiente describirá y explicara la creación de analizador léxico como parte de un trabajo practico propuesto por la catedra de Diseño de Compiladores.  
Al grupo le fue asignado como actividad particular:

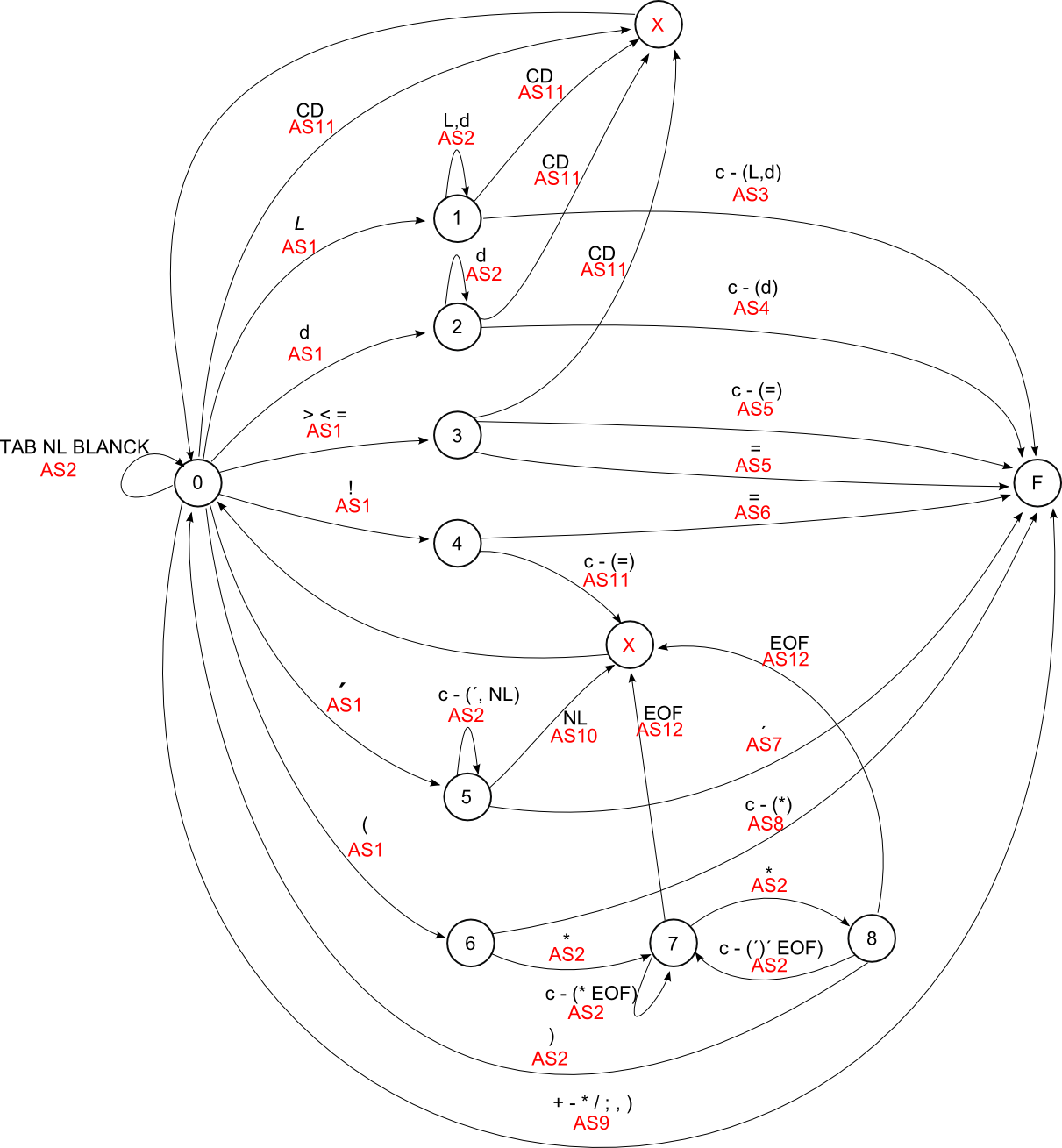
* funciones con pasajes de parámetro por referencia(4)
* Unsigned int
* Iteración de LOOP y UNTIL (1)
* Comentarios mixtilíneas
* Cadenas de una línea

**INTRODUCCION**  
**¿Qué es un analizador léxico?**   
Un analizador léxico (AL) es un componente de un compilador de lenguaje de programación, que se encarga de analizar y encontrar errores léxicos en un código fuente dado. El AL buscará palabras reservadas propias del lenguaje, además de detectar identificadores, constantes, comentarios, etc. A estos elementos se los denominan *Tokens.*

**¿Cómo funciona?**  
El AL funciona en conjunto con otro componente llamado analizador sintáctico (AS). En éste caso el sintáctico se desarrollada utilizando una herramienta ya implementada que se denomina YACC. El AS es quien pone en funcionamiento al procesador léxico solicitándole un token.  
El AL trabaja recorriendo un archivo de texto caracter a caracter procesándolos en un autómata finito. Cada vez que se arribe a un estado final el analizador habrá encontrado un token que deberá ser entregado al procesador sintáctico.  
Es importante aclarar que el AL se pone en funcionamiento solo cada vez que el AS le solicita un token y que una vez encontrado y entregado, se detiene hasta nuevo pedido.  
Cada token está compuesto por un número que lo identifica y un valor de posición que indica en que parte de la tabla de símbolos está almacenado.

**Desarrollo de la aplicación**Como primer medida el grupo comenzó por la creación del autómata que describe los estados por los que atraviesa el AL cada vez que lee un carácter y hasta encontrar un token. (Imagen 1)  
  
L = carácter letra  
d = carácter digito   
TAB = carácter de tabulación  
NL = carácter de Nueva Línea  
EOF = fin de archivo  
BLANK = carácter de espacio  
CD = carácter desconocido  
c – () = cualquier carácter menos los que se encuentran especificados entre paréntesis  
Cualquier otro símbolo o carácter se representa así mismo.

Los escritos en rojos hacen referencia a las Acciones Semánticas que se ejecutan en cada transición de estados.



*Imagen 1 correspondiente al autómata que representa el Analizador Léxico  
Los estados de error (X) son iguales pero por cuestiones de disposición espacial se dibujaron separados*

Matriz de transición de estados

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Carácter** | | | | | | | | | | | | | | | | | | | |
| **Est** | L | d | > | < | = | ! | ‘ | ( | ) | + | - | \* | / | ; | , | NL | TAB | BLANK | EOF | CD |
| 0 | 1 | 2 | 3 | 3 | 3 | 4 | 5 | 6 | F | F | F | F | F | F | F | 0 | 0 | 0 | F | **X** |
| 1 | 1 | 1 | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | **X** |
| 2 | F | 2 | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | **X** |
| 3 | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | **X** |
| 4 | **X** | **X** | **X** | **X** | F | **X** | **X** | **X** | **X** | **X** | **X** | **X** | **X** | **X** | **X** | F | **X** | **X** | F | **X** |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | F | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | **X** | 5 | 5 | F | **X** |
| 6 | F | F | F | F | F | F | F | F | F | F | F | 7 | F | F | F | F | F | F | F | **X** |
| 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 7 | 7 | 7 | 7 | 7 | 7 | **X** | 7 |
| 8 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 0 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | **X** | 7 |
| F |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Los estados marcados con **X** son estados de error.

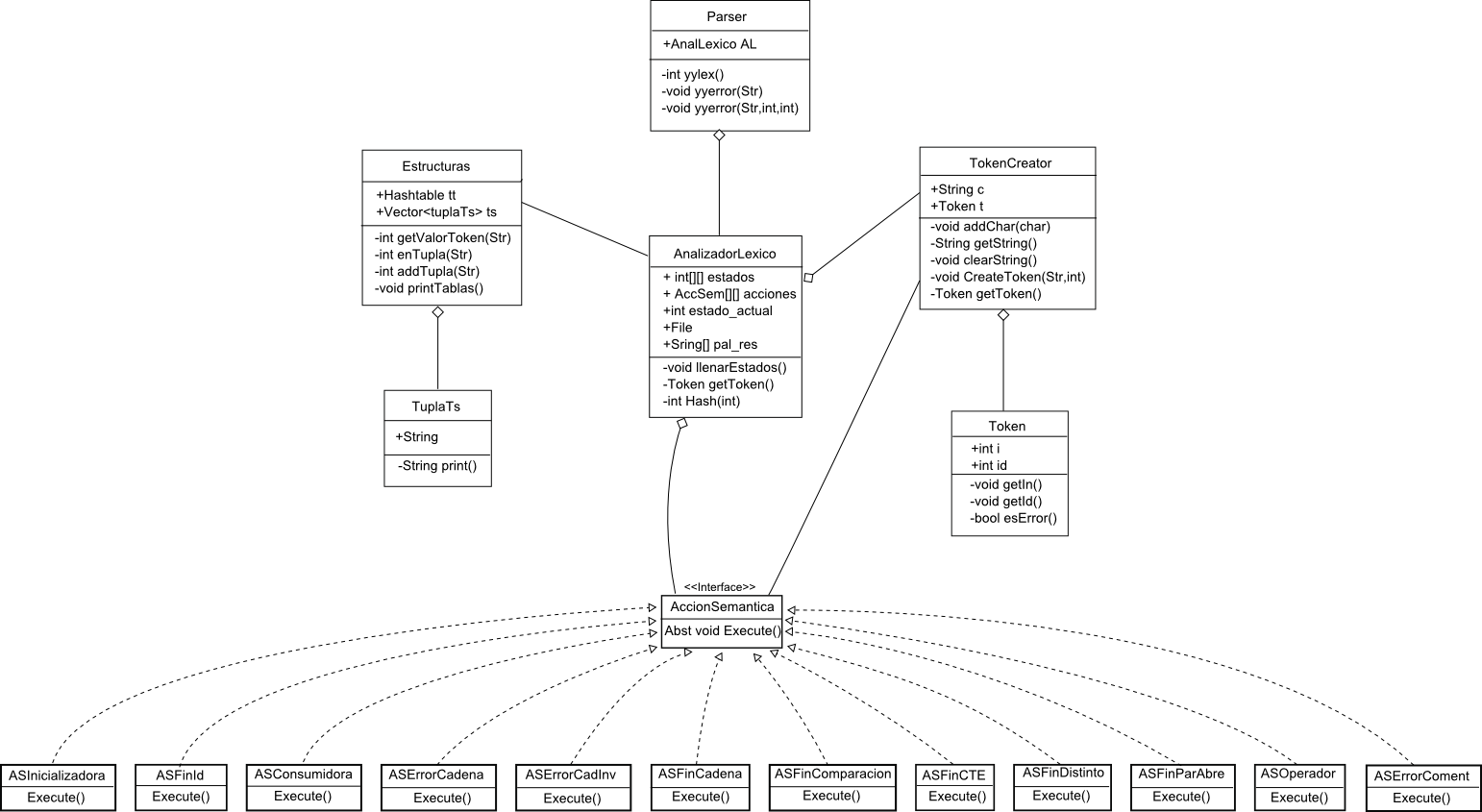
Matriz de Acciones semánticas

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Est** | **Carácter** | | | | | | | | | | | | | | | | | | | |
|  | L | d | > | < | = | ! | ‘ | ( | ) | + | - | \* | / | ; | , | NL | TAB | BLK | EOF | CD |
| 0 | AS 1 | AS 1 | AS 1 | AS 1 | AS 1 | AS 1 | AS 1 | AS 1 | AS 9 | AS 9 | AS 9 | AS 9 | AS 9 | AS 9 | AS 9 | AS 2 | AS 2 | AS 2 | AS 2 | AS 11 |
| 1 | AS 2 | AS 2 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 3 | AS 11 |
| 2 | AS 4 | AS 2 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 4 | AS 11 |
| 3 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 5 | AS 11 |
| 4 | AS 11 | AS 11 | AS 11 | AS 11 | AS 6 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 | AS 11 |
| 5 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 7 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 10 | AS 2 | AS 2 | AS 2 | AS 11 |
| 6 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 2 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 8 | AS 11 |
| 7 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | A S2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 12 | AS2 |
| 8 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 2 | AS 12 | AS2 |
| F |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Una vez delineado el autómata y las matrices que lo representan se comenzó con la implementación. El grupo eligió el lenguaje Java para desarrollarse.  
Se crearon las clases:

* Analizador Léxico: Encargada de recorrer un archivo de texto (código fuente) carácter a carácter y encontrar tokens y devolverlos. Una vez encontrado algún token se detiene hasta nuevo pedido.
* Estructuras: Contiene la tabla de símbolos (implementada como un vector<TuplaTs>) y la tabla de tokens (implementada con una hashtable). Contiene funcionalidades para agregar y consultar entradas en ambas tablas.
* TuplaTs: Implementa la tupla de la tabla de símbolos. Los atributos de ésta pueden variar
* TokenCreator: Se encarga de crear los tokens. Almacena los caracteres que se le van pasando y cuando se le solicita crea un token con los caracteres que venía almacenando
* Token: Clase que representa un token. Está compuesta de un entero que representa el número de token y un puntero a la posición de la tabla de símbolos donde se encuentra almacenado.
* Interface Acción Semántica: Contiene el comportamiento común de toda acción Semántica.
* ASInicializadora (AS1): Lee el primer carácter, marca el archivo en la posición donde lo leyó y lo almacena
* ASConsumidora (AS2): Consume un carácter y lo almacena.
* ASFinId (AS3): Detecta un carácter que no pertenece a la cadena, no lo consume y vuelve hacia atrás una posición en el archivo. Crea un token y además trunca el tamaño del token a 15 caracteres.
* ASFinCTE (AS4): Detecta un carácter que no pertenece a la constante, no lo consume y vuelve hacia atrás una posición en el archivo. Crea un token y controla que el valor de la constante no supere el límite permitido
* ASFinComparacion (AS5): Detecta si el/los caracter/es de comparación está/n bien escrito/s. Se ejecuta siempre después de haber leído un > < o =. Luego si el carácter que viene es un =, puede saber si se trata de un >=, <=, ==. Caso contrario vuelve una posición en el archivo.
* ASFinDistinto (AS6): Detecta el token != (distinto). Si lo encuentra lo crea.
* ASFinCadena (AS7): Detecta la comilla simple final y de esta manera reconoce un fin de cadena. La consume y crea el token correspondiente.
* ASFinParAbre (AS8): Si luego de leer un paréntesis que abre viene cualquier otro carácter que no es un asterisco estaremos entonces en el inicio de un nuevo token. Esta acción semántica retrocede una posición en la lectura del archivo y crea un token con el carácter “(”.
* ASOperador (AS9): Ésta acción detecta los operadores + - \* / ; , ) y crea el token correspondiente.
* ASErrorCadena (AS10): Detecta e informa el error provocado por cadenas que contienen en su interior el carácter de nueva línea.
* ASErrorCarInv (AS11): Detecta inversiones de caracteres. Se ejecuta luego de leer un ! cuando el analizador léxico espera un = y esto no ocurre.
* ASErrorComentario (AS12): Detecta el carácter de fin de archivo (EOF) en medio de un comentario.

La clase Parser y ParseVal surgen de ejecutar la aplicación YACC que recibe como parámetro la gramática del lenguaje creado por el grupo. La gramática se encuentra en el archivo Gramatica.y que se adjunta en ésta entrega.



**Problemas surgidos durante la implementación**  
A lo largo del trabajo el grupo se topó con problemas de distinta índole, siendo en su mayoría los generados por la gramática.  
Durante la generación del Parser, surgió un conflicto reduce/reduce dado que nos enfrentábamos a un problema con la lista de parámetros informada dentro de un llamado a función. Este problema se debe a que la lista de parámetros dentro de una función puede estar vacía, por lo que se debe añadir en la gramática la posibilidad de que esto ocurra. Se solucionó añadiendo a la declaración de la llamada a función, ID '('')', y de esta forma se aceptan tanto parámetros, lista de parámetros o ningún parámetro.  
Otro fue la aparición de un shift/reduce, provocado en las sentencias IF y ELSE dado que llegado a ese punto la gramática no especificaba que camino escoger. Se solucionó dando precedencia a una sentencia sobre la otra.

**Errores léxicos detectados por el compilador**

* Carácter invalido: Se da cuando espera un “=” luego de leer un “!” para intentar crear un token “!=”. Cuando el carácter no es un “=” se muestra el siguiente mensaje:  
  "Linea #: Caracter invalido, se esperaba '=' luego del '!'"
* Constante fuera de rango: cuando una constante posee un valor menor a 0 y mayor a 65535 se muestra el mensaje:  
  "Linea #: El entero se encuentra fuera del rango permitido"
* Cadena mixtilínea: ocurre cuando dentro de una cadena se encuentra el carácter “\n”. Se muestra el mensaje:  
  "Linea #: No se admiten cadenas mixtilínea"
* Error por fin de archivo: Sucede cuando se lee el carácter EOF dentro de un comentario. El mensaje que se muestra es:  
  "Linea # : Error EOF leído en medio de un comentario"
* Error de Carácter desconocido: Se da cuando el Analizador Léxico lee una carácter que no es reconocido por el lenguaje. Se muestra el siguiente mensaje:  
  "Linea #: El Caracter leído “Caracter”no es válido"
* Truncamiento de ID: ocurre cuando el tamaño de un identificador es mayor a 15 caracteres. El mensaje que se muestra es:  
  "Linea #: WARNING: El identificador tiene un nombre muy largo, será truncado"

**Errores sintáctico detectados por el compilador**Error que se produce cuando el parser recibe un token diferente al que está esperando según la gramática.  
Se informa el mensaje:  
“Token leído: # y se esperaba $”

**Lista de no terminales de la gramática**programa: no terminal de entrada a la gramática.  
lista\_declaraciones: Se refiere al listado de declaraciones del programa.  
Declaración: Es una declaración en el programa.  
bloque\_sent: Un bloque de sentencias (ya sean ejecutables o declarativas).  
lista\_sent: Lista de sentencias ejecutables.  
sentencia\_ejec : Sentencia ejecutable que acepta el programa.  
llamada\_funcion: Sentencia ejecutable de llamado a función.  
sentencia\_declarativa\_tipo: Sentencia declarativa de tipo de constante.  
lista\_var: Lista de variables.  
bloque\_IF : Bloque de un IF (ya sea con o sin ELSE).  
cond : Condición de un IF o LOOP-UNTIL.  
comparador : Operador de comparación.  
sentencia\_declar\_funcion: Sentencia declarativa de una función.  
parámetros: Parámetros, o bien de sentencia declarativa de función, o de llamado a una función.  
lista\_parametros : Lista de parámetros.  
parametro : Parámetro expresado como Expresión.  
bloque\_funcion : Lista de sentencias dentro de una función.  
lista\_sent\_declar: Lista de sentencias declarativas.  
asignacion : Asignación de una variable.  
expresion : Expresión matemática de suma o resta.  
termino : Término matemático de producto o división.  
factor : Reductor de constante o identificador.  
tipo : Tipo de variable.

**Tratamientos de errores**Para el tratamiento de errores, el grupo optó primero por el modo pánico. Esto quiere decir que para las sentencias simples, cuando se encuentra un error, se buscará el token ';' y se recuperará a partir de allí.  
Pero existen otras situaciones de posibles errores, donde el token ‘;’ no es útil para recuperarse rápidamente, por ejemplo:

*if(a<<b) then  
 begin   
 sentencia1;  
 sentencia2;  
end*

En éste caso va a recuperarse recién luego de la sentencia 1, lo que implica que también va a ignorar el *Begin*, por lo tanto va a encontrar otro error cuando lea *end.*

Luego de eso, el grupo empezó a buscar la manera de tratar los errores más complejos, empezando por la falta del ';'. Cuando falta este token al final de una sentencia, el analizador llega a un estado de error, pero no puede salvarse buscando el token ';' porque es el que falta, por lo tanto ignora todo lo que sigue luego del error hasta encontrarlo y esto genera varios problemas:

1. si hay error antes de la *sentencia2;*, ignora la *sentencia2*, y si ésta tenia error, no lo detecta
2. si se da justo antes de un *if* o un *loop* sucede lo mismo que antes , ignora todo el *if* o *loop*
3. si se da al final de un bloque de sentencias, ignora el end.

Básicamente trata como un error grande todo desde el momento que encuentra el error hasta el próximo ‘;’.

Fue ahí cuando nos inclinamos por trabajar analizando todos los casos de error por separado.  
Se pensó en separar las sentencias bien escritas de las mal escritas, y dentro de las que se encontraban con error especificar los diferentes casos.  
Por ejemplo en una sentencia *IF (if (condición) then <bloqueIF>)* se puede dar:  
 1) que falte la condición o que esté mal escrita, falta de algún paréntesis

2) que falte el *then*

En un bucle *loop (loop <bloqueLOOP> until (condicion)):*

1. que falte el *Until*
2. algún error en la condición

Por el lado de las sentencias, una correcta es o bien una incompleta con el ‘;’ o una con error y un ‘;’, eso es para evitar errores shift/reduce.

**Conclusión**  
El trabajo fue de mucha utilidad para la comprensión del funcionamiento de un compilador. Tanto la creación del Analizador léxico como el del Sintáctico fueron una experiencia clara que ayudaron al entendimiento del problema y comprensión del tratamiento de errores.

**2da Parte  
Código intermedio y Assembler**

**Introducción**

Para la segunda parte, se asigno como tarea el desarrollo de las partes faltantes del compilador, referidas a la generación de código. Para la misma se pidió desarrollar una generación de código intermedio y a continuación desarrollar el generador de código Assembler MASM.

**Temas Asignados**

Al grupo se le asignaron los siguientes puntos específicos:  
\*Polaca inversa para el código intermedio  
\*Pasaje de parámetros por referencia.

**Código Intermedio**

La polaca inversa fue implementada con un arreglo, la misma contiene dentro, operandos (referencias a la tabla de símbolos), operadores, etiquetas (de funciones y de sentencias de control) y referencias a posiciones de la propia polaca.  
Para llenar la polaca se tuvieron en cuenta ciertas situaciones:  
\*operadores y operandos: se insertan directamente en la polaca, estas inserciones se realizan siempre al final, por lo tanto desde la gramática se tuvo especial cuidado para que las inserciones se hagan en el orden adecuado 🡪 a+b, primero se inserta ‘a’ luego ‘b’ luego ‘+’.  
Dentro de nuestra gramática tenemos la regla “factor: ID|CTE;”, es aquí cuando agregamos a la polaca el token leído ($1.ival, su posición en la tabla de símbolos).  
Gracias a la recursividad a izquierda, siempre vamos a encontrar que el primer operando llega a la regla mencionada antes que el segundo. Luego en las reglas referidas a operaciones, ya sea en operadores lógicos como aritméticos, se insertan en la polaca los operadores correspondientes. (\*1)  
\*Precedencia: gracias a que se tuvo en cuenta la precedencia de operadores desde un principio, no se necesito tomar recaudos para que se refleje correctamente la precedencia en la polaca.  
\*Asignación: “ID ‘=’ expresión”, llegados a este punto, la expresión esta correctamente insertada en la polaca, faltan insertar los token ID e ‘=’, ahora se debe insertar primero el ID y luego el ‘=’, por lo tanto insertamos, $1.ival e “=” en ese orden. (2\*)

(\*1) : en la gramática se cambio la regla condición, antes era “cond: expresión comparador expresión” donde luego la regla del comparador era “comparador: <|>|<=|>=|==|!=. Esto se cambió y reemplazó en condición para poder insertar correctamente en la polaca (ver gramática actual)  
(\*2) la asignación es un caso especial de la polaca, dado que para a=b por ejemplo, la polaca resultante es b|a|=, por lo tanto se debe tener en cuenta al hacer el código assembler.

\*Funciones: en la declaración de funciones, se agrega a la polaca la etiqueta de la misma (en la regla se corresponde con $2.sval), además se almacena la posición de la misma en una estructura que será luego accedida para conocer la posición de la misma.

\*Call, Return y fin de funciones: para el llamado de funciones, dado el ID ($2.sval) se busca en la estructura antes mencionada la posición del arreglo donde se encuentra el label, y se agregan este número y el operador de CALL. En los return simplemente se añade el operador RETURN, que indicara el fin de la función. A su vez se añade un operador return al final de la función, haya otro antes o no. Esta decisión se basó en que la lógica necesitada para evitar el doble Return era compleja, y es inaceptable que no haya ninguno.

\*