Forma

Descripción generada automáticamente con confianza media

**ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA**

**DOBLE GRADO EN INGENIERÍA INFORMÁTICA E INGENIERÍA DEL SOFTWARE**

**Curso Académico 2023/2024**

**Procesadores de Lenguajes**

**8 de abril de 2024**

**Práctica Obligatoria PL**

**Equipo 6**:

Ángel Marqués García

Tarek Elshami Ahmed

Álvaro Serrano

# Índice

[Índice I](#_Toc162089216)

[1. Introducción 2](#_Toc162089217)

[2. Parte Aprobado 3](#_Toc162089218)

[2.1 Reducción de la gramática a LL(1) 3](#_Toc162089219)

[2.1.1 Eliminar recursividad por la izquierda 3](#_Toc162089220)

[2.1.2 Factorizar 3](#_Toc162089221)

[2.1.3 Generar los conjuntos de CAB, SIG 4](#_Toc162089222)

[2.2 Generar la gramática de ANTLR 5](#_Toc162089223)

[3. Parte Notable 6](#_Toc162089224)

[3.1 Sentencias de control de flujo 6](#_Toc162089225)

[3.2 Reconocimiento y notificación de errores 6](#_Toc162089226)

[4. Parte sobresaliente 6](#_Toc162089227)

[5. Probar el correcto funcionamiento del reconocedor 6](#_Toc162089228)

[5.1 Correctos 7](#_Toc162089229)

[5.1.1 Caso 1 7](#_Toc162089230)

[5.1.2 Caso 2 7](#_Toc162089231)

[5.1.3 Caso 3 7](#_Toc162089232)

[5.1.4 Caso 4 7](#_Toc162089233)

[5.2 Erróneos 7](#_Toc162089234)

[5.2.1 Caso 1 7](#_Toc162089235)

[5.2.2 Caso 2 7](#_Toc162089236)

[5.2.3 Caso 3 7](#_Toc162089237)

[5.2.4 Caso 4 7](#_Toc162089238)

# Introducción

Esta memoria redacta el proceso seguido por el **grupo 6** en la realización de la práctica obligatoria de Procesadores de Lenguajes. El objetivo de esta práctica es diseñar e implementar un programa capaz de traducir código de un lenguaje fuente a un lenguaje final. El lenguaje fuente será similar a C, mientras que el lenguaje final será similar a PASCAL.

La práctica consta de tres sets de requisitos diferenciados, completar cada uno permite optar a una mejor calificación:

1. Requisitos Aprobado (parte obligatoria)
2. Requisitos Notable (sentencias de control de flujo y notificación de errores)
3. Requisitos Sobresaliente (recuperación de errores y toda la parte opcional)

A la misma vez, la práctica consta de dos entregas. La primera consta de los analizadores léxico y sintáctico, y la segunda de la práctica completa.

Por tanto, el primer paso que tomamos en el desarrollo de la práctica ha sido desarrollar los analizadores léxico y sintáctico. Este proceso consta de 3 fases principales:

1. Reducir la gramática a LL (1)
2. Trasladar la gramática a ANTLR
3. Probar su corrección y completitud

Y de nuevo este proceso se ha dividido entre los distintos sets de requisitos, de forma que primero ha sido aplicado a la parte de aprobado, y luego a las partes de notable y sobresaliente.

Las siguientes secciones detallan, primero, el proceso seguido para reducir la gramática a LL(1), seguido de una explicación de los casos de prueba que se han utilizado para comprobar que el traductor funciona correctamente (estos casos de prueba se incluirán en los anexos). Por último, se describirá cómo se ha implementado la notificación y recuperación de errores.

# Parte Aprobado

## Reducción de la gramática a LL(1)

La gramática del lenguaje fuente es la siguiente:

program ::= defines partes

defines ::= ʎ | "#define" IDENT ctes defines

ctes ::= CONSTINT | CONSTFLOAT | CONSTLIT

partes ::= part partes | part

part ::= type restpart

restpart ::= IDENT "(" listparam ")" blq

| IDENT "(" "void" ")" blq

blq ::= "{" sentlist "}"

listparam ::= listparam "," type IDENT | type IDENT

type ::= "void" | "int" | "float"

sentlist ::= sentlist sent | sent

sent ::= type lid ";" | IDENT "=" exp ";" | IDENT "(" lexp ")" ";"

| IDENT "(" ")" ";" | "return" exp ";"

lid ::= IDENT | lid "," IDENT

lexp ::= exp | lexp "," exp

exp ::= exp op exp | factor

op ::= "+" | "-" | "\*" | "/" | "%"

factor ::= IDENT "(" lexp ")" | IDENT "(" ")"

| "(" exp ")" | IDENT | ctes

Para reducirla a LL(1) hemos seguido una serie de pasos:

### Eliminar recursividad por la izquierda

Hemos identificado que las siguientes reglas tenían recursividad por la izquierda:

* listparam
* sentlist
* lid
* lexp
* exp

Y hemos eliminado la recursividad por la izquierda modificando estas reglas y añadiendo nuevas, de forma que ahora son de la siguiente manera:

listparam: type IDENT listparam2;  
listparam2: ',' type IDENT listparam2 | ʎ;

sentlist: sent sentlist2;  
sentlist2: sent sentlist2| ʎ;  
lid: IDENT lid2;  
lid2: ',' IDENT lid2 | ʎ;  
lexp: exp lexp2;  
lexp2: ',' exp lexp2 | ʎ;  
exp: factor exp2;  
exp2: op exp exp2| ʎ;

Una vez la gramática no tiene recursividad por la izquierda, el siguiente paso es:

### Factorizar

Hemos identificado que las siguientes reglas podían ser factorizadas:

* restpart
* sent
* factor
* partes

Y las hemos factorizado modificándolas y añadiendo nuevas reglas, de forma que ahora son de la siguiente manera:

restpart: IDENT '(' restpart2;  
restpart2: listparam ')' blq | 'void' ')' blq;  
sent: type lid ';'  
 | 'return' exp ';'  
 | IDENT sent2  
sent2: '=' exp ';'| '(' sent3;  
sent3: ')'';'| lexp ')'';';

factor: '(' exp ')' | ctes | IDENT factor2;  
factor2: '(' factor3 | ;  
factor3: lexp ')' | ')';

partes: part partes2;  
partes2: partes | ;

Ahora que la gramática no tiene recursividad por la izquierda ni contiene reglas que puedan ser factorizadas a simple vista, el siguiente paso es:

### Generar los conjuntos de CAB, SIG

Para generar estos conjuntos utilizado la herramienta JFLAP. Dado que JFLAP sólo permite usar reglas que consten de una letra, hemos tenido que mapear cada regla a una letra del alfabeto. Así, la representación de las reglas en JFLAP es la siguiente:

A -> program  
B -> defines  
C -> ctes

D -> partes  
Ñ -> partes2  
E -> part  
F -> blq

G -> type  
H -> op  
I -> listparam  
J -> listparam2  
K -> sentlist  
L -> sentlist2  
M -> lid  
N -> lid2  
O -> lexp

P -> lexp2  
Q -> exp  
R -> exp2  
S -> factor  
T -> factor2  
U -> factor3  
V -> restpart  
W -> restpart2  
X -> sent   
Y -> sent2  
Z -> sent3

Y los símbolos terminales han sido codificados de la siguiente manera:

a -> '#defines'  
b -> IDENT  
c -> CONSTINT  
d -> CONSTFLOAT  
e -> CONSTLIT  
f -> '{'  
g -> '}'  
h -> 'void'  
i -> 'int'  
j -> 'float'  
k -> '+'  
l -> '-'  
m -> '\*'  
n -> '/'  
o -> '%'  
p -> ','  
q -> '('  
r -> ')'  
s -> ';'

t -> 'return'

Con esta traducción, JFLAP genera los siguientes conjuntos de Cabecera y Siguiente:  
Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Junto con la siguiente tabla de directores:

Imagen que contiene mucho, cubierto, llenado, luz

Descripción generada automáticamente

Una vez completado este paso, podemos proceder a generar la gramática en ANTLR.

## Generar la gramática de ANTLR

Para generar la gramática en ANTLR hemos copiado las reglas de la gramática final a ANTLR, ajustando el formato para que sea el adecuado, y creando además los fragmentos léxicos necesarios para el correcto funcionamiento del reconocedor. Se han creado los siguientes fragmentos léxicos:

IDENT: (LETTER | '$') (LETTER |DIGIT|'\_'| '$')\*;  
  
CONSTINT: DECIMAL\_INT | OCTAL\_INT | HEXA\_INT;  
CONSTFLOAT: DECIMAL\_REAL | OCTAL\_REAL | HEXA\_REAL;  
  
DECIMAL\_INT: ('+' | '-')? DIGIT+;  
DECIMAL\_REAL: ('+' | '-')? DIGIT+ '.' DIGIT+;  
  
OCTAL\_INT: '0' ('+' | '-')? OCT\_DIGIT+;  
OCTAL\_REAL: '0' ('+' | '-')? OCT\_DIGIT+ '.' OCT\_DIGIT+;

HEXA\_INT: '0x' ('+' | '-')? HEX\_DIGIT+;  
HEXA\_REAL: '0x' ('+' | '-')? HEX\_DIGIT+ '.' HEX\_DIGIT+;  
  
CONSTLIT: '\'' (ESC\_SEQ | ~['])\* '\'' ;  
  
GPC: '//' ~[\r\n]\* -> skip ;  
GPC2: '/\*' .\*? '\*/' -> skip;  
  
JUMP: ('\r'? '\n' | ' ')+ -> skip;  
  
fragment ESC\_SEQ: '\\' '\'';  
fragment LETTER: [a-zA-Z];  
fragment HEX\_DIGIT: [0-9A-F];

fragment OCT\_DIGIT: [0-7];  
fragment DIGIT: [0-9];

# Parte Notable

Los requisitos de esta parte incluyen reconocer sentencias de control de flujo, y reconocer y notificar errores. Lo primero que hemos hecho ha sido modificar la gramática para añadir reconocimiento de las sentencias de control de flujo. Para ello hemos partido de una gramática de estas sentencias, que hemos reducido a LL(1), e incorporado al resto de la gramática.

## Sentencias de control de flujo

Partimos de las siguientes reglas:

lcond ::= lcond opl lcond | cond | "!" cond

opl ::= "||" | "&&"

cond ::= exp opr exp

opr "==" | "<" | ">" | ">=" | "<="

sent ::= …

| "if" "(" lcond ")" blq "else" blq

| "while" "(" lcond ")" blq

| "do" blq "until" "(" lcond ")"

| "for" "(" IDENT "=" exp ";" lcond ";" IDENT "=" exp ")" blq

Identificamos que la regla lcond tiene recursividad por la izquierda, y la eliminamos, obteniendo:

lcond ::= cond lcond2 | "!" cond lcond2;  
lcond2 ::= opl lcond lcond2 | ʎ;

Identificamos que ninguna regla puede ser factorizada a simple vista, y procedemos a generar los conjuntos de cabecera, siguiente, y directores. Cabe destacar que, al estar modificando la regla sent, será necesario recalcular los conjuntos de todas las reglas, pues al modificar la gramática podrían haber cambiado.

PONER LOS CONJUNTOS ACTUALIZADOS SI HACEN FALTA

## Reconocimiento y notificación de errores

Para poder notificar los errores que surjan de una manera más concreta/útil/concisa, hemos sustituido el ErrorListener que proporciona ANTLR por defecto por un ErrorListener personalizado. Para ello, el primer paso ha sido crear este CustomErrorListener. Para que sea reconocido como un ErrorListener y para que no sea necesario implementar cada método de cero, deberá extender la clase BaseErrorListener. Luego, se sobreescribe el método syntaxError, implementando en su código lo que debe ocurrir cuando se encuentre un error. Para no tener que crear una clase para errores sintácticos y otra para léxicos, reutilizaremos la misma clase, ayudándonos de un atributo type, que recibe su valor en el constructor, y valdrá 0 para el analizador léxico, y 1 para el sintáctico. Luego, el mensaje de error indicará, en función de este atributo, si el error era léxico o sintáctico, y en caso de ser sintáctico, añadirá un mensaje mostrando la línea en la que haya ocurrido el error con el token que lo ha causado subrayado.

Una vez creada esta clase, sólo queda utilizarlo como ErrorListener en los analizadores léxico y sintáctico. Para ello lo primero es borrar los *ErrorListeners* por defecto de ambos reconocedores.



Una vez eliminado el listener por defecto, le añadimos el listener personalizado:



Una vez hecho esto, cuando haya un error, se mostrarán mensajes de error como este:



Cabe destacar que el mensaje de error de esta imagen es el que aparecerá cuando el error sea sintáctico. Si ocurre un error léxico, sólo se verá la primera línea, anunciando la presencia de un error léxico junto con su posición y posible causa.

# Parte sobresaliente

La parte sobresaliente tiene tres requisitos principales: haber completado la parte notable, haber completado toda la parte opcional, y haber proporcionado recuperación de errores léxica y sintáctica. La parte notable ya se ha explicado en su propia sección, así que esta la dedicaremos a explicar cómo se han implementado los elementos restantes de la parte opcional y la recuperación de errores. Quedaban dos partes por implementar de la sección opcional: distinguir entre librerías y programas, y mejorar la generación de declaraciones. El método que hemos seguido para realizar estas tareas está estrechamente enlazado a el procedimiento que hemos seguido para la traducción, por lo que también incluiremos en esta sección una explicación de este proceso.

## Traducción

El enfoque utilizado para la traducción es el siguiente:

1. Hemos estudiado ambas gramáticas para familiarizarnos con ellas y entenderlas a fondo
2. Identificar los principales elementos del lenguaje final, ayudándose para ello de la gramática proporcionada. Estos elementos son cosas como asignaciones, declaraciones de variables, llamadas a procedimientos/funciones, tipos, expresiones, etc
3. Para cada elemento, crear una clase. Se creará una nueva instancia de esta clase cada vez que se identifique que hay un elemento de este tipo, y esa instancia guardará la información relevante. Por ejemplo, para una asignación, al encontrar “someVar = 2;”, crearemos una instancia de asignación, que almacenará el nombre de la variable y su valor.
4. Para cada clase, hacer que implemente una interfaz translation, que consta en un método getTranslation() que devuelve una string. Este método es el responsable de, partiendo de la información almacenada en la clase, formar la string que corresponde al elemento en la gramática final. Volviendo al ejemplo anterior, el método getTranslation de la clase asignación revolvería algo similar a “return variable + ‘:=’ + exp.getTranslation()’, asumiendo que el nombre se almacena en una propiedad tipo String llamada variable, y su valor en una variable de tipo <? extends Translation> llamada exp. En nuestro caso, la clase del valor es de tipo Expression.
5. Al haber reconocido la gramática completa, hacer System.out.println(program.getTranslation()). Program es la clase que almacena toda la información del programa. Al llamar a su getTranslation, este irá llamando de manera recursiva al resto de elementos para obtener su traducción, resultando al final en una string que contiene la traducción completa.

La imagen siguiente muestra un diagrama de clases ilustrando las clases creadas y cómo estas están relacionadas entre sí.

Interfaz de usuario gráfica

Descripción generada automáticamente

### Utilización de interfaces

Durante el proceso de diseñar la traducción, observamos que la gramática final tiene una serie de listas heterogéneas. Por ejemplo, un bloque consiste en una lista de declaraciones y una lista de sentencias. Pero una sentencia puede ser de varios tipos: puede ser una asignación, una llamada a otra función, un return, e incluso, incluyendo la parte opcional, algún tipo de condicional o bucle. Algo similar ocurre con las declaraciones. En casos como estos, para poder almacenar en una misma lista estos objetos distintos hemos creado interfaces. De forma que el objeto Blq tiene una lista de objetos que implementan Declaration y otra de objetos que implementan Sent. Esto ha sido esencial para permitir almacenar los valores de manera sencilla, y ha facilitado mucho el proceso.

## Distinguir entre librerías y programas

Para distinguir entre librerías y programas, la clase Program tenía un atributo booleano hasMain. Cada vez que se reconocía una declaración de una función, se comprobaba su nombre. En caso de que fuera “main”, se le asignaba true a hasMain, que por defecto vale false. Cuando se llame program.getTranslation(), si hasMain vale true, se traducirá con el formato de un programa, y en caso contrario se traducirá como una unidad.

## Mejora de la generación de declaraciones

Este punto consistía en identificar casos en los que se declaran, en dos líneas consecutivas, declaraciones de variables del mismo tipo, para, al traducir, poder declararlas todas en una misma sentencia. De igual manera, era necesario identificar casos en los que, al declarar una función, se declaraban parámetros del mismo tipo de manera consecutiva, con el mismo fin de agruparlos en la traducción.

Para conseguir esto hemos modificado los métodos getTranslation de Blq, que contiene la lista de declaraciones, y de MethodDcl, que es la clase que contiene la sentencia de declaración de un método, y por tanto, sus parámetros. Sin embargo, hemos usado métodos ligeramente distintos en cada uno.

La traducción de Blq consta de dos partes, la traducción de la lista de declaraciones y la traducción de la lista de sentencias. Para esta sección sólo nos interesa la primera parte. Esta tiene un funcionamiento algo peculiar. Dado que en el lenguaje final las declaraciones de constantes y variables van agrupadas, creamos tres partes de la traducción: constants, vars, y other. Mientras iteramos por la lista de declaraciones, comprobamos sus tipos. Si la declaración es una VarDcl, y su isConstant devuelve true, añadimos su traducción a constants. Si es VarDcl y isConstant devuelve false, su traducción se añade a vars, y en otro caso, se añade a other. Para implementar esta sección, sólo es necesario modificar la sección en la cual la declaración es una instancia de VarDcl, y su isConstant devuelve false.

Se ha modificado siguiendo la siguiente lógica:

* Si esta variable es la primera declaración, o la primera variable de su tipo, se añade una tabulación a la traducción.
* Si esta variable no es la última declaración, y la siguiente declaración en la lista es una variable del mismo tipo que la variable actual, se añade a la traducción el nombre de la variable seguido por una coma.
* En caso contrario, se añade la traducción de la variable devuelta por getTranslation().

En lo referente a las listas de parámetros, antes de implementar esta mejora, al traducir la declaración de un método, al llegar a sus parámetros simplemente iterábamos por la lista de parámetros del método llamando a getTranslation sobre cada parámetro, e incluyendo manualmente comas entre ellos. Para incluir esta mejora, hemos modificado notablemente este proceso. Se sigue el siguiente algoritmo:

* Crear una variable lastType, inicializada a “”. Esta variable guarda el tipo de la variable en la posición i-1.
* Añadir a la traducción el nombre del parámetro actual, seguido por una coma, y actualizar lastType
* Al principio del bucle, si lastType no coincide con el tipo del parámetro actual, eliminar la coma introducida en el paso anterior, y escribir “: ” seguido de lastType.
* Si el parámetro actual es el último, realizar el mismo proceso de arriba (eliminar coma, escribir : lastType)

## Recuperación de errores

No hemos considerado necesario modificar la recuperación de errores por defecto de ANTLR, ya que cubre todos los errores comunes, y dada la complejidad de las estrategias utilizadas, no creemos que podamos aportar una implementación mejor o más completa.

# Probar el correcto funcionamiento del reconocedor

Por último, hemos realizado una serie de pruebas para comprobar que el reconocedor funciona de la forma esperada. Se ha probado su funcionamiento con 8 casos de prueba en total, cuatro de los cuales son correctos, y el resto erróneos. Los casos de prueba son los siguientes:

## Correctos

### Caso 1

En este primer caso tenemos una función sola con un “for” correctamente definido:

int sum(int a, int b) {  
 int result;  
 int i;  
 result = 0;  
  
 for (i = a; i < b; i = i + 1) {  
 result = result + i;  
 }  
 return result;  
}

Tras ejecutar el programa, la traducción mostrará el programa en pascal con las dos variables con el mismo tipo “result” y “i” en la misma línea.

### Caso 2

En este caso, hemos cambiado un poco el caso 1 de tal forma que el “for” ya no sea el caso más sencillo posible:

int sum(int a, int b) {  
 int result;  
 int i;  
 result = 0;  
  
 for (i = a; i <= b; i = i + 2) {  
 result = result + i;  
 }  
 return result;  
}

El resultado de traducir este código es uno muy parecido al del caso 1, solo que en este el “for” se ha convertido en un “while-do”.

### Caso 3

En este caso tenemos un código con una función “main” y varias constantes:

#define decimalFloatConst +9.32  
#define octalFloatConst 0-3.06  
#define hexFloatConst 0x+A9F.7D  
  
void main(void voidParam, int intParam, float floatParam) {  
 void voidVar1, voidVar2;  
  
 someVar=((puedesAnidarTantosComoQuieras));  
 someOtherVar=32;  
 yetAnotherVar=someFunct(3, someVar, anotherFunction());  
 lastVar= 32 + times(2,3);  
  
 int intVar1, intVar2;  
  
 someFunction();  
  
 return value(var1,var2,2) / 3;  
  
 unreachableCall();  
 float unreachableVar;  
 unreachableAssignment = 0x9.58;  
}

### Caso 4

Y por último, un ejemplo más grande con un poco de todo:

#define decimalFloatConst +9.32  
#define octalFloatConst 0-3.06  
#define hexFloatConst 0x+A9F.7D  
  
int functionName(void){  
 if (0+7 < var1 || !var2==f() && 0x9A>(-10\*0+5)){  
 noPermiteBloquesVacios();  
 } else {  
 noPermiteBloquesVacios();  
 }  
  
 while ((2 + 2)==$\_5) { //$\_5 is a variable  
 noPermiteBloquesVacios();  
 }  
  
 do {  
 int i;  
 int j;  
 i = 0;  
 j = i;  
 } until (f()>+3)  
  
 for (var0 = -07; var0>f(0xFF); var0 = var0 + 1) {  
 noPermiteBloquesVacios();  
 }  
  
}  
  
void main(void voidParam, int intParam, float floatParam) {  
 void voidVar1, voidVar2;  
  
 someVar=((puedesAnidarTantosComoQuieras));  
 someOtherVar=32;  
 yetAnotherVar=someFunct(3, someVar, anotherFunction());  
 lastVar= 32 + times(2,3);  
  
 int intVar1, intVar2;  
  
 someFunction();  
  
 return value(var1,var2,2) / 3;  
  
 unreachableCall();  
 float unreachableVar;  
 unreachableAssignment = 0x9.58;  
}

## Erróneos

### Caso 1

Como primer caso de “input” erróneo tenemos una función en la que al “if” le faltan corchetes y la variable “y” se declara y es asignada un valor en la misma línea:

void example() {  
 if (x > 0)  
 int y = 2;  
 else  
 y = 3;  
}

### Caso 2

Es este segundo caso de error, el problema esta en que nuestra gramática no reconoce “string” como un tipo valido:

string getName() {  
 return ANTLR4;  
}

### Caso 3

En este caso 3, hemos añadido a un “for” más argumentos de los que nuestra gramática permite:

void loopExample() {  
 int i;  
 for (i = 0, j = 10; i < j; i = i + 1 , j = j - 1) {  
 i = j + hola  
 }  
}

### Caso 4

Y por último, tenemos un código con la función “switch” que nuestra gramática no soporta:

void checkGrade(int score) {  
 switch (score) {  
 case 90 ... 100:  
 hola(1);  
 break;  
 case 80 ... 89:  
 adios();  
 break;  
 default:  
 cosas = 1 + 1  
 }  
}