Jonathan Salisbury Vega

Joan Sansó Pericàs

Julián Wallis Medina

Imagen que contiene Forma

Descripción generada automáticamente

Curso 2021-2022

Práctica Final

Compiladores - 21780

Índice

[Analizador Léxico 2](#_Toc93874573)

[Tokens 2](#_Toc93874574)

[Generación del Escáner 3](#_Toc93874575)

[Restricciones y Peculiaridades 3](#_Toc93874576)

[Analizador Sintáctico 4](#_Toc93874577)

[Gramática 4](#_Toc93874578)

[Generación del Analizador 7](#_Toc93874579)

[Restricciones y Peculiaridades 8](#_Toc93874580)

[Analizador Semántico 8](#_Toc93874581)

[Tabla de Símbolos 8](#_Toc93874582)

[Recorrido del Árbol Sintáctico 9](#_Toc93874583)

[Generación de Código 9](#_Toc93874584)

[Referencias 10](#_Toc93874585)

# Analizador Léxico

## Tokens

Se han definido todos los *tokens* necesarios para la gramática del lenguaje de nuestro compilador, a continuación, se encuentra la lista completa:

* *Keywords*:

|  |  |
| --- | --- |
| * Int * Bool * Const * If * Else * While * Crotofunc * Break | * Return * And * Or * Not * True * False * Main |

En primer lugar, están las palabras reservadas por el lenguaje para funciones internas, estas palabras no podrán ser utilizadas como identificadores por el usuario ya que el escáner las detectará como *tokens* especiales.

* *Operators*:

|  |  |
| --- | --- |
| * = * + * - * \* * / * == | * != * > * < * >= * >= |

En segundo lugar, encontramos los distintos operadores del lenguaje, estos incluyen los operadores aritméticos y de comparación.

* *Separators*:

|  |  |
| --- | --- |
| * ( * ) * { | * } * ; * , |

Por último, tenemos los distintos separadores del lenguaje, estos tokens se utilizan para distinguir diferentes partes del código fuente como, por ejemplo, funciones, bloques de código o finales de línea.

Cabe destacar que también están definidos tokens para poder representar los valores literales de los distintos tipos de datos del lenguaje.

## Generación del Escáner

Para el apartado del analizador léxico del compilador se ha utilizado la herramienta de generación de escáner *JFlex* [1]. Como se puede observar en el código del generador localizado en */jflex/Lexer.flex* existen tres grandes apartados para la generación del escáner que explicaremos a continuación.

* **Código de Usuario**:

En este apartado inicial se puede encontrar código fuente de *Java* que se utilizara posteriormente en el archivo escáner generado. El código está formado en su totalidad de *imports* de clases para la creación de los símbolos correspondientes a partir de los *tokens*.

* **Opciones y Declaraciones**:

En este apartado se indican diferentes parámetros a la herramienta *JFlex*, como por ejemplo el nombre final del escáner generado, el tipo de caracteres utilizados o que mantenga una cuenta de la línea y columna de cada *token*. Además, también se aplican los ajustes necesarios para que sea compatible con la herramienta *CUP* que se utiliza más adelante para el analizador sintáctico. A continuación, también definimos una sección de código *Java* donde creamos algunas funciones auxiliares para poder crear correctamente las instancias de los Símbolos y gestionar errores. Por último, se definen algunas variables que servirán para definir las reglas léxicas a continuación, algunos ejemplos son las expresiones regulares para los comentarios, identificadores o literales.

* **Reglas Léxicas**:

En este último apartado se encuentran las reglas léxicas del escáner, donde se define como detectar y tratar los distintos *tokens* definidos anteriormente. En general se define la estructura de cada *token*utilizando su representación literal o utilizando las variables de expresiones regulares definidas en el apartado anterior. Una vez identificado el *token* en cuestión, se define el código *Java* a ejecutar en cada caso. Este código utiliza los métodos definidos anteriormente para generar las instancias de símbolo correctas para cada *token*, indicando el *token* en cuestión, el identificador del token y un valor en los casos necesarios.

Cabe destacar que uno de los métodos definidos en el apartado de declaraciones, es para la muestra correcta de los errores léxicos en caso de haber alguno. Este método utiliza la información de la línea y columna del token en cuestión para indicar al usuario de manera clara el error encontrado.

## Restricciones y Peculiaridades

El lenguaje de este compilador se ha basado en la gramática del lenguaje *Java* por tanto la mayoría de los tokens necesarios para definir dicha gramática se comparten. Hay algunas excepciones:

* **Identificadores:** Los identificadores tienen más restricciones en este lenguaje. En primer lugar, deben comenzar obligatoriamente por un carácter del alfabeto y después pueden estar seguidos de más caracteres o dígitos numéricos. Se ha restringido el uso de “\_” para gestionar internamente el apartado de generación de Código.
* **Comentarios:** Como se ha explicado anteriormente el lenguaje tiene un gran parecido con *Java* y en este caso comparte la misma definición de los distintos comentarios, tanto los de una línea, como los de múltiples líneas.
* ***Keywords*:**
  + **Operadores Lógicos:** los operadores *and, or y not* se representan con las mismas palabras, en lugar de utilizar los tokens de *Java* (&&, ||, !).
  + **Tipo Booleano:** para definir variables de tipo *booleano* se ha definido un *token*especial “*bool*”. Además, se ha modificado la representación de los literales *booleanos* “*True*” y “*False*” para que comiencen por mayúscula de una forma parecida al lenguaje *Python*.
  + **Otros:** finalmente se han definido algunos tokens especiales del lenguaje como “*crotofunc”* para la definición de funciones o “*const*” para la definición de variables constantes.

Cualquier carácter que no esté definido dentro de un *token*o en una expresión regular lanzará un error de carácter no reconocido.

# Analizador Sintáctico

## Gramática

A continuación, se explica la gramática definida para el lenguaje del compilador. En primer lugar, se explica la estructura básica de un programa en CROTO.

1. start with program;
2. program → main | methods main
3. main → CROTOFUNC MAIN LPAREN RPAREN LBRACE code\_block RBRACE
4. methods → method | methods method

La unidad mínima de compilación definida es *program*, un programa puede estar formado por una lista opcional de métodos (*methods*), y obligatoriamente un método *main* al final.

La lista de métodos *methods* puede contener un único método o una lista de varios, el *main* tendrá la forma crotofunc main**(){** code\_block **},** donde *code\_block* es un bloque de código que contiene todas las instrucciones a ejecutar.

La estructura de un método está definida en la gramática de la siguiente forma:

1. method → CROTOFUNC type id LPAREN param\_list RPAREN LBRACE code\_block RBRACE

Donde el *token*“*crotofunc*” es obligatorio para indicar el inicio de una función, *type* hace referencia al tipo de retorno de la función que puede ser *bool, int* o omitirlo en caso de que la función no retorne ningún valor. Seguidamente está el *string* identificador con el nombre de la función seguido de un paréntesis abierto, una definición opcional de parámetros y un paréntesis cerrado. Para finalizar hay un conjunto de llaves que abarca todo el *code\_block* que contiene todas las instrucciones a llevar a cabo dentro del método.

1. code\_block → | instructions
2. instructions → instruction | instructions instruction
3. instruction → statement | var\_declaration

Como se puede observar en las reglas anteriores, un *code\_block* puede estar **vacío**, o ser una lista de instrucciones. Una instrucción es una línea de Código que puede ser una declaración de variables o una sentencia. De esta manera dentro de cualquier bloque de código se pueden intercalar estos tipos de instrucciones y no separarlo en dos apartados diferentes como en algunos lenguajes por ejemplo *Ada*. A continuación, se puede ver la estructura de una declaración de variable

1. var\_declaration → CONSTANT type id SEMICOLON
2. | CONSTANT type id ASSIGNMENT expression SEMICOLON

En primer lugar, hay un *token* opcional para indicar si la variable es constante (solo se puede leer, después de declararla) o no, a continuación, está el tipo de datos de la variable que solo puede ser *bool* o *int*, seguido del *string* identificador con el nombre de la variable. Para finalizar, una declaración puede acabar aquí con un punto y coma, o asignar directamente un valor a la variable creada. Esto se haría indicado con un *token* “=” (llamado ASSIGNMENT) seguido de una expresión que devuelva un valor.

A continuación, tenemos la estructura del otro tipo de instrucción, las sentencias:

1. statement → id ASSIGNMENT expression SEMICOLON
2. | id LPAREN RPAREN SEMICOLON
3. | id LPAREN expression\_list RPAREN SEMICOLON
4. | if\_statement
5. | WHILE LPAREN expression RPAREN LBRACE code\_block RBRACE
6. | RETURN SEMICOLON
7. | RETURN expression SEMICOLON
8. | BREAK SEMICOLON
9. | error;

Una sentencia es una instrucción para realizar alguna función específica del lenguaje. En la gramática están definidas las siguientes:

* **Asignación:**

Sirve para dar o actualizar valores a variables declaradas. Comienza por el identificador de la variable, seguido del *token* de asignación (“=”) y la expresión a calcular para obtener el valor.

* **Llamada a funciones:**

Sirve para realizar llamadas a funciones que no tengan un valor de retorno o que no se quiera asignar a alguna variable. Su estructura es: el identificador de la función a llamar seguido de un paréntesis abierto, una lista opcional de argumentos que pasar a la función en caso de que se haya definido con parámetros y por último el cierre de paréntesis.

* ***If:***

Sirve para realizar saltos condicionales en el código, tiene la siguiente estructura:

1. if\_statement → IF LPAREN expression RPAREN LBRACE code\_block RBRACE
2. | IF LPAREN expression RPAREN LBRACE code\_block RBRACE ELSE if\_statement
3. | IF LPAREN expression RPAREN LBRACE code\_block RBRACE ELSE LBRACE | code\_block RBRACE LBRACE code\_block RBRACE

Empieza con el *token* especial *if* para indicar la sentencia, seguido de la expresión a evaluar entre paréntesis. A continuación, puede seguir opcionalmente un *code\_block* dentro de un *else if* con otra expresión a evaluar, o un *code\_block* en un *else* que se ejecutaría solo si no se cumple ninguna de las condiciones anteriores, pudiendo formar así cadenas de condicionales.

* ***While***:

Sirve para realizar un bucle condicional en el código, tiene una estructura muy parecida al *if* básico, pero el código del *code\_block* se ejecuta hasta que deje de cumplirse la condición definida en la expresión.

* ***Return*:**

Sirve para retornar de una llamada a función, se debe utilizar con una expresión en las funciones que tengan declaradas un tipo de retorno.

* ***Break:***

Por último, tenemos la sentencia *break* que sirve para salir de bucles en cualquier momento, su estructura es el *token* especifico *break* seguido de un punto y coma para terminar la línea.

Por último, tenemos las expresiones, un elemento muy necesario en la gramática, que se evalúan para obtener un valor. La gramática tiene las siguientes:

1. expression → expression ADDITION expression
2. | expression SUBTRACTION expression
3. | expression MULTIPLICATION expression
4. | expression DIVISION expression
5. | LPAREN expression RPAREN
6. | boolean\_expression
7. | id
8. | id LPAREN expression\_list RPAREN
9. | id LPAREN RPAREN
10. | literal
11. | SUBTRACTION expression
12. | error;

* **Aritméticas:**

Son expresiones que contienen un operador aritmético, estos pueden ser suma (“+”), resta (“-”), división (“/”), multiplicación (“\*”) y paréntesis (“()”). Estos operadores siguen el orden matemático de las operaciones combinadas indicando la precedencia. También está incluida la operación de negación, todas estas operaciones deben tener operandos del tipo *Int* y siempre devuelven un valor del mismo tipo.

* **Expresiones Booleanas:**

Son expresiones que solo se pueden evaluar a un valor booleano (*True* o *False*), la gramática consta de las siguientes:

1. boolean\_expression → expression EQUAL expression
2. | expression DIFFERENT expression
3. | expression GREATER expression
4. | expression LOWER expression
5. | expression GREATER\_EQUAL expression
6. | expression LOWER\_EQUAL expression
7. | expression AND expression
8. | expression OR expresión
9. | NOT expresión

Existen diferentes operadores de comparación y los operadores lógicos básicos. Este tipo de expresiones se puede usar como condición en los saltos y bucles, así como para asignar valores a variables de tipo *bool*.

* **Identificadores y Literales:**

Son expresiones que se evalúan directamente, en el caso de los literales se obtiene el valor del símbolo, y en el caso de variables se accede a la tabla para obtener el valor.

* **Llamadas a Funciones:**

Es un tipo de instrucción muy parecido al de las sentencias, aunque en este caso solo se puede utilizar en funciones que retornen un valor, dicho valor luego se puede utilizar como parte de una expresión.

## Generación del Analizador

Para el apartado del analizador sintáctico del compilador se ha utilizado la herramienta de generación de *parser* *CUP* [2]. Esta herramienta hace uso de una clase *Symbol* para representar a los tokens, y se puede enlazar con la herramienta utilizada en el apartado léxico *JFlex.* Para poder tener más control sobre los símbolos hemos definido nuestra propia clase *CrotoSymbol* y un *Factory* para crear los objectos *CrotoSymbolFactory* localizados en el directorio */src/sintactic/symbols/*.

Como se puede observar en el código del generador localizado en */cup/Grammar.cup* también existen tres grandes apartados para la generación del *parser* que explicaremos a continuación.

* **Código de *Java*:**

En este apartado inicial se puede encontrar código fuente de *Java* que se utilizara posteriormente en el archivo *parser* generado. El código está formado por *imports* de librerías utilizadas para la creación y manejo de símbolos, así como de funciones sobrescritas de *CUP* para poder gestionar los errores de una manera más limpia.

* **Declaraciones:**

En este apartado se encuentran tanto las declaraciones de los terminales como la de los no terminales que usamos en la gramática del lenguaje. Por cada símbolo terminal definido se genera un identificador en el archivo */src/sintactic/sym.java* que se puede utilizar para identificar los *tokens* en el analizador léxico. Para poder realizar posteriormente el análisis semántico utilizando el método de gestión con árbol sintáctico explícito se ha creado una clase para cada símbolo no terminal y de esta manera poder ir generando el árbol a medida que se recorren las producciones.

* **Reglas Sintácticas:**

Este apartado consta de todas las producciones que definen la gramática del lenguaje explicada anteriormente. Para cada posible producción tenemos un apartado de código *Java* donde creamos una instancia de la clase que representa la producción, de esta manera se va generando el árbol sintáctico. En cada instancia se almacena la línea y columna del inicio de la producción para poder mostrar información extendida en los errores semánticos (Se debe ejecutar el .cup con el *flag* (-*locations*).

## Restricciones y Peculiaridades

* Debido a la definición de *program* la unidad mínima de compilación, el método principal (*main*) debe estar siempre después de cualquier declaración de método.
* En las declaraciones de métodos, siempre debe haber un *return* al final de la función, pero puede haber otros dentro del *code\_block* si el usuario desea retornar en un punto anterior del codigo.
* También cabe destacar que no existe la posibilidad de declarar variables globales que se puedan utilizar
* Por último, una pecularidad del lenguaje es que puede haber bloques de codigo vacios, por tanto, se pueden declarar métodos sin codigo interno, o *if’s* y *while’s*.

# Analizador Semántico

Para el análisis del código fuente se ha utilizado el método de gestión con árbol sintáctico explicito. En el apartado anterior se ha ido generando el árbol sintáctico a medida que se recorren las producciones, donde cada nodo del árbol es una instancia de una clase que representa la producción en cuestión.

Dentro del directorio */src/semantic/symbols* esta la Clase *Structure,* que representa cualquier nodo del árbol sintáctico, junto con todas sus clases hijas que representan las distintas producciones sintácticas. Para producciones complejas, como por ejemplo *Statement* o *Expression* se han hecho subclases para cada uno de los distintos casos específicos.

Además, está definida la interfaz *Visitor* que representa los métodos que debe tener una clase que quiera realizar el recorrido del árbol sintáctico con un método *visit* para cada clase distinta. En la clase *Structure* existe un método abstracto *check* que deben implementar todas las demás clases, este método recibe un objeto visitante y debe indicarle que visite ese nodo en cuestión. Se ha seguido este proceso basándose en un ejemplo del repertorio del *CUP* [3].

## Tabla de Símbolos

Se han creado tres clases para implementar la Tabla de Símbolos:

* ***Variable***: esta clase contiene tres atributos: un tipo (entero o booleano), un Valor de tipo *Object* (para así poder almacenar ambos tipos de valor) y un booleano que indica si es constante o no. Las instancias de esta clase se irán creando y almacenando en la tabla de variables a medida que vayamos identificando las distintas variables del Programa.
* ***Table***: Es una clase que contiene un *HashMap* de (*String*, *Variable)* y un puntero a otra ***Table*** que es su tabla padre. La “tabla padre” de la tabla contiene las declaraciones de variables del ámbito inmediatamente superior a esta. Tiene un método para insertar una variable y un método para buscar si una variable está en esa tabla en concreto.
* ***SymbolTable***: tiene un puntero a la Tabla de variables del ámbito actual, un *HashMap* de parámetros (ya que vamos a tener **una tabla de símbolos para cada función**, al no tener variables globales), un contador de la profundidad de ámbito, y un tipo de retorno, que guarda el tipo de retorno de la función.

Hay definida una función para insertar parámetros, otra para obtenerlos, una función para entrar en un ámbito (crea una nueva Tabla con la tabla actual como padre de la nueva), una función para salir del ámbito (cambia la tabla actual por la tabla padre de la actual). Finalmente, una función para añadir variables (añade la variable a la tabla actual), y una para buscar variables.

La función para buscar variables se encarga de buscar la variable en el ámbito actual, y, si no la encuentra, se va al ámbito anterior (su tabla padre), y así hasta encontrar la variable o hasta llegar al último ámbito y no encontrarla, en ese caso devolverá *null*.

## Recorrido del Árbol Sintáctico

Para hacer el recorrido del árbol sintáctico y realizar las comprobaciones necesarias en cada caso se ha creado la clase *SemanticAnalyzer* que implementa la interfaz *Visitor*. En la clase hay definidas dos variables globales que se utilizaran a lo largo del recorrido, en primer lugar, tenemos una tabla de métodos de tipo *HashMap* donde la llave es el nombre del método y el elemento es un objeto *SymbolTable.* La segunda variable es *string* que guarda el nombre del método actual.

El análisis comienza con el nodo más alto del árbol devuelto por el analizador sintáctico, que es un objeto *Program*. Para comprobar semánticamente este nodo solo se deben pasar a los nodos hijos que en este caso son los métodos declarados, en el caso de que haya, y el método *main.*

Para el análisis de los métodos, primero se comprueba que no haya un método ya declarado con el mismo identificador o utilice el nombre de las funciones reservadas *print* y *scan*. A continuación, se comprueban cada uno de los parámetros, y se actualiza el tipo de retorno. Por último, se comprueba el *CodeBlock* si después de la comprobación no se ha encontrado un *return* y la función no era *void* se muestra un error semántico. Para la comprobación de los parámetros se revisa que no haya ningún parámetro anterior definido con ese nombre en el mismo método.

El análisis semántico de un *CodeBlock* consiste solamente en recorrer la lista de instrucciones y realizar la comprobación de cada una. Las comprobaciones que se realizan en el primer tipo de instrucción *VarDeclaration* son, asegurar que no haya ya una variable con el nombre indicado y en el caso que se quiera inicializar directamente asignado un valor que este sea del tipo esperado.

A continuación, se explican las distintas instrucciones del segundo tipo *Statement*:

* ***Assignment*:**

En primer lugar, se comprueba que la variable que se quiere actualizar este declarada correctamente, a continuación, se comprueba que no sea una variable constante, ya que estas no se pueden modificar. Por último, se comprueba que el valor que se quiere asignar sea del mismo tipo que la declaración de la variable.

* ***FunctionCall*:**

En este caso primero se comprueba si el nombre de la función a ejecutar es alguna de las

funciones reservadas del lenguaje *print* o *scan*. En el caso de *scan* se comprueba que tenga

un único argumento y este sea el identificador de una variable donde almacenar el valor. En

el caso de *print* se comprueba la expresión del argumento para obtener un valor. En

cualquier otro caso primero se comprueba que el método con ese identificador exista, en

caso contrario se muestra el error semántico. A continuación, se comprueba que el numero

de argumentos sea el mismo que el número de parámetros en la declaración del método y

que el tipo de cada uno corresponda.

* ***If* y *While*:**

Para las sentencias *if* y *while solo* se realiza la comprobación de que la expresión de condición sea de tipo booleano, en caso contrario se devuelve el error semántico correspondiente.

* ***Return:***

En primer lugar, se comprueba si el *return* va seguido de una expresión, si no es el caso y el método actual tenia un tipo de retorno distinto a *void*, se muestra un error semántico. Si hay una expresión de retorno se comprueba que sea del tipo esperado.

* ***Break:***

En este caso solo se comprueba que la sentencia este dentro de un bucle para poder salir, en caso contrario mostrar el error semántico correspondiente.

A continuación, se explican la gestión de las distintas expresiones del lenguaje:

* ***Arithmetic*:**

Se comprueba que ambos operadores sean de tipo entero, en caso contrario se muestra un error.

* ***Boolean:***

En primer lugar, se comprueba la expresión de la derecha al operador, si la operación es un *not* lógico se comprueba el resultado con este único operando. En cualquier otro caso se comprueba la expresión de la izquierda al operador y se asegura de que el resultado sea booleano. **Revisar classcastexception**

* ***FunctionCall*:**

Primero se comprueba que el método con ese identificador exista, en

caso contrario se muestra el error semántico. A continuación, se comprueba que el numero

de argumentos sea el mismo que el número de parámetros en la declaración del método y

que el tipo de cada uno corresponda.

* ***Literal*:**

Se indica el tipo de retorno del valor en función al literal

* ***Id*:**

# Generación de Código

# Referencias

|  |  |
| --- | --- |
| [1] | [En línea]. Available: https://www.jflex.de/. |
| [2] | [En línea]. Available: http://www2.cs.tum.edu/projects/cup/. |
| [3] | [En línea]. Available: https://versioncontrolseidl.in.tum.de/parsergenerators/cup/-/tree/master/testgrammars/minijava2/src/minijava. |