INSTITUTO TECNOLÓGICO DE COSTA RICA

Proyecto #1: Scanner

Escuela de Ingeniería en Computación
Compiladores e Intérpretes IC-5701

Alonso Navarro Carrillo, c. 2022236435

Carlos Venegas Masis, c. 2022153870

Valeria, c.

Ing. Ericka Marín Schumann
II Semestre 2024

Tabla de contenidos

Introducción

Este proyecto se sitúa en la primera etapa del desarrollo de un compilador para el lenguaje de programación C, conocida como el Análisis Léxico. El objetivo principal de esta etapa es diseñar y construir un scanner que sea capaz de identificar y clasificar los diferentes tokens que conforman un programa en C. Para lograr esto, se utilizó la herramienta JFlex, la cual permite definir expresiones regulares que describen los patrones de los tokens a reconocer.

Estrategia de solución

Después de leer detenidamente la documentación de JFlex, se comenzó a diseñar las expresiones regulares para los tokens que debía reconocer el escáner. Aquí surgió el primer problema: el escáner reconoce los tokens según el orden de prioridad. Es decir, si la primera expresión regular es un punto, ningún otro token será reconocido, ya que este metacaracter coincide con cualquier carácter, interpretándolo como un error. Por lo tanto, fue crucial definir adecuadamente el orden de las expresiones regulares.

Una vez definido el orden de las expresiones regulares, procedimos a diseñar la estructura de los tokens y sus tipos. Para ello, decidimos crear un mapa que facilitara la búsqueda de tokens repetidos en el documento. De manera similar, cada token guarda las líneas en las que aparece en un mapa, lo que permite incrementar el contador de ocurrencias de un token en una misma línea. Finalmente, los tipos de tokens se almacenan en un enum.

Por último, se implementaron errores definidos, como un número seguido de un identificador o números flotantes que comienzan con un punto. Era necesario definir estos errores como tokens para que pudieran ser reconocidos por el escáner. Los errores no se almacenan en una estructura de datos; en su lugar, se imprimen directamente y no se agregan al mapa de tokens.

Análisis de resultados

Actividad	Porcentaje realizado	Justificación
Desplegar lista de errores léxicos	100%	
Desplegar listado de tokens encontrados	100%	
Mostrar tipo de token, línea y cantidad de apariciones por cada token	100%	
Manejar 4 tipos grandes (operadores, literales, ids, palabras reservadas) de tokens	100%	
Ignorar comentarios en línea y bloque	100%	
Identificar todos los operadores válidos de C	100%	
Identificar todos los literales válidos de C	100%	
Identificar todos los identificadores válidos de C	100%	
Identificar todas las palabras reservadas de C	100%	
Definir errores léxicos	100%	

Lecciones aprendidas

Casos de prueba

Caso de prueba 1: Comentarios

```
#include <stdio.h> // Comentario después de directiva de preprocesador
#define MAX(a, b) ((a) > (b) ? (a) : (b)) // Macro con comentario en línea

// Este es un comentario en línea
int main() {
    // Comentario con caracteres especiales: !@#$%^&*()
    /* Comentario en bloque con caracteres especiales: !@#$%^&*() */
    // Comentario con código incorrecto: int x = ;
    /* Comentario en bloque con código incorrecto: int y = ; */
```

```
        // Comentario con    espacios
        /*  Comentario en     bloque con espacios */
        //      Comentario con tabulaciones
        int a = 10; // Comentario al final de una línea de código
        int b = 20; /* Comentario en bloque
                       que se extiende en varias líneas */
        int d = a + b;
        int c = a + b; // Comentario después de una expresión
        char *str = "Este es un string con // comentario en línea";
        char *str2 = "Este es otro string con /* comentario en bloque */";
        b = 20;
        /* Comentario en bloque sin cerrar
        b = 20; // Comentario en línea anidado
        return 0;
    }
```

Errores esperados:

- Error en la línea 1: Operador inválido '#'.
- Error en la línea 2: Operador inválido '#'.
- Error en cáscada en línea 20: Comentario de bloque sin cerrar.

Es importante resaltar que aunque el proyecto indique que no se pueden tener errores en cáscada, al encontrar un comentario de bloque sin cerrar el comportamiento usual será seguir buscando hasta encontrar un "*/".

Resultados:

```
Errors:
Character unknown: # in 1
Character unknown: # in 2
Block comment without closure: /* Comentario en bloque sin cerrar
    b = 20; // Comentario en linea anidado
    return 0;
} in 20
+------------+-----------------+-----------------------------------------+
| Token      | Tipo de Token   | Linea                                   |
+------------+-----------------+-----------------------------------------+
| char       | KEYWORD         | 17, 18                                  |
+------------+-----------------+-----------------------------------------+
```

int	KEYWORD	5, 13, 14, 15, 16
MAX	ID	2
a	ID	2(3), 13, 15, 16
b	ID	2(3), 14, 15, 16, 19
c	ID	16
d	ID	15
define	ID	2
h	ID	1
include	ID	1
main	ID	5
stdio	ID	1
str	ID	17
str2	ID	18
(	OPERATOR	2(6), 5
)	OPERATOR	2(6), 5
,	OPERATOR	2
.	OPERATOR	1
;	OPERATOR	13, 14, 15, 16, 17, 18, 19
=	OPERATOR	13, 14, 15, 16, 17, 18, 19
{	OPERATOR	5

*	OPERATOR_ARITHMETIC	17, 18
+	OPERATOR_ARITHMETIC	15, 16
:	OPERATOR_RELATIONAL	2
<	OPERATOR_RELATIONAL	1
>	OPERATOR_RELATIONAL	1, 2
?	OPERATOR_RELATIONAL	2
10	LITERAL_INT	13
20	LITERAL_INT	14, 19
"Este es otro string con /* comentario en bloque */"	LITERAL_STR	18
"Este es un string con // comentario en linea"	LITERAL_STR	17

Caso de prueba 2: Operadores

```
#include <stdio.h>

int main() {
    int a = 5 + 3;          // Suma
    int b = a - 2;          // Resta
    int c = a * b;          // Multiplicación
    int d = c / 2;          // División
    int e = c % 2;          // Módulo
    int f = a & b;          // AND bit a bit
    int g = a | b;          // OR bit a bit
    int h = a ^ b;          // XOR bit a bit
    int i = ~a;             // NOT bit a bit
    int j = a << 2;         // Desplazamiento a la izquierda
    int k = b >> 1;         // Desplazamiento a la derecha

    if (a == b) {           // Igualdad
        printf("a es igual a b\n");
    }

    if (a != b) {           // Desigualdad
        printf("a no es igual a b\n");
    }

    if (a < b) {            // Menor que
        printf("a es menor que b\n");
    }

    if (a > b) {            // Mayor que
        printf("a es mayor que b\n");
    }

    if (a <= b) {           // Menor o igual que
        printf("a es menor o igual que b\n");
    }

    if (a >= b) {           // Mayor o igual que
        printf("a es mayor o igual que b\n");
```

```
        }

        if (a && b) {           // AND lógico
            printf("a y b son verdaderos\n");
        }

        if (a || b) {           // OR lógico
            printf("a o b es verdadero\n");
        }

        if (!a) {               // NOT lógico
            printf("a es falso\n");
        }

        // Errores intencionales
        int l = a @ b;          // Operador inválido
        int n = a // b;         // Comentario mal formado
        int p = a +;            // Operador sin operando
        int q = ;               // Declaración sin inicialización

        return 0;
    }
```

Errores esperados:

- Error en la línea 1: Operador inválido '#'.
- Error en la línea 53: Operador inválido '@'.

También se espera que suponga el '//' como un comentario y no como dos operadores de división.

Resultados:

```
Errors:
Character unknown: # in 1
Character unknown: @ in 53
+------------+-----------------+-----------------------------------------+
| Token      | Tipo de Token   | Linea                                   |
+------------+-----------------+-----------------------------------------+
| if         | KEYWORD         | 16, 20, 24, 28, 32, 36, 40, 44, 48      |
```

int	KEYWORD	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 53, 54, 55, 56
return	KEYWORD	58
a	ID	4, 5, 6, 9, 10, 11, 12, 13, 16, 20, 24, 28, 32, 36, 40, 44, 48, 53, 54, 55
b	ID	5, 6, 9, 10, 11, 14, 16, 20, 24, 28, 32, 36, 40, 44, 53
c	ID	6, 7, 8
d	ID	7
e	ID	8
f	ID	9
g	ID	10
h	ID	1, 11
i	ID	12
include	ID	1
j	ID	13
k	ID	14
l	ID	53
main	ID	3
n	ID	54
p	ID	55

printf	ID	17, 21, 25, 29, 33, 37, 41, 45, 49
q	ID	56
stdio	ID	1
(	OPERATOR	3, 16, 17, 20, 21, 24, 25, 28, 29, 32, 3 3, 36, 37, 40, 41, 44, 45, 48, 49
)	OPERATOR	3, 16, 17, 20, 21, 24, 25, 28, 29, 32, 3 3, 36, 37, 40, 41, 44, 45, 48, 49
.	OPERATOR	1
;	OPERATOR	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17 , 21, 25, 29, 33, 37, 41, 45, 49, 53, 55 , 56, 58
=	OPERATOR	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 53 , 54, 55, 56
{	OPERATOR	3, 16, 20, 24, 28, 32, 36, 40, 44, 48
}	OPERATOR	18, 22, 26, 30, 34, 38, 42, 46, 50, 59
%	OPERATOR_ARITHMETIC	8
*	OPERATOR_ARITHMETIC	6
+	OPERATOR_ARITHMETIC	4, 55
-	OPERATOR_ARITHMETIC	5
/	OPERATOR_ARITHM	7

	ETIC	
!	OPERATOR_LOGICAL	48
&&	OPERATOR_LOGICAL	40
||	OPERATOR_LOGICAL	44
!=	OPERATOR_RELATIONAL	20
<	OPERATOR_RELATIONAL	1, 24
<=	OPERATOR_RELATIONAL	32
==	OPERATOR_RELATIONAL	16
>	OPERATOR_RELATIONAL	1, 28
>=	OPERATOR_RELATIONAL	36
&	OPERATOR_BITWISE	9
<<	OPERATOR_BITWISE	13
>>	OPERATOR_BITWISE	14
^	OPERATOR_BITWISE	11

	E	
|	OPERATOR_BITWIS E	10
~	OPERATOR_BITWIS E	12
0	LITERAL_INT	58
1	LITERAL_INT	14
2	LITERAL_INT	5, 7, 8, 13
3	LITERAL_INT	4
5	LITERAL_INT	4
"a es fals o\n"	LITERAL_STR	49
"a es igua l a b\n"	LITERAL_STR	17
"a es mayo r o igual que b\n"	LITERAL_STR	37
"a es mayo r que b\n"	LITERAL_STR	29
"a es meno r o igual que b\n"	LITERAL_STR	33
"a es meno r que b\n"	LITERAL_STR	25
"a no es i	LITERAL_STR	21

```
| gual a b\n |                 |                                          |
| "          |                 |                                          |
+------------+-----------------+------------------------------------------+
| "a o b es  | LITERAL_STR     | 45                                       |
| verdadero\ |                 |                                          |
| n"         |                 |                                          |
+------------+-----------------+------------------------------------------+
| "a y b son | LITERAL_STR     | 41                                       |
|  verdadero |                 |                                          |
| s\n"       |                 |                                          |
+------------+-----------------+------------------------------------------+
```

Manual de usuario

Instalación

Para construir y ejecutar el proyecto, es necesario tener Java instalado en tu sistema. Sigue estos pasos para configurar el proyecto:

1. Clona el repositorio:

   ```
   git clone https://github.com/AlonsoNav/CCompilerJFlex.git
   cd your-repo
   ```

2. Genera el archivo `CLexer`:

   ```
   java -jar lib/jflex-full-1.9.1.jar src/scanner/CLexer.flex
   ```

3. Compila el proyecto:

   ```
   javac -d bin -sourcepath src src/app/Main.java
   src/scanner/CLexer.java .\src\scanner\Token.java
   .\src\scanner\TokenType.java
   ```

Uso

Para ejecutar el compilador con un archivo de entrada, utiliza el siguiente comando:

```
java -cp bin app.Main input_file
```

También puedes enviar la salida a un archivo .txt con

```
java -cp bin app.Main input_file > output.txt
```

Bitácora

Fecha: 26-08-2024

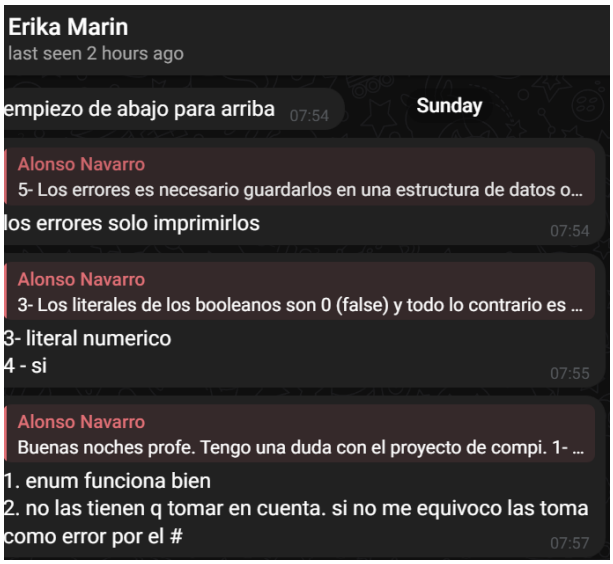
En la primera reunión del equipo de trabajo, se acordó que CV se encargará de los expresiones regulares de los operadores y del formato de impresión de la tabla. AN diseñará la estructura de los tokens y sus errores, así como las expresiones regulares de los identificadores y palabras reservadas. VG se responsabilizará de los literales. Por último, se decidió que la documentación se realizará en LaTeX y que GitHub será utilizado como sistema de control de versiones.

Fecha: 01-09-2024

La profesora responde las siguientes consultas:

- Nosotros ya manejamos los diferentes tipos de tokens en un enum. Queremos agregar subtipos pensando a futuro, pero no sé si sea mejor manejar cada subtipo de token como una clase aparte o si todo chorreado en un enum funciona. ¿Usted qué me recomienda?
- ¿Qué hacemos con las directivas para el procesador como `#include` o `#define`?
- Los literales de los booleanos son 0 (false) y todo lo contrario es true eso lo tomamos como un literal numérico no importa?
- Manejamos también sufijos (U: unsigned, L: long...)?
- Los errores es necesario guardarlos en una estructura de datos o solo con imprimirlos basta?

Lo siguiente son las respuestas de la profesora:

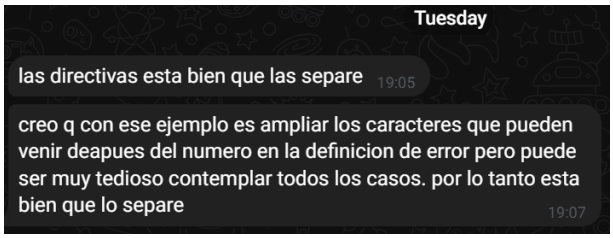


Fecha: 03-09-2024

La profesora responde la siguiente consulta:

> Usted comentó que cosas como ”1ejemplo”, debería ser un error y no que los separe como ”1”: literal y ”ejemplo”: id. Hay otras cosas que se pegan, por ejemplo con las directivas:
>
> #include <stdio.h >
>
> lo separa como
>
> #: error
>
> include: id
>
> <operador
>
> stdio: id
>
> . operador
>
> h: id
>
> >operador
>
> Entonces yo ya tengo definido el error genérico de ”1ejemplo”, pero con esto estaba pensando en meter más caracteres en ese mismo error, pero es que si fuera un ”5<identificador” en algún condicional ya croma.

La respuesta de la profesora es:



Fecha: 04-09-2024

Luego de tener todo el scanner corriendo correctamente se decide que lo siguiente es finalizar la documentación. CV se encargará de la introducción, AN de la estrategia de solución y VG de las lecciones aprendidas. Las demás secciones son redactadas en conjunto. Finalmente, se acuerda que cada integrante hará dos casos de prueba y documentará lo encontrado.

Bibliografía

[1] Klein, G., Rowe, S., & Décamps, R. (marzo de 2023). *JFlex User's Manual*. JFlex Team. En: `https://www.jflex.de/manual.html`.