Universidad Rafael Landívar
Facultad de Ingeniería
Ingeniería en Informática y Sistemas
Catedrático: Moises Alonso

# PROYECTO PRÁCTICO FASE I

## Generador de Scanner

Daniel Fernando Cabrera Reyes
1117121

José Mario Marroquín Roldán
1234621

Diego Andres Bautista Cruz
1181821

Guatemala, 27 de febrero del 2023

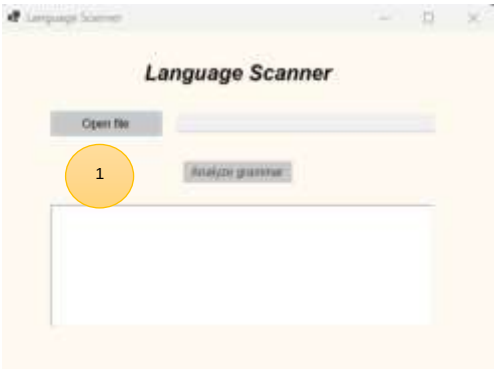
# I. Descripción del Proyecto

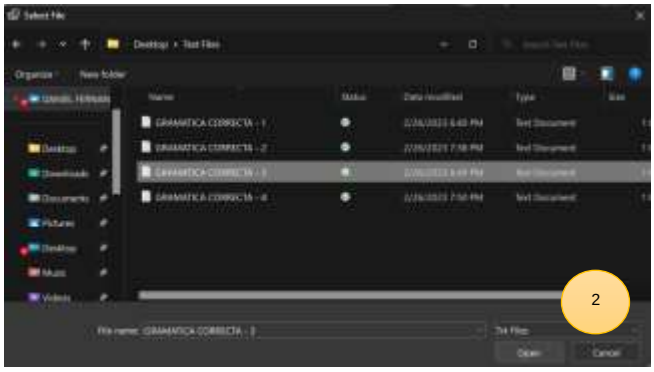
El proyecto presentado consiste en una aplicación generadora de Scanner, la cual tiene como objetivo analizar la estructura léxica de la definición de una gramática contenida en un archivo de texto. El análisis consiste en la validación y comprobación del formato específico para los elementos que conforman a la gramática, los cuales son: SETS, TOKENS, ACTIONS y ERROR. Cada una de las secciones cuenta con su propio objetivo, formato y elementos, requerido en conjunto para una correcta gramática. La correcta verificación del formato de la gramática a utilizar permitirá en un futuro la correcta construcción del lenguaje que describe esta gramática.

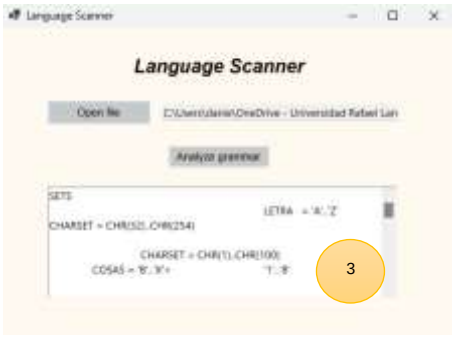
# II. Manual de Usuario

### 1. Abrir archivo de texto

Al pulsar el botón “Open file”, se abrirá una ventana de diálogo, donde se podrá buscar el archivo a analizar. Al encontrar el archivo, hacer clic sobre él y presionar “Abrir” en la esquina inferior derecha de la ventana de diálogo. Al finalizar este procedimiento, se mostrará el texto del archivo en la caja de texto del programa.





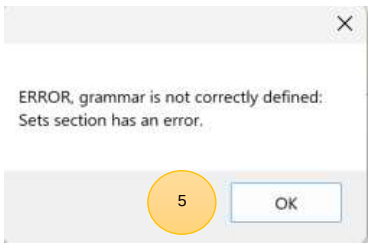

En el repositorio de GitHub existe una carpeta “Inputs” el cual contiene dos carpetas, una para los textos con los lenguajes con sintaxis correcta y el otro con sintaxis incorrecta.

Repositorio de GitHub: https://github.com/danieelfcr/LanguageScanner.git

## 2. Analizar gramática

Al presionar el botón “Analyze grammar”, se comenzará a analizar la gramática del texto incluido en el archivo. Al finalizar el análisis, se observará una ventana de diálogo la cuál mencionará si la gramática es correcta o incorrecta. En caso de ser incorrecta, se mostrará la sección en la cuál la gramática es errónea.

Al cerrar las ventanas emergentes, la caja de texto cambiará de color dependiendo si la gramática del texto analizado es correcto (verde) o si es incorrecto (rojo).


# III. Manual Técnico

## *a) Especificaciones*

- **Lenguajes de programación utilizado:** C#
- **Versión:** .NET 6.0
- **Sistema operativo:** Windows
- **Versión sistema operativo:** 7.0 en adelante
- **Recursos utilizados:** Archivo de texto (.txt)

## *b) Características de la Gramática*

### a. SETS

La sección de SETS corresponde a la definición de un conjunto de símbolos no terminales a los que se les asigna un conjunto de símbolos terminales, los cuáles pueden ser caracteres o dígitos. Entre las características de esta sección se encuentran:

- ✓ Esta sección puede no venir dentro del archivo de texto.
- ✓ Si esta sección viene, debe poseer al menos un SET.
- ✓ La palabra SETS debe de venir en mayúscula.
- ✓ Se puede utilizar la función CHR(Número del 0 al 255).
- ✓ Puede haber muchos espacios entre el identificador, el símbolo “=” y la definición.
- ✓ Puede haber muchos saltos de línea.
- ✓ Para identificar que se desea de un carácter a otro (ya sea con la función CHR() o con el carácter en sí), se debe de colocar el no terminal seguido de dos puntos (..) y el otro terminal sin espacios.
- ✓ Los sets pueden estar concatenados con el signo “+”. En este caso sí pueden existir espacios.

### b. TOKENS

La sección TOKENS consiste en un conjunto de definiciones para una expresión regular, que se utilizará en versiones posteriores para un análisis sintáctico de la gramática. Es una sección de gran importancia, pues combina algunas características de las demás secciones. Sus principales características son:

- ✓ Debe ser una sección obligatoria.
- ✓ Tiene que comenzar con la palabra TOKENS.
- ✓ Cada elemento dentro de la sección debe comenzar con la palabra TOKEN, seguido de un número de identificación y un signo igual.
- ✓ Al menos un TOKEN debe existir.
- ✓ Pueden haber indefinidos saltos de línea y espacios o tabulaciones entre dos TOKENS
- ✓ Cada TOKEN puede contener uno o más símbolos.
- ✓ Los símbolos permitidos son no terminales (en mayúsculas) o terminales (un solo símbolo que puede ser o no una letra mayúscula y está encerrado entre comillas simples).
- ✓ Cada símbolo puede ser acompañado o no de uno de los tres operadores básicos de una expresión regular (+, *,?).
- ✓ Pueden existir paréntesis para agrupar símbolos, pero cada apertura debe contar con su respectivo cierre.
- ✓ La disyunción entre dos símbolos se representa con el operador (|), el cual debe tener obligatoriamente un símbolo a cada lado para ser válido.
- ✓ Se pueden agregar llamadas a funciones definidas en la sección ACTIONS, siempre y cuando estén dentro de una apertura y cierre de los símbolos de llaves ({ , }).
- ✓ Cada función debe escribirse en mayúsculas y finalizar con una apertura y cierre de paréntesis.

### c. ACTIONS

La sección de ACTIONS corresponde a un conjunto de funciones que definen las palabras reservadas del lenguaje. Estas definiciones contenidas en diversas funciones serán útiles para brindarle un significado implícito a palabras escritas de forma específica. Algunas de las características son:

- ✓ Las funciones contenidas poseer: un identificador, paréntesis abierto y cerrado y tokens,
- ✓ Los tokens se forman de un identificador numérico, un signo igual y la definición de la palabra contenida en apóstrofes,
- ✓ El identificador de los tokens debe ser únicamente numérico y en la definición de la palabra solo puede utilizarse letras,
- ✓ La función “RESERVADAS()” siempre debe estar contenida en la sección una sola vez

### d. ERROR

La sección de ERROR se refiere al contenido de todos los posibles mensajes de error que se utilizarán para el lenguaje. En la gramática no cuenta con ningún encabezado, ya que son una lista secuencial de errores con las siguientes características:

- ✓ Cada error está conformado por: un identificador, un signo de igualdad y un número
- ✓ El identificador puede ser conformado únicamente por letras y este debe tener por los menos la palabra “ERROR” como sufijo
- ✓ Del lado derecho de la igualdad solamente se definen valores numéricos

## *c) Expresiones regulares utilizadas*

Para la validación del archivo de texto que contiene la gramática mediante el scanner se utilizaron expresiones regulares. Las cuales validan exclusivamente la construcción léxica de la gramática; esto quiere decir que se verifica el cumplimiento de los formatos específicos de cada sección.

Para la construcción de las expresiones regulares en el lenguaje de programación se importó una librería que contiene la definición de una clase para la escritura y revisión de expresiones regulares, propia de C#, “Regex”. Esto conlleva al uso de términos específicos del lenguaje, tales como:

- ‘\\S’: cualquier carácter que no se un dígito.
- ‘\\t’: tabulación
- ‘\\n’: salto de línea
- ‘\\s’: cualquier tipo de espacio en blanco (\\n, \\t, \\r, \\b)
- ‘[X-X]’: la definición de rangos de números y letras
- ‘*’: cerradura estrella
- ‘+’: cerradura positiva
- ‘|’: alternación, OR
- ‘()’: signos de agrupación

A continuación, se presentará la expresión regular utilizada para cada sección, así como una breve explicación del funcionamiento de cada una:

### ✓ Expresión regular: “SETS”

```
(\\s*SETS(\\s)+((\\s)*([A-Z])+( |\\t)*=( |\\t)*('([a-z])'\\.\\.'([a-z])'|'([A-
Z])'\\.\\.'([A-Z])'|'.'|'([0-9])'\\.\\.'([0-9])'|CHR\\((([0-9])|([0-9])([0-9])|0([0-
9])([0-9])|1([0-9])([0-9])|2([0-5])([0-5]))\\)(\\.\\.CHR\\((([0-9])|([0-9])([0-
9])|0([0-9])|0([0-9])([0-9])|1([0-9])([0-9])|2([0-5])([0-5]))\\))?)(( |\\t)*\\+(
|\\t)*('([a-z])'\\.\\.'([a-z])'|'([A-Z])'\\.\\.'([A-Z])'|'.'|'([0-9])'\\.\\.'([0-
9])'|CHR\\((([0-9])|([0-9])([0-9])|0([0-9])([0-9])|1([0-9])([0-9])|2([0-5])([0-
5]))\\)(\\.\\.CHR\\((([0-9])|([0-9])([0-9])|0([0-9])|0([0-9])([0-9])|1([0-9])([0-
9])|2([0-5])([0-5]))\\))?))*)+\\s*)?
```

Esta sección, según la expresión regular presentada, puede o no venir dentro del archivo de texto. Si la sección viene, debe de contener obligatoriamente la palabra “SETS” y un set. Luego de la palabra “SETS” debe de venir un salto de línea. Luego, pueden venir muchos saltos de línea, espacios o tabulaciones para encontrarse con el primer símbolo no terminal. Los símbolos no terminales deben estar escritos en mayúscula obligatoriamente. Pueden venir muchos espacios entre el no terminal y el símbolo “=”, al igual que entre el símbolo “=” y el primer símbolo terminal. Los símbolos terminales pueden ser definidos como un solo carácter entre comillas simples, puede definirse un rango entre caracteres, por ejemplo ‘A’..’Z’, en donde los terminales deben de estar escritos entre comillas y deben de separarse por dos puntos “..” sin espacios. También puede definirse un carácter con ASCII con la función CHR(), tomando en cuenta que los caracteres ASCII van desde 0 hasta 255. También se puede definir un rango con las funciones CHR() utilizando las mismas reglas descritas anteriormente. Los símbolos terminales pueden estar concatenados utilizando el símbolo “+”, siguiendo las mismas reglas, pudiendo haber muchos espacios entre los símbolos concatenados.

## ✓ Expresión regular de la sección “TOKENS”

```
\\s*TOKENS\\s*\\n(\\s*TOKEN( |\\t)*([1-9][0-9]*)( |\\t)*=( |\\t)*(((( |\\t)*{(
|\\t)*([A-Z]+( |\\t)*\\(( |\\t)*\\)( |\\t)*)+( |\\t)*}( |\\t)*)|(([A-Z]+(
|\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))|(( |\\t)*(([A-Z]+(
|\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))+( |\\t)*\\|( |\\t)*(([A-Z]+(
|\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))( |\\t)*(\\|( |\\t)*(([A-Z]+(
|\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?)( |\\t)*)+)*)|(\\(( |\\t)*((((A-
Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))( |\\t)*)|(( |\\t)*(([A-
Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))+( |\\t)*\\|( |\\t)*(([A-
Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))( |\\t)*(\\|( |\\t)*(([A-
Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?)( |\\t)*)+)*)|(\\((
|\\t)*((((A-Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))( |\\t)*)|((
|\\t)*(([A-Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))+( |\\t)*\\|(
|\\t)*(([A-Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?))( |\\t)*(\\|(
|\\t)*(([A-Z]+( |\\t)*(\\*|\\+|\\?)?)|('\\S'( |\\t)*(\\*|\\+|\\?)?)(
|\\t)*)+)*))+\\)( |\\t)*(\\*|\\+|\\?)?))+\\)( |\\t)*(\\*|\\+|\\?)?))( |\\t)*)+)+\\s*
```

TOKENS es una sección obligatoria que tiene el objetivo de representar expresiones regulares siguiendo un formato riguroso. La primera de sus características es que debe de contar con la palabra específica “TOKENS”, seguida de caracteres de espacio y un salto de línea obligatorio. Cada TOKEN es acompañado de un número de identificación que puede repetirse y no tiene limitantes, seguido del signo “=”. A partir de dicho formato pueden ocurrir cuatro casos distintos que ayudarán a formar una expresión regular válida, distinta a las expresiones usadas para validar el formato de la gramática. El primer caso es para determinar símbolos terminales y no terminales, siendo los primeros siempre encerrados entre comillas simples, ambos tipos de símbolos pueden ser acompañados, o no, de uno de los tres operadores básicos de una expresión regular (*, +,?).

El segundo caso permite utilizar el símbolo de la disyunción para expresiones regulares (|) y puede generar cadenas de varias disyunciones, obligando a que el formato de entrada respete que ambos lados de la disyunción deben de existir. El tercer caso se enfoca en los símbolos de agrupación o paréntesis, con los cuales son válidos tanto el primer caso como el segundo, además, es posible contener una pareja de paréntesis dentro de otra. La expresión regular mostrada se encarga de validar que una apertura de paréntesis siempre esté acompañada de su respectivo cierre, evitando errores de formato. El cuarto caso es el encargado de permitir el uso de funciones dentro de un TOKEN, para lo cual siempre deberá abrirse y cerrarse una pareja de llaves ({,}) en donde dentro pueden existir varias llamadas a funciones acompañadas de una pareja abierta y cerrada de paréntesis.

## ✓ Expresión regular: “ACTIONS”

```
ACTIONS\\s*RESERVADAS\\s*\\(\\s*\\)\\s*{\\s*([0-9]|([1-9][0-9]*))\\s*=\\s*'([A-Z]|[a-z])+'\\s*)+}\\s*([A-Z]|[a-z])+\\s*\\(\\s*\\)\\s*{\\s*([0-9]|([1-9][0-9]*))\\s*=\\s*'([A-Z]|[a-z])+'\\s*)+}\\s*)*
```

La sección de ACTIONS debe ser definida en la gramática. Al comienzo de esta debe tener la palabra ‘ACTIONS’, a la cual puede seguir luego muchos o ningún espacio en blanco, que ahora en adelante se referirá a estos como “espacios”.

Luego comienza la definición de funciones con su identificados con letras mayúsculas, espacios, paréntesis abierto, espacios, paréntesis cerrar espacios y se abren llaves, lo que se refiere al contenido de las funciones que son tokens propios de esta sección. En cuanto a la definición de funciones, la primer en ser declarada es la función “RESERVADAS()” que debe existir obligatoriamente en la gramática, después pueden venir o no más funciones que cumplan el formato.

En cuanto al contenido de las funciones, los tokens de la función son definidos mediante 3 elementos: un identificador numérico, un signo de igualdad y una palabra contenida en apóstrofes. La forma de evaluación es así: espacios, números como identificador, espacios, signo igual, espacios, palabra contenida en apóstrofes y espacios; así en cada token. La función debe tener por lo menos un token y puede tener varios.

### ✓ Expresión regular: “ERROR”

```
([A-Z]*ERROR\\s*=\\s*[1-9][0-9]*\\s*)+
```

La sección de ERROR se refiere a la definición de errores utilizados por el lenguaje generado por la gramática analizada correctamente. Esta sección comienza inmediatamente después de la sección de ACTIONS, luego de la última llave cerrada de la última función. En cualquier gramática siempre debe haber definido al menos un error y puden definirse muchos. La estructura de los errores es de la forma: el identificador del error, un signo de igualdad y un número. El identificador del error se compone únicamente de letras y al menos la palabra “ERROR” al final, o bien solo esta. La forma de evaluarlo es así: espacios, identificador de error, espacios, signo de igualdad, espacios, número, espacios.