Análisis lexicográfico / léxico

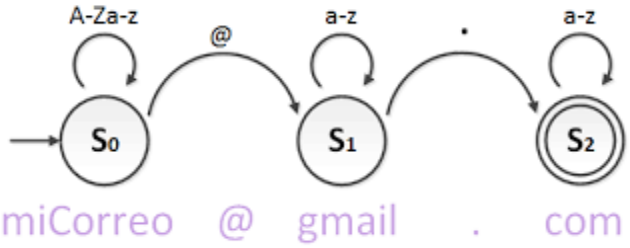
Visión general.

Las técnicas utilizadas para construir analizadores léxicos también se pueden aplicar a otras áreas como, por ejemplo, a lenguajes de consulta y sistemas de recuperación de información.

Caracteres Regex Básicos	Significado
.	Cualquier carácter individual
[]	Cualquier carácter especificado
[^]	No el carácter especificado
*	Cero o más de los carácter previos
^	Si ^ es el primer carácter en el patrón, entonces el patrón debe estar al principio de la linea que debe coincidir, de lo contrario solo un literal ^.
$	Si $ es el último carácter en el patrón, entonces el patrón debe estar al final de la linea que debe coincidir, de lo contrario solo un literal $.

AulaFacil.com

En cada aplicación, el problema de fondo es la especificación y diseño de programas que ejecuten las acciones activadas por palabras que siguen ciertos patrones dentro de las cadenas a reconocer.

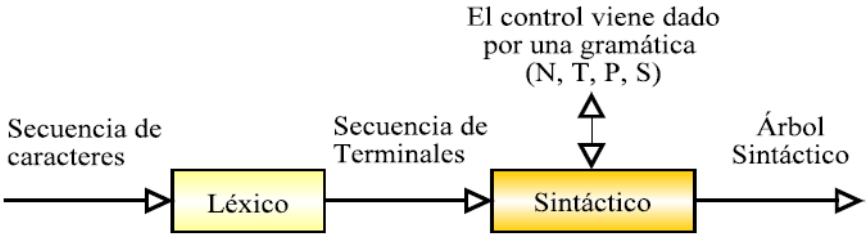


Como la programación dirigida por patrones está ampliamente extendida y resulta de indudable utilidad, existen numerosos metalenguajes que permiten establecer pares de la forma patrón-acción, de manera que la acción se ejecuta cada vez que el sistema se encuentra una serie de caracteres cuya estructura coincide con la del patrón. En concreto y para este curso se utilizará Lex con el objetivo de especificar los analizadores léxicos. En este lenguaje, los patrones se especifican por medio de expresiones regulares, y el metacompilador Lex genera un reconocedor de las expresiones regulares mediante un autómata finito determinista eficiente.

Concepto de Analizador Léxico

Se encarga de buscar los componentes léxicos o palabras que componen el programa fuente, según unas reglas o patrones. Este analizador también genera la tabla de símbolos.

La entrada del analizador léxico podemos definirla como una secuencia de caracteres. El analizador léxico divide esta secuencia en palabras con significado propio y después las convierte a una secuencia de terminales desde el punto de vista del analizador sintáctico.

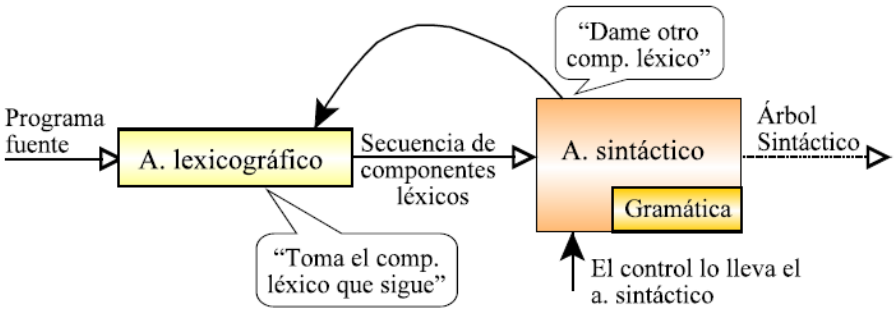


donde:

N ≡ No terminales
T ≡ Terminales
P ≡ Reglas de producción
S ≡ Axioma inicial

Funciones del análisis Léxico

Su principal función consiste en leer los caracteres de entrada y elaborar como salida una secuencia de componentes léxicos que utiliza el analizador sintáctico para hacer el análisis.



Otras funciones:

• Eliminar los comentarios del programa.

• Eliminar espacios en blanco, tabuladores, retorno de carro, etc, y en general, todo aquello que carezca de significado según la sintaxis del lenguaje.

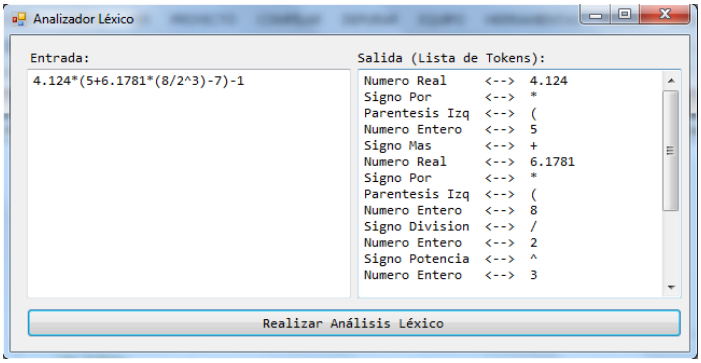
• Reconocer los identificadores de usuario, números, palabras reservadas del lenguaje, etc., y tratarlos correctamente con respecto a la tabla de símbolos (solo en los casos en que este analizador deba tratar con dicha estructura).

• Llevar la cuenta del número de línea por la que va leyendo, por si se produce algún error, dar información acerca de dónde se ha producido.

• Avisar de errores léxicos. Por ejemplo, si el carácter '@' no pertenece al lenguaje, se debe emitir un error.

• También puede hacer funciones de preprocesador.

Necesidad del analizador léxico



Simplificación

Se hace patente cuando es necesario realizar modificaciones o extensiones al lenguaje inicialmente ideado.

Eficiencia

Gran parte del tiempo de compilación se invierte en leer el programa fuente y dividirlo en componentes léxicos. Con técnicas especializadas de manejo de buffers para la lectura de caracteres de entrada y procesamiento de patrones se puede mejorar significativamente el rendimiento de un compilador.

Portabilidad

Las peculiaridades del alfabeto de partida, del juego de caracteres base y otras anomalías propias de los dispositivos de entrada pueden limitarse al analizador léxico.

Patrones complejos

Permite realizar el reconocimiento de componentes básicos complejos, todo dependerá de los patrones establecidos.

Token, patrón y lexema

Token	Lexema	Patron	Reservada
identificador	temp, aux, a	letra seguida de letra o digito	No
numero	16, 30, 2	digito seguido de mas digitos	No
if	if	letra i seguida de f	Si
op_asig	=	simbolo =	Si

Patrón: es una expresión regular.

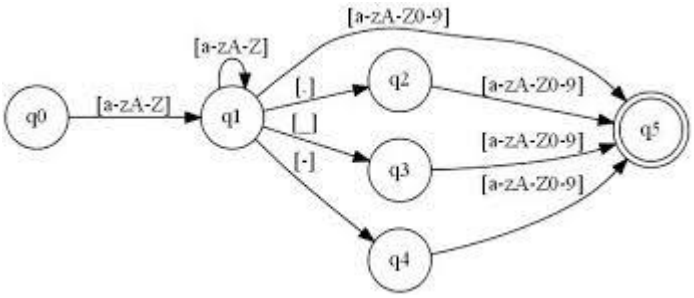
Lexema: Es cada secuencia de caracteres concreta que encaja con un patrón.

Token: es la categoría léxica asociada a un patrón. Cada token se convierte en un número o código identificador único. Este concepto coincide con el "terminal" de la fase sintáctica.

Patrón -> Lexema -> Token

¿Cómo construir un analizar léxico?

Ad hoc. Consiste en la codificación de un programa reconocedor que no sigue los formalismos propios de la teoría de autómatas. Este tipo de construcciones es muy propenso a errores y difícil de mantener.



Mediante la implementación manual de los **autómatas finitos (máquinas de estado)**. este método es sistemático y no propenso a errores, pero cualquier actualización de los patrones reconocedores implica la modificación del código que los implementa, por lo que el mantenimiento se hace muy costoso.

*AFD = Autómatas Finitos Determinísticos

Mediante un **metacompilador**. Éste genera todos los autómatas finitos, los convierte a autómata finito determinista, y lo implementa en algún lenguaje de programación.