Analizador Léxico

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Aspectos Generales

- Un analizador léxico es la especificación y el diseño de programas que ejecuten las acciones activadas por patrones dentro de las cadenas.
- La principal función es leer los caracteres de entrada y elaborar como salida una secuencia de componentes léxicos que se utilizaran.
- Convierte una cadena de caracteres en una cadena de palabras.
- Una forma sencilla de crear un analizador léxico consiste en la elaboración de un diagrama que muestre la estructura de los componentes léxicos del archivo fuente, y después hacer la traducción “a mano” del diagrama a un programa para encontrar los componentes léxicos.

Aspectos Generales (cont.)

- Una herramienta de software que automatiza la construcción de analizadores léxicos, permite que personas con diferentes conocimientos utilicen la concordancia de patrones en sus propias áreas de aplicación.
- La gran ventaja de un generador de analizadores léxicos es que puede utilizar los algoritmos más conocidos de concordancia de patrones, con lo cual crea analizadores léxicos eficientes para los no especialistas en dichas técnicas.

Algunas Reglas en RI

- *Dígitos*. En general números no son buenos candidatos de palabras claves. Sin embargo, la normalización de ciertos números en el contexto de ciertas palabras pueden ser relevantes para la recuperación de información.
- *Guiones*. Puede que sea o no sea relevante la eliminación de guiones. En general, se adopta una regla y se agregan excepciones.
- *Tildes y caracteres especiales*. Puede que sea o no sea relevante la eliminación de tildes y caracteres especiales. En general, se adopta una regla y se agregan excepciones.

Algunas Reglas en RI (cont.)

- *Las etiquetas HTML* por lo general son removidas.
- *Los signos de puntuación* son generalmente removidos.
- *Generalmente el texto es transformado a mayúscula o minúscula.*

Algunas Reglas en RI (cont.)

- *Dígitos:*
  - Generalmente son palabras poco específicas.
  - Hay casos en que si son importantes:
    - Vitamina B12
    - Efecto 2000
  - No hay una solución general pues depende del dominio.
  - En colecciones generales se consideran los que empiezan por letra:
    - 2000 NO
    - B12 SI

Algunas Reglas en RI (cont.)

- *Guiones:*
  - Al final de palabra se puede eliminar y juntar los dos trozos.
  - Separador:
    - Es bueno que siga junto: F-16.
    - Son dos palabras: Jean-Claude.
  - En un dominio específico tenemos otros elementos de juicio.
  - En dominios generales podemos usar:
    - F no es término y 16 no es término. Por tanto F-16 es término.
    - Jean es término y Claude es término. Por tanto "–" es un separador.

Algunas Reglas en RI (cont.)

- *Otros signos:*
  - Puntuación, barras, etc.
  - En la mayoría de los casos son separadores.
  - Aunque no siempre:
    - Ej.: OS/2
- *Mayúsculas/Minúsculas:*
  - Normalmente da lo mismo.
  - Se suelen poner todo en mayúsculas o todo en minúsculas.

Algunas Reglas en RI (cont.)

- *Palabra:*
  - Regla general (colección genérica):
    - Palabra: Toda cadena de caracteres alfanuméricos o numéricos en que el primer carácter es una letra. Todos los caracteres se pasan a mayúsculas (o minúsculas) y todos los demás son separadores.
  - En dominios específicos se tiene una lista de términos candidatos.
    - Ej.: B12 no lo separes
- En las preguntas se suele hacer el mismo análisis léxico:
  - Ej.: "Quiero documentos que tengan 13 **animales** de **pelo verde** y **rojo**".

Componentes Léxicos, Patrones y Lexemas

- Se dice que hay un conjunto de cadenas en la entrada para el cual se produce como salida el mismo componente léxico. Este conjunto de cadenas se describe mediante una regla llamada *patrón al componente léxico*. Se dice que el patrón concuerda con cada cadena del conjunto.
- Un *lexema* es una secuencia de caracteres en el archivo fuente, el cual debe concordar con el patrón para un componente léxico.
- Los *componentes léxicos* se tratan como símbolos terminales de la gramática del archivo fuente. Los lexemas para el componente léxico que concuerden con el patrón representan cadenas de caracteres en el archivo fuente.

Componentes Léxicos, Patrones y Lexemas (cont.)

- Un *patrón* es una regla, la cual describe el conjunto de lexemas que pueden representar a un determinado componente léxico en los archivos fuentes.
- Cuando concuerda un lexema con un patrón, el analizador léxico, proporcionara información adicional sobre el lexema concreto que concordó.
- Las expresiones regulares son una notación importante para especificar patrones. Cada patrón concuerda con una serie de cadenas, de modo que las expresiones regulares sirvan como nombres para conjuntos de cadenas.

Componentes Léxicos, Patrones y Lexemas (cont.)

- *Alfabeto* o *clase de carácter* denota cualquier conjunto finito de símbolos.
- *Cadena* es una secuencia finita de símbolos tomados de un alfabeto.
- Los términos *frase* o *palabra* a menudo se utilizan como sinónimos del termino cadena.

Analizador Léxico JFlex

- JFlex es un generador de analizadores léxicos para Java y está escrito en Java.
- Instalación:
  - Unzip los archivos zip del JFlex y Java_Cup en el directorio que guste
  - Crear las variables de ambiente:
    - JAVA_HOME = C:\j2sdk1.4.2_05
    - JFLEX_HOME = C:\jflex-1.4
    - JAVACUP_HOME = C:\javacup
  - Agregar a la variable de ambiente PATH
    - %JAVA_HOME%\bin;%JFLEX_HOME%\bin
  - Agregar a la variable de ambiente CLASSPATH
    - %JFLEX_HOME%\lib\JFlex.jar
    - %JAVACUP_HOME%\java_cup.jar

Analizador Léxico JFlex (cont.)

- Para correr JFlex ejecute la siguiente instrucción:
  - java JFlex.Main <options> <inputfiles>
- Un archivo JFlex está organizado en tres secciones, separado por directivas de porcentaje ("%%").
- Una especificación de JFlex apropiada tiene el siguiente formato:

  Código del usuario

  %%

  Directivas JFlex

  %%

  Reglas de expresiones regulares

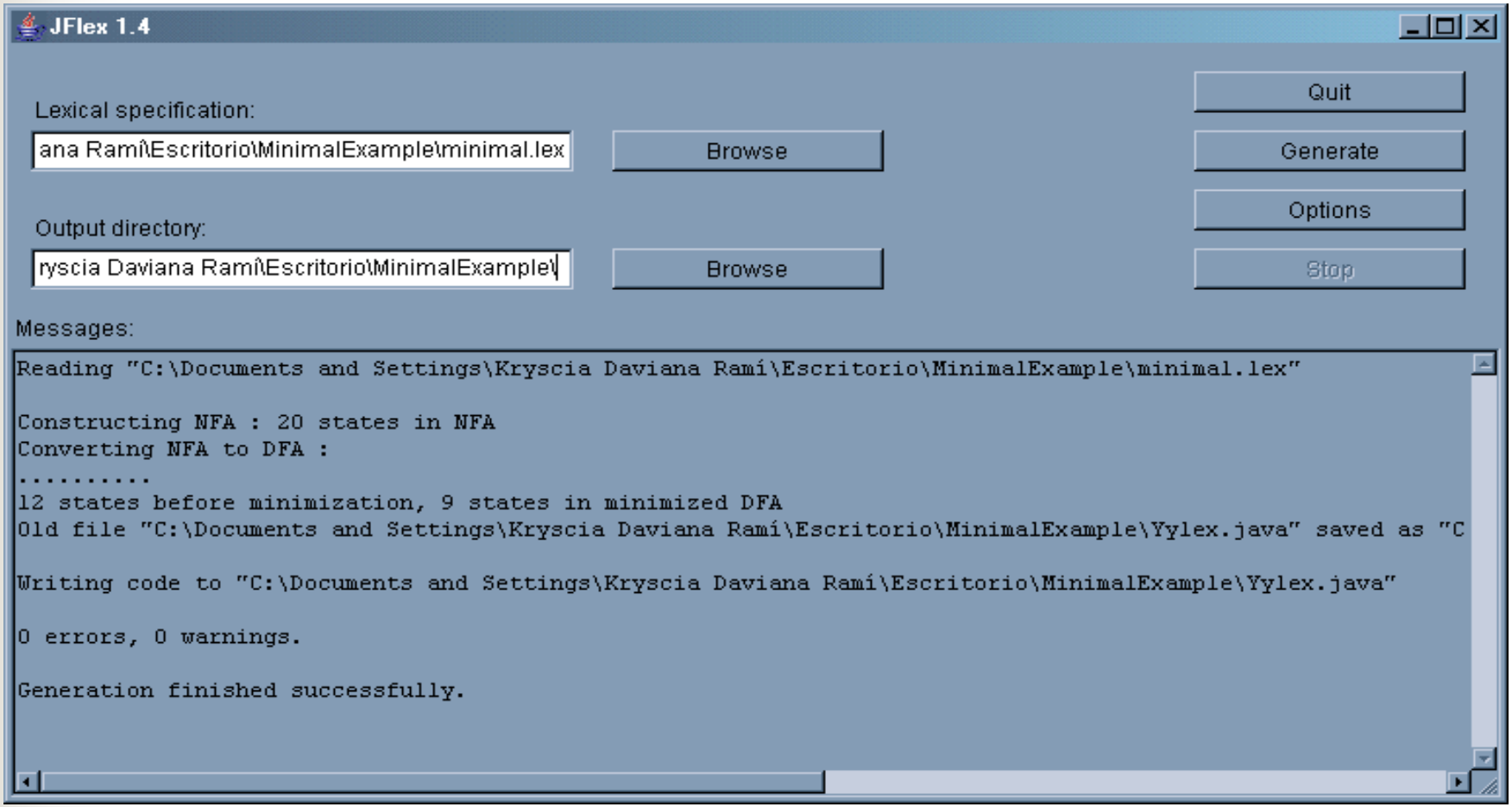
Analizador Léxico JFlex (cont.)

- Las directivas "%%" distinguen las secciones del archivo y van al principio de la línea. El resto de la línea contenida en "%%" pueden desecharse y no deben usarse para alojar declaraciones adicionales o código.
- La sección del código del usuario es copiada directamente en el archivo resultante. Este área de la especificación provee espacio para la implementación de clases o tipos retornados.
- La sección de directivas JFlex es donde se dan las definiciones de los macros y se declaran nombres de estado.
- La tercer sección contiene las reglas del analizador léxico, cada una consiste en tres partes: una lista de estados optativa, una expresión regular y una acción. El formato es el siguiente: [<*states*>] <*expression*> { <*action*> }

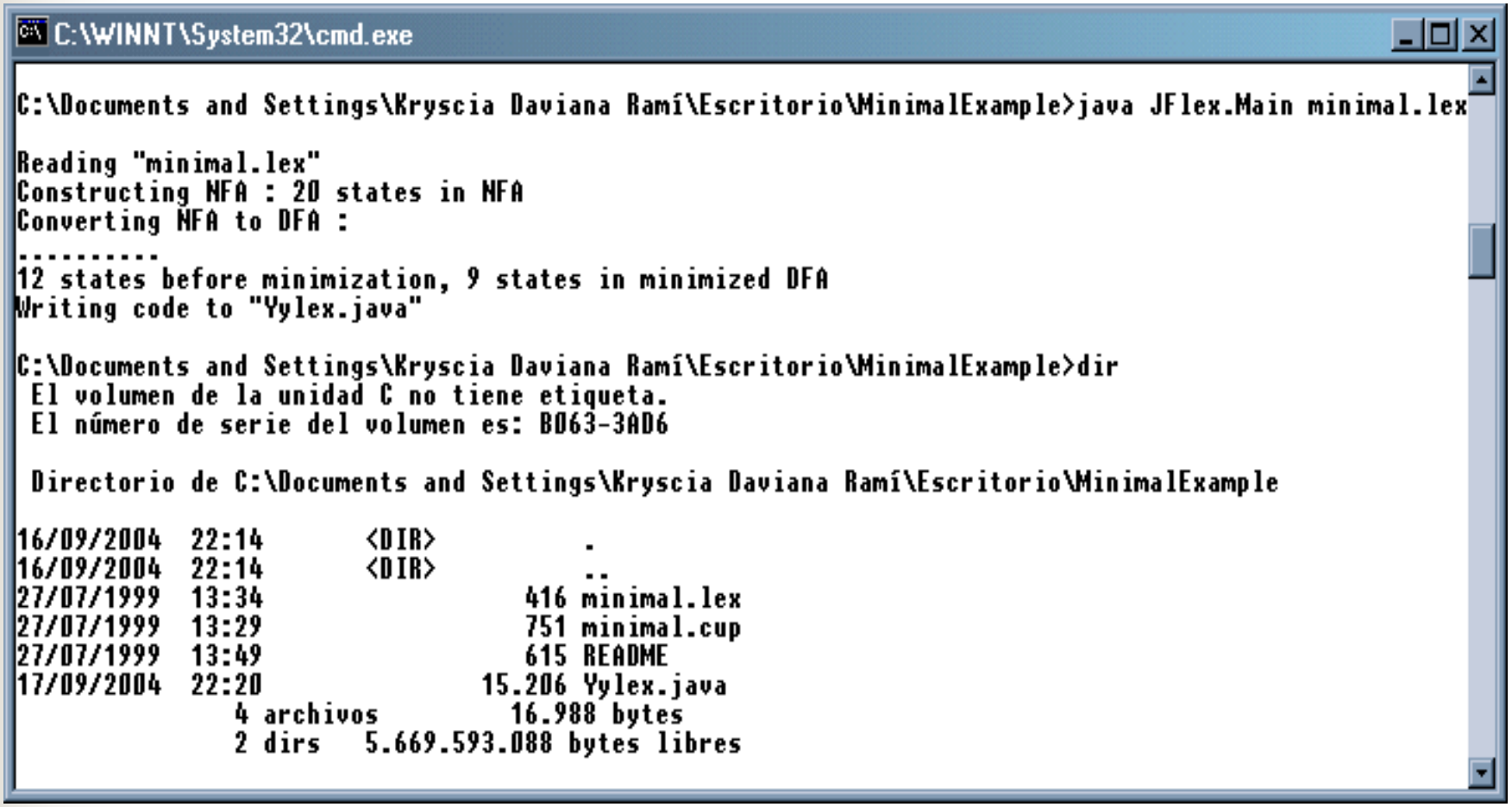
Analizador Léxico JFlex (cont.)

- Se muestra un pequeño ejemplo donde se utiliza JFlex y Java CUP. Este ejemplo requiere la versión de JFlex 1.2.5 o mayor y la versión de Java Cup 0.10.
- El ejemplo es una simple calculadora, solamente suma y multiplica; recibe de entrada diferentes chars.
- Para compilar y construir el ejemplo se ejecutan las siguientes instrucciones:
  - java JFlex.Main minimal.lex
  - java java_cup.Main < minimal.cup
  - javac -d . parser.java sym.java Yylex.java

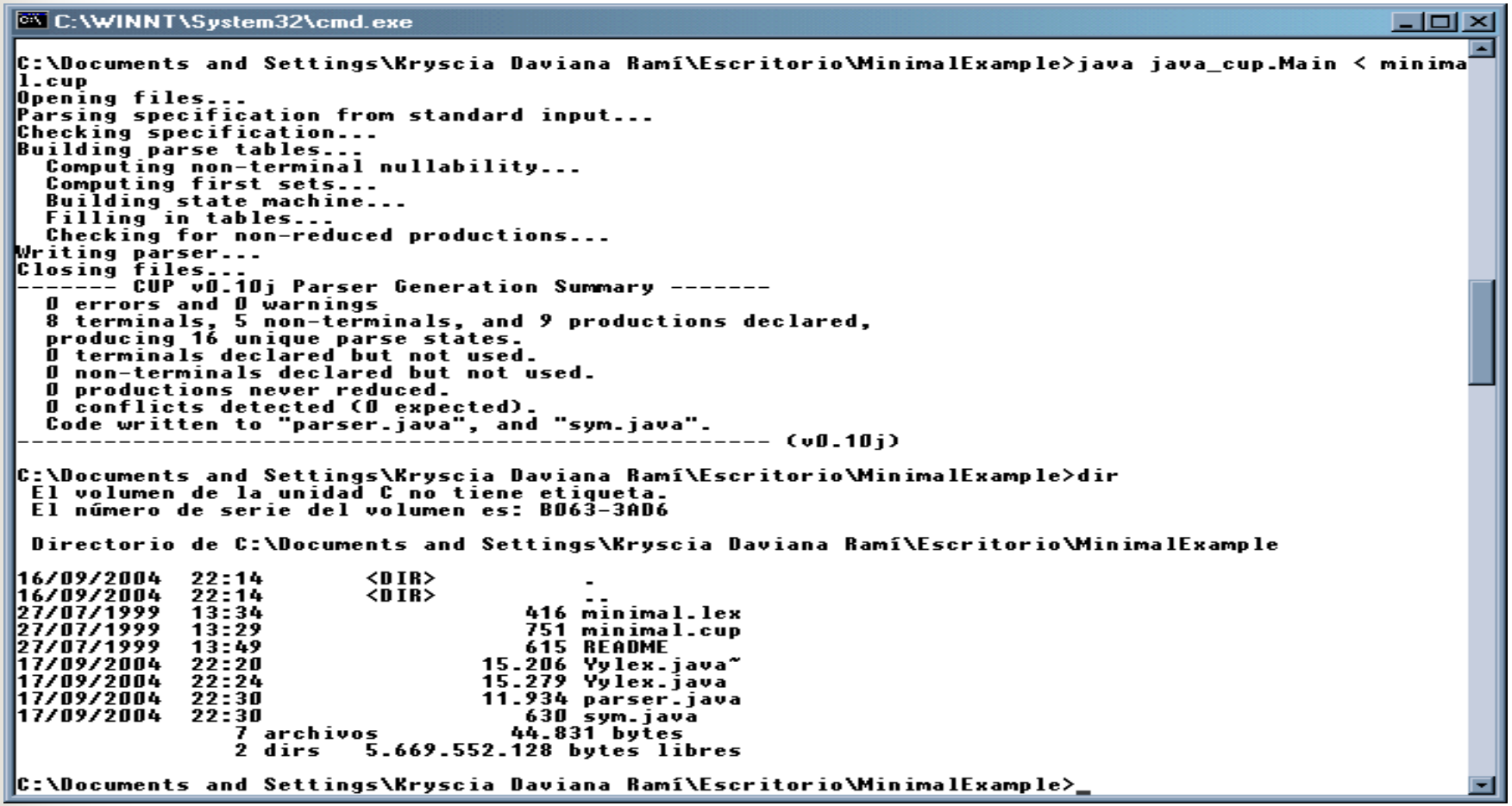
Analizador Léxico JFlex (cont.)



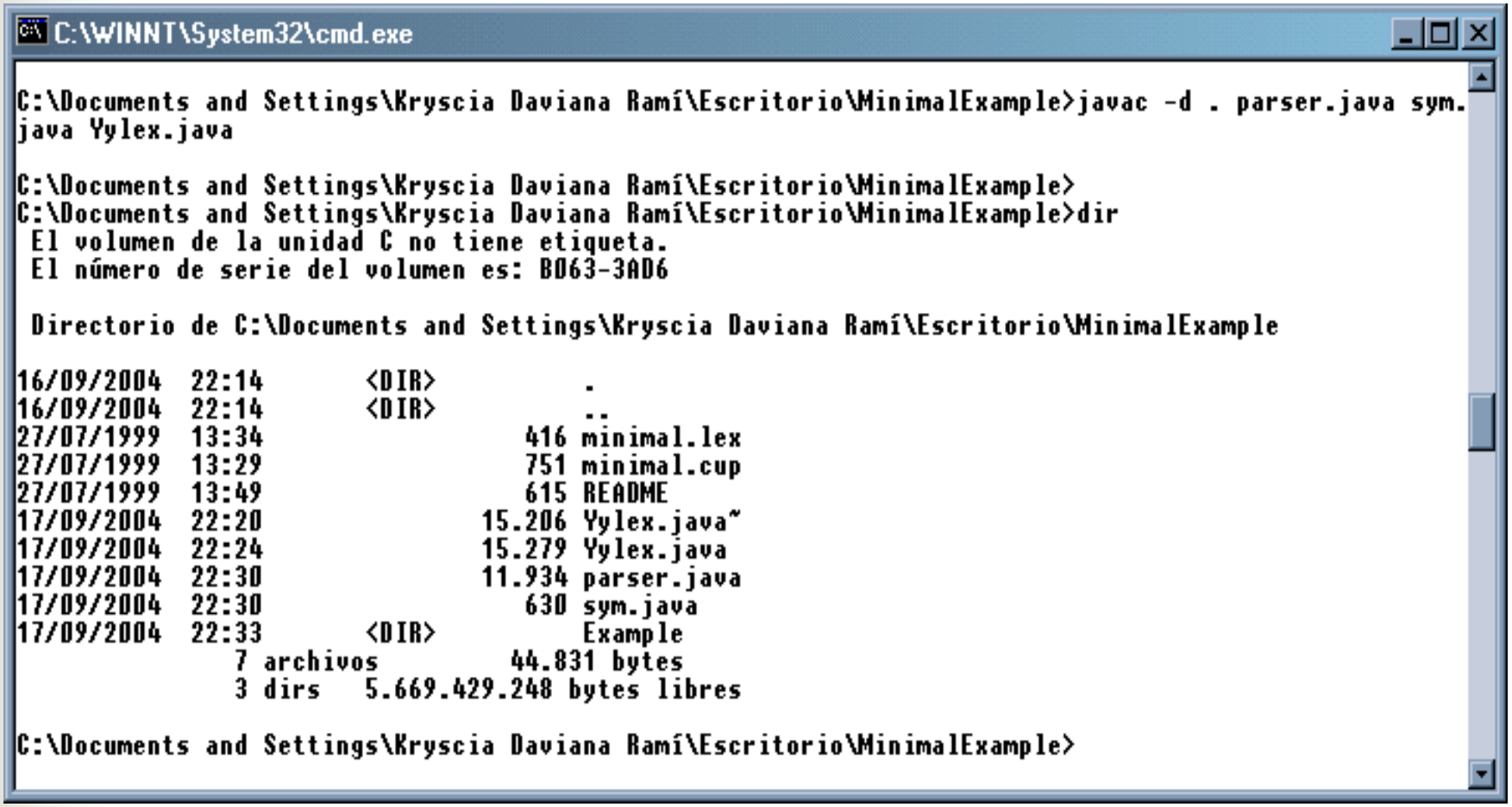
Analizador Léxico JFlex (cont.)



Analizador Léxico JFlex (cont.)

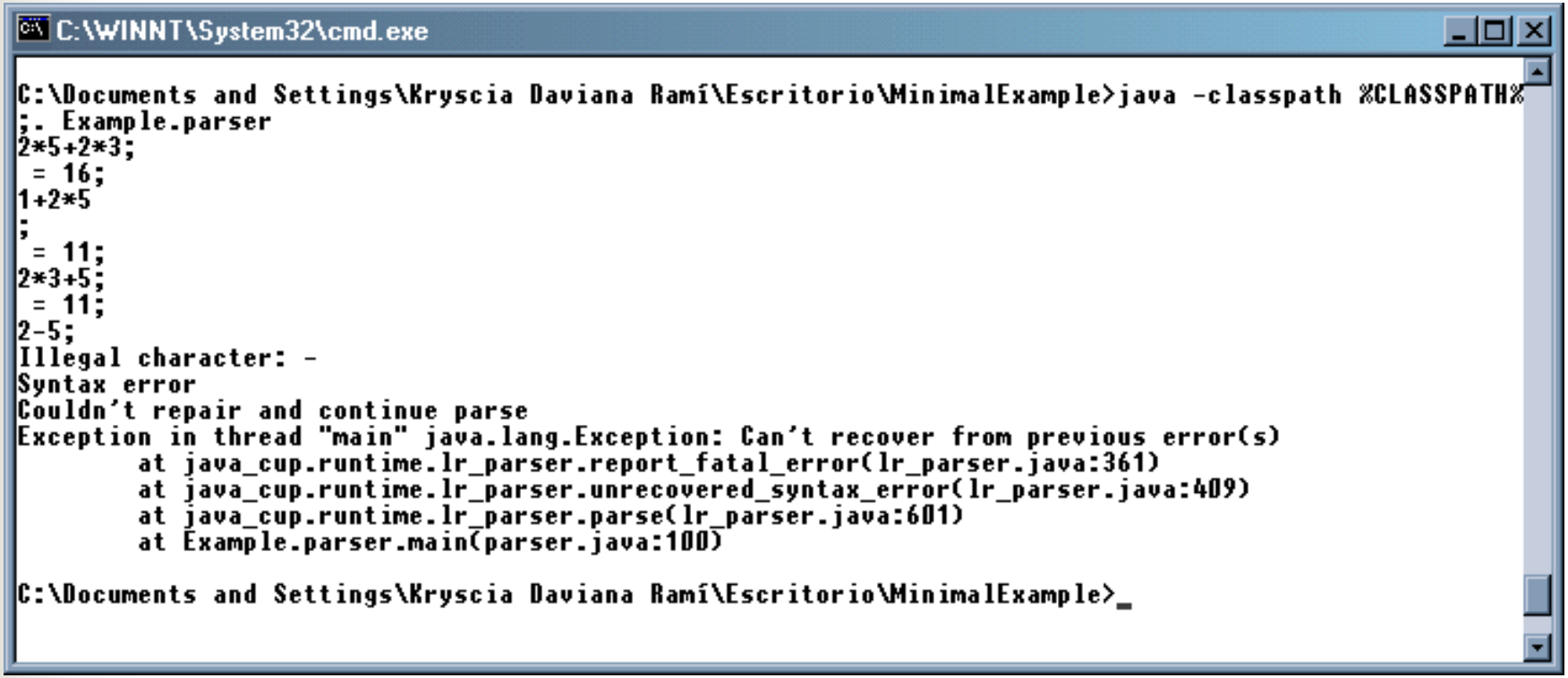


Analizador Léxico JFlex (cont.)



Analizador Léxico JFlex (cont.)

- Para correrlo se realiza lo siguiente:
  - java -classpath %CLASSPATH%;. Example.parser



Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.
  - http://www.gedlc.ulpgc.es/docencia/seminarios/rit/.