

UN GENERADOR DE PARSERS PARA MICROS

Agustín Sánchez Mambrilla
Grupo Golem de Valencia

En el grupo GOLEM hemos desarrollado un sistema para generar gramáticas libres de contexto que puede ser utilizado como herramienta en las asignaturas lingüísticas de bachillerato, donde la relación entre lenguaje y computación es prácticamente desconocida.

El paquete tiene las siguientes prestaciones:

- Permite definir gramáticas libres de contexto.
- Genera aleatoriamente oraciones y subestructuras propias de la gramática definida.
- Genera el parser que permite analizar oraciones de dicha gramática, confeccionando el árbol morfosintáctico correspondiente.
- Permite visualizar y modificar cómodamente las estructuras definidas, proporcionando el vocabulario utilizado, y las estructuras existentes.
- Procesa las contracciones AL y DEL y los signos de puntuación, produciendo las separaciones necesarias.
- Facilita el almacenamiento y lectura de las gramáticas en diskette.
- Produce salida por pantalla y/o impresora

El programa está escrito en LOGO, y rueda en IBM PC o compatible. Su tamaño es de 16 kb y requiere una memoria mínima de 192 kb.

FUNCIONAMIENTO

Para definir una gramática el programa parte de una única estructura inicial, DRACION, cuyas partes son especificadas por el usuario de forma interactiva.

Al dar la información al programa, los terminales se distinguen de las estructuras porque van precedidos por comillas. Cada vez que nombramos una nueva estructura, el programa pregunta por su composición, admitiendo varias posibilidades alternativas.

A medida que se va definiendo la gramática se van generando los procedimientos que realizan el matching con las diferentes estructuras, confeccionando un parser que da una ejecución como la que mostramos más adelante.

UTILIZACION DIDACTICA

Estamos desarrollando con el generador ejemplos en Latín y en lenguas modernas, para determinar su utilidad como herramienta didáctica. Por el momento los resultados son alentadores, ya que los usuarios se divierten a la vez que se ven obligados a reflexionar sobre las estructuras gramaticales cuando la generación de oraciones o el parsing no dan los resultados previstos.

En este sentido las limitaciones del sistema son, quizá, una ventaja, ya que es responsabilidad del usuario jerarquizar las diferentes alternativas de cada estructura. Por ejemplo, en la gramática de los NÚMEROS, dentro de la estructura HASTA.MIL1 la subestructura CENTENAS HASTA.CIEN1 ha de ir necesariamente antes de CENTENAS, que es menos exigente, ya que de lo contrario el matching tendría éxito prematuramente, evolucionando el control al procedimiento que busca "MIL. Este podría encontrarse con HASTA.CIEN1 y fracasaría.

Igualmente, al no manejar raíces ni sufijos, sino solamente palabras enteras, para concordar géneros y números hay que diseñar con sumo cuidado la estructura gramatical.

La mecánica del matching genera abundantes mensajes de error no significativos, ya que es posible que tengan éxito posteriores alternativas. Por ello se suele trabajar con los mensajes inhibidos a menos que fracase la totalidad del análisis.

SEMANTICA

Hay otra versión que procesa también la semántica de las oraciones, entendiendo por significado de una estructura cualquier combinación de los siguientes elementos:

- Un literal (precedido por comillas)
- Un nombre de variable en la sintaxis de LOGO
- Una primitiva del lenguaje LOGO
- Un nombre de procedimiento
- Un número

con tal de que sea significativa para el sistema LOGO anfitrión.

El manejo de este sistema requiere algunos conocimientos de la sintaxis LOGO, por lo que su utilización es algo más restrictiva.

La principal aplicación de este sistema es la de crear con facilidad programas en los que la comunicación hombre-máquina se haga en lenguaje natural escrito.

EL MUNDO DE LOS BLOQUES

Como ejemplo tomemos una versión reducida de la gramática del Mundo de los Bloques de Winograd, en la que pueden procesarse oraciones como:

PON EL ROJO GRANDE A LA DERECHA DEL VERDE PEQUEÑO

¿QUE BLOQUE ESTA A LA DERECHA DEL ROJO PEQUEÑO?

¿ESTA EL BLOQUE GRANDE VERDE A LA DERECHA DEL ROJO PEQUEÑO?

¿POR QUE PUSISTE EL BLOQUE QUE ESTA A LA DERECHA DEL ROJO PEQUEÑO A LA IZQUIERDA DEL GRANDE VERDE?

La estructura gramatical sería:

ORACION es:

- 1 IMPERATIVA
- 2 INTERROGATIVA

IMPERATIVA es:

- 1 VERBO.IMPERATIVO FRASE.NOMINAL CALIFICATIVO

VERBO.IMPERATIVO es:

- 1 "PON

FRASE.NOMINAL es:

- 1 ARTICULO GRUPO.NOMINAL

ARTICULO es:

- 1 "EL

GRUPO.NOMINAL es:

- 1 NOMBRE ADJETIVOS
- 2 ADJETIVOS
- 3 NOMBRE FRASE.RELATIVO
- 4 FRASE.RELATIVO

NOMBRE es:

- 1 "BLOQUE

ADJETIVOS es:

- 1 COLOR TAMAÑO
- 2 TAMAÑO COLOR

COLOR es:

- 1 "ROJO
- 2 "VERDE

TAMAÑO es:

- 1 "GRANDE
- 2 "PEQUEÑO

FRASE.RELATIVO es:

1 "QUE VERBO.COPULATIVO CALIFICATIVO

VERBO.COPULATIVO es:

1 "ESTA

CALIFICATIVO es:

1 "A "LA LUGAR "DE FRASE.NOMINAL

LUGAR es:

1 "IZQUIERDA

2 "DERECHA

INTERROGATIVA es:

1 "& INTERROGACION "?

INTERROGACION es:

1 "QUE NOMBRE VERBO.COPULATIVO CALIFICATIVO

2 VERBO.COPULATIVO FRASE.NOMINAL CALIFICATIVO

3 "POR "QUE PRETERITO FRASE.NOMINAL CALIFICATIVO

PRETERITO es:

1 "FUSISTE

Las estructuras pueden ser recursivas. Así FRASE.NOMINAL contiene a CALIFICATIVO, y este a FRASE.NOMINAL, lo que permite procesar oraciones de complejidad ilimitada (...EL BLOQUE QUE ESTA A LA DERECHA DEL QUE ESTA A LA IZQUIERDA DEL...)

La ejecución del parsing da una salida como la que sigue:

PON EL BLOQUE ROJO PEQUEÑO A LA DERECHA DEL ROJO GRANDE

PON

VERBO.IMPERATIVO

EL

ARTICULO

BLOQUE

NOMBRE

ROJO

COLOR

PEQUEÑO

TAMANO

ROJO PEQUEÑO

ADJETIVOS

BLOQUE ROJO PEQUEÑO

GRUPO.NOMINAL

EL BLOQUE ROJO PEQUEÑO

FRASE.NOMINAL

DERECHA

LUGAR

EL

ARTICULO

ROJO

COLOR

GRANDE

TAMANO

ROJO GRANDE

ADJETIVOS

ROJO GRANDE

GRUPO.NOMINAL

EL ROJO GRANDE

FRASE.NOMINAL

A LA DERECHA DE EL ROJO GRANDE

CALIFICATIVO

PON EL BLOQUE ROJO PEQUEÑO A LA DERECHA DE EL ROJO GRANDE

IMPERATIVA

PON EL BLOQUE ROJO PEQUEÑO A LA DERECHA DE EL ROJO GRANDE

ORACION CORRECTA

Pueden activarse mensajes de error, mostrando este aspecto

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
PON

VERBO.IMPERATIVO

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
EL

ARTICULO

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
BLOQUE

NOMBRE

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
ROJO

COLOR

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE

No encuentro GRANDE

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
PEQUEÑO

TAMAÑO

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
ROJO PEQUEÑO

ADJETIVOS

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
BLOQUE ROJO PEQUEÑO

GRUPO.NOMINAL

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE
EL BLOQUE ROJO PEQUEÑO

FRASE.NOMINAL

PON EL BLOQUE ROJO PEQUEÑO LA DERECHA DE EL ROJO GRANDE

No encuentro A

.....

ORACION INCORRECTA

Hemos utilizado esta gramática para desarrollar una versión reducida del SHRDLU de Winograd, que en un universo formado por cuatro bloques rotados de COLOR, TAMAÑO y POSICION, contesta a las preguntas anteriormente escritas y ejecuta los movimientos que se le ordenan, resolviendo el eventual problema de falta de sitio entre bloques mediante movimientos preliminares.

NUMEROS

Otro problema clásico es la gramática de los números, mucho más complicada en español que en inglés.

En efecto, decimos MIL CIEN y CIEN MIL pero también CIEN QUINCE. Diferentes problemas crean el UNO y el VEINTIUNO:

VEINTIUN MIL / MIL VEINTIUNO
NOVENTA Y UN MILLONES / CIENTO NOVENTA Y UNO

Hasta un millón, la estructura gramatical sería:

ORACION es:

1 HASTA.UN.MILLON

HASTA.UN.MILLON es:

1 HASTA.MIL1 "MIL HASTA.MIL2

2 HASTA.MIL1 "MIL

3 "MIL HASTA.MIL2

4 "MIL

5 HASTA.MIL2

HASTA.MIL1 es:

1 "CIENTO HASTA.CIEN1

2 CENTENAS HASTA.CIEN1

3 CENTENAS

4 HASTA.CIEN1

5 "CIEN

HASTA.CIEN1 es:

1 DECENAS "Y UNIDADES1

2 DECENAS

3 UNIDADES.N

4 "DIEZ

5 "VEINTE

6 HASTA.VEINTE

7 HASTA.TREINTA1

DECENAS es:

1 "TREINTA

.....
7 "NOVENTA

UNIDADES1 es:
1 "UN
2 UNIDADES.N

UNIDADES.N es:
1 "DOS
.....
8 "NUEVE

HASTA.VEINTE es:
1 "ONCE
.....
9 "DIECINUEVE

HASTA.TREINTA1 es:
1 "VEINTIUN
2 HASTA.TREINTA.N

HASTA.TREINTA.N es:
1 "VEINTIDOS
.....
8 "VEINTINUEVE

CENTENAS es:
1 "DOSCIENTOS
.....
8 "NOVECIENTOS

HASTA.MIL2 es:
1 "CIENTO HASTA.CIENZ
2 CENTENAS HASTA.CIENZ
3 CENTENAS
4 HASTA.CIENZ
5 "CIEN

HASTA.CIENZ es:
1 DECENAS "Y UNIDADES2
2 DECENAS
3 UNIDADES2
4 "DIEZ
5 HASTA.VEINTE
6 HASTA.TREINTA2

UNIDADES2 es:
1 "UNO
2 UNIDADES.N

HASTA.TREINTA2 es:
1 "VEINTIUNO
2 HASTA.TREINTA.N