El diccionario electrónico español

Carlos Subirats Rüggerberg

Universidad Autónoma de Barcelona

EL DICCIONARIO ELECTRONICO DEL ESPAÑOL*

1. CARACTERISTICAS GENERALES

Un diccionario electrónico consiste en una lista de palabras con información gramatical codificada, la cual puede ser utilizada por programas de generación automatizada de las formas flexivas de la lengua. Dicha información gramatical consiste en una especificación codificada de la categoría gramatical de cada una de las entradas, así como de sus correspondientes propiedades morfológicas de flexión, en el caso de las palabras que son susceptibles de poseerlas, como los verbos, los nombres y los adjetivos. La especificación de las propiedades morfológicas de flexión se hace mediante un código alfanumérico que remite a unos algoritmos de flexión morfológica; así, la aplicación de los programas morfológicos de flexión permite generar automáticamente todas las formas flexivas de la lengua, con una especificación de un conjunto de propiedades morfológicos, a partir de un número limitado de formas de base, que son las únicas que contiene el diccionario electrónico. La obtención de las formas flexivas mediante programas, y no mediante un listado, tiene una importancia crucial para un diccionario electrónico del español, ya que la abundancia de formas flexivas que presenta esta lengua hace que el número de formas de base se multiplique por un factor 10, tras la aplicación de los programas de generación morfológica de formas flexivas a verbos, nombres y adjetivos.

(*) Quisiera dar las gracias a José M. Blecua, a Annibale Elia y a todos los miembros del CCUAB, por la ayuda que nos han proporcionado. La colaboración de Blandine Courtois, así como la de todos los demás miembros del LADL, ha sido crucial. La realización de este proyecto no hubiese sido posible sin la ayuda de Maurice Gross.

Este proyecto se está realizando con una subvención de la CAICYT (PB85-0371-C02-01) y una ayuda de Acción Integrada Hispano-Francesa (9/322).

Es importante aclarar la ambigüedad que entraña la utilización del término "electrónico" o "informático", aplicado a un diccionario. En la actualidad, prácticamente todos los diccionarios modernos están disponibles en soporte magnético, ya que los sistemas de fotocomposición modernos se realizan con medios informáticos. En efecto, la confección de los originales de los textos e, incluso, la realización del montaje de la edición se pueden llevar a cabo sobre terminales especiales que generan una cinta magnética en la que se encuentra el texto, así como los códigos especiales de fotocomposición, los cuales varían en función de las características concretas del entorno informático que se utilice, así como de las características específicas del programa de fotocomposición. No obstante, una vez eliminados los códigos de fotocomposición, estas cintas magnéticas no contienen nada más que el texto íntegro del diccionario. Así, el hecho de que el texto de un diccionario esté disponible sobre soporte informático no conlleva ninguna alteración sustancial con respecto a las características del texto original, el cual está pensado única y exlusivamente para ser utilizado por lectores humanos. Así, una cinta magnética del texto de un diccionario no tiene ninguna utilidad para la realización de un diccionario de aplicaciones informáticas. En resumen, un diccionario electrónico para aplicaciones informáticas y un diccionario usual --incluso un diccionario usual preparado para ser consultado electrónica o telemáticamente-- poseen unas características que los convierten en objetos radicalmente distintos.

Es importante, asimismo, aclarar las diferencias que existen entre un diccionario electrónico y los diccionarios que utilizan algunos procesadores o programas de tratamiento de textos, los cuales están destinados, fundamentalmente, a servir de base para la corrección ortográfica automática. Estos diccionarios ortográficos presentan una serie de diferencias importantes con respecto al diccionario electrónico del español que estamos realizando en la UAB. En primer lugar, señalemos que el número de entradas que poseen los diccionarios ortográficos es muy reducido en relación a la talla real del léxico de una lengua. Por otro lado, las entradas de los diccionarios ortográficos no poseen información ni sobre la categoría gramatical a la que pertenece cada una de ellas, ni tampoco sobre otras propiedades morfológicas, como la flexión, etc.; ello es

debido a que dichos diccionarios están concebidos exclusivamente como ayudas para los tratamientos de textos, y, por lo tanto, no está prevista, por ejemplo, la aplicación de programas de generación automatizada de formas flexivas, ni tampoco la utilización de dichos diccionarios para aplicaciones tecnológicas más avanzadas, como por ejemplo, la generación o el análisis automático de textos, etc. Por el contrario, el diccionario electrónico del español, no solo posee información codificada con respecto a la categoría gramatical a la que pertenece cada una de sus entradas, sino también de las propiedades de flexión y aun sobre otras propiedades morfológicas, como el género gramatical, etc.

En resumen, los diccionarios ortográficos, cuya talla léxica, como ya hemos señalado, es muy limitada en relación al léxico real de una lengua, están pensados para ser utilizados como complementos para los programas de tratamiento de textos y consitituyen, por lo tanto, objetos radicalmente distintos a los diccionarios electrónicos.

2. LOS PROGRAMAS DE GENERACION MORFOLOGICA AUTOMATIZADA DEL DICCIONARIO ELECTRONICO DEL ESPAÑOL

En el paquete de programas de generación morfológica, hay que distinguir, por un lado, los programas de generación de las formas nominales y adjetivas y, por otro, los programas de generación de las formas verbales. El conjunto de programas de generación de las formas nominales y adjetivas consta de un fichero con los algoritmos de flexión y del programa de generación propiamente dicho. El trabajo de implementación del fichero de algoritmos y del mencionado programa está en curso de elaboración.

El paquete de programas de generación de las formas verbales está formado por dos ficheros de datos y tres programas. Los ficheros de datos son los siguientes:

(1) un fichero donde se especifican todas las desinencias verbales, establecidas y clasificadas en función de los tiempos verbales a las que pertenecen;

(2) un fichero con los algoritmos de la conjugación, tanto regular como irregular.

Los programas que integran el paquete de generación de las formas verbales son los siguientes:

(1) un programa de pretratamiento que actúa sobre el fichero de desinencias, generando un nuevo fichero que es el que realmente utiliza el programa de generación de formas verbales;
(2) un programa de pretratamiento que actúa sobre el fichero de conjugaciones y que, al igual que en el caso anterior, genera otro fichero, el cual es el que utiliza el programa de generación de formas verbales;
(3) el programa de generación de las formas verbales propiamente dichas, el cual interpreta la codificación alfanumérica de las entradas verbales del diccionario electrónico para proceder a su conjugación automática;

El programa de generación verbal actúa, en primer lugar, seleccionando las entradas del diccionario que llevan el código alfabético correspondiente a los verbos, es decir, las que están marcadas con el código *V* e interpreta, además, la codificación numérica --que va junto a la alfabética--, la cual remite al fichero de conjugaciones; a partir del algoritmo de la conjugación especificada en dicho fichero, el programa hace el cálculo de las raíces, que combina con las desinencias correspondientes, tras efectuar la consulta al fichero de desinencias verbales.

Un ejemplo del resultado final de la aplicación del programa de generación de las formas verbales, lo podemos observar en las *figuras 1* y *2*. Como se puede observar en dichas figuras, cada una de las formas verbales generadas lleva una especificación de la relación que mantiene con la forma de base del diccionario electrónico, es decir, con el infinitivo, y, además, una especificación de las categorías de *tiempo*, *modo*, *persona* y *número*. El *tiempo/modo* se represeta mediante un código alfabético, concretamente, mediante una letra mayúscula y, a su vez, la *persona/número* se representa mediante un código alfanumérico. El *Programa de generación de formas verbales*, el *Programa de pretratamiento del*

fichero de desinencias verbales, así como el *Programa de pretratamiento del fichero de conjugaciones* han sido realizados por Blandine Courtois en el LADL (Université Paris VII).

3. INVESTIGACION TEORICA Y APLICACIONES TECNOLOGICAS DEL DICCIONARIO ELECTRONICO

La creación de un diccionario electrónico del español de las características del nuestro permitirá realizar, por primera vez, tanto aplicaciones de tipo tecnológico, como investigaciones de tipo teórico *en las que se podrá trabajar sobre un léxico exhaustivo del español*.

Desde el punto de vista de sus aplicacones tecnológicas, el diccionario electrónico del español puede constituir, por ejemplo, una base de datos léxicos para la realización de un diccionario de corrección ortográfica, que constituya un producto *fiable*, es decir, un producto con un porcentaje de error nulo. En efecto, el porcentaje de error de un diccionario de corrección ortográfica viene dado por el número de entradas que se pueda encontrar al procesar un texto y que, a su vez, no estén especificadas en dicho diccionario. No obstante, si se parte de un diccionario como el nuestro que posee un léxico exhaustivo de la lengua, su porcentaje de error, es decir, la posibilidad de que se pueda encontrar alguna forma que no esté especificada en él, sería nula.

Desde el punto de vista del estudio teórico, del mismo modo que la investigación léxico-gramatical, centrada en la creación de una gramática electrónica del español (cf. Subirats 1986), ha permitido demostrar la inadecuación empírica de las hipótesis pretendidamente "explicativas" de la gramática generativa y ha obligado, incluso, a revisar el propio concepto de *regla* en la gramática (cf. Gross, 1975 y Harris, 1982), así también la investigación de la estructura de la palabra, partiendo de una consideración exhaustiva del léxico español, permitirá establecer una teoría morfológica realmente *explicativa*, que se aparte radicalmente de las especulaciones escolásticas de la morfológía generativa, que --al igual que la

sintaxis generativa-- trata de establecer principios "generales" e hipótesis "explicativas" a partir de la consideración de un número absolutamente insuficiente de elementos léxicos.

```
estar,**********************.V:W00
estando,**********************-3r.V:G00
estado,**********************-2r.V:Kms
estoy,**********************-2ar.V:P1s
estás,**********************-2ar.V:P2s
está,**********************-1ar.V:P3s:Y2s
estamos,**********************-3r.V:P1p
estáis,**********************-3ar.V:P2p
están,**********************-2ar.V:P3p
estaba,**********************-2r.V:I1s:I3s
estabas,**********************-3r.V:I2s
estábamos,**********************-6ar.V:I1p
estabais,**********************-4r.V:I2p
estaban,**********************-3r.V:I3p
estuve,**********************-3ar.V:J1s
estuviste,**********************-6ar.V:J2s
estuvo,**********************-3ar.V:J3s
estuvimos,**********************-6ar.V:J1p
estuvisteis,**********************-8ar.V:J2p
estuvieron,**********************-7ar.V:J3p
estaré,**********************-1.V:F1s
estarás,**********************-2.V:F2s
estará,**********************-1.V:F3s
estaremos,**********************-4.V:F1p
estaréis,**********************-3.V:F2p
estarán,**********************-2.V:F3p
esté,**********************-1ar.V:S1s:S3s
estés,**********************-2ar.V:S2s
estemos,**********************-4ar.V:S1p
estéis,**********************-3ar.V:S2p
estén,**********************-2ar.V:S3p
estuviera,**********************-6ar.V:T1s:T3s
estuvieras,**********************-7ar.V:T2s
estuviéramos,**********************-9ar.V:T1p
estuvierais,**********************-8ar.V:T2p
estuvieran,**********************-7ar.V:T3p
estuviese,**********************-6ar.V:U1s:U3s
estuvieses,**********************-7ar.V:U2s
estuviésemos,**********************-9ar.V:U1p
estuvieseis,**********************-8ar.V:U2p
estuviesen,**********************-7ar.V:U3p
estad,**********************-1r.V:Y2p
estaría,**********************-2.V:C1s:C3s
estarías,**********************-3.V:C2s
estaríamos,**********************-5.V:C1p
estaríais,**********************-4.V:C2p
estarían,**********************-3.V:C3p
```

Figura 1

Conjugación automática del verbo *estar*

```
ir,**************************.V:W00
yendo,************************-5ir.V:G00
ido,*************************-2r.V:Kms
voy,*************************-3ir.V:P1s
vas,*************************-3ir.V:P2s
va,**************************-2ir.V:P3s
vamos,***********************-5ir.V:P1p
vais,************************-4ir.V:P2p
van,*************************-3ir.V:P3p
iba,*************************-2r.V:I1s:I3s
ibas,************************-3r.V:I2s
íbamos,**********************-6ir.V:I1p
ibais,***********************-4r.V:I2p
iban,************************-3r.V:I3p
fui,*************************-3ir.V:J1s
fuiste,**********************-6ir.V:J2s
fue,*************************-3ir.V:J3s
fuimos,**********************-6ir.V:J1p
fuisteis,********************-8ir.V:J2p
fueron,**********************-6ir.V:J3p
iré,*************************-1.V:F1s
irás,************************-2.V:F2s
irá,*************************-1.V:F3s
iremos,**********************-4.V:F1p
iréis,***********************-3.V:F2p
irán,************************-2.V:F3p
vaya,************************-4ir.V:S1s:S3s
vayas,***********************-5ir.V:S2s
vayamos,*********************-7ir.V:S1p
vayáis,**********************-6ir.V:S2p
vayan,***********************-5ir.V:S3p
fuera,***********************-5ir.V:T1s:T3s
fueras,**********************-6ir.V:T2s
fuéramos,********************-8ir.V:T1p
fuerais,*********************-7ir.V:T2p
fueran,**********************-6ir.V:T3p
fuese,***********************-5ir.V:U1s:U3s
fueses,**********************-6ir.V:U2s
fuésemos,********************-8ir.V:U1p
fueseis,*********************-7ir.V:U2p
fuesen,**********************-6ir.V:U3p
ve,**************************-2ir.V:Y2s
id,**************************-1r.V:Y2p
iría,************************-2.V:C1s:C3s
irías,***********************-3.V:C2s
iríamos,*********************-5.V:C1p
iríais,**********************-4.V:C2p
irían,***********************-3.V:C3p
```

Figura 2

Conjugación automática del verbo *ir*.

REFERENCIAS

CASAJUNA, R. & C. RODRIGUEZ. 1985. Verificación ortográfica en castellano; la realización de un diccionario en ordenador, *Español Actual* 44.

COURTOIS, Blandine. 1984. DELAS: *Dictionnaire Electronique du LADL. Mots Simples*, Rapport Technique du LADL No.12.

DANLOS, Laurence. 1987. *The Linguistic Basis of Text Generation*. Cambridge: Cambridge University Press.

ELIA, Annibale. 1984. *Le dictionnaire electronique de l'italien* ILUS, Salerno.

GROSS, Maurice. 1975. *Méthodes en syntaxe*. Paris: Cantilène, 1987.

_____ . 1968-1987. *Grammaire franciase. Le verbe, le nom et l'adverbe*, 3 vols. Paris: Cantilène.

HARRIS, Zellig. 1982. *A Grammar of English on Mathematical Principles*. New York: Wiley-Interscience.

SOPEÑA, L. de. Diccionario del castellano en ordenador para composición y verificación de textos, *Procesamiento del lenguaje natural* 4.27-33.

SUBIRATS RÜGGEBERG, C. 1986. *Sentential Complementation in Spanish. A lexico-grammatical study of three classes of verbs*. Amsterdam/Philadelphia: John Benjamins.

_____. 1987. *El Diccionario Electrónico de Formas Simples del Español. Informe Técnico No.2*. Bellaterra, Universidad Autónoma.