

EL DICCIONARIO ELECTRONICO DEL ESPAÑOL

CARLOS SUBIRATS RÜGGEBERG

UNIVERSIDAD AUTONOMA DE BARCELONA
FACULTAD DE LETRAS
DEPARTAMENTO DE FILOGIA ESPAÑOLA
08193 Bellaterra (Barcelona), ESPAÑA

RESUMEN

El diccionario electrónico consiste en una lista exhaustiva del léxico español; cada una de las entradas de dicha lista va acompañada de una codificación alfanumérica, que es utilizada por el paquete de programas de generación automática de las formas flexivas. Dichos programas utilizan la información del código alfanumérico para generar todas las formas flexivas; a su vez, para cada una de las formas generadas, los programas especifican:

- a. las principales propiedades morfológicas de flexión;
- b. su relación formal con la forma de base del diccionario electrónico.

La posibilidad de disponer, por primera vez, de un léxico codificado y exhaustivo tanto de las formas de base como de las formas flexivas de la lengua española, permitirá realizar tanto investigaciones teóricas como aplicaciones tecnológicas de carácter radicalmente nuevos.

EL DICCIONARIO ELECTRONICO DEL ESPAÑOL(1)

1. CARACTERISTICAS GENERALES

1. La lingüística y la informática se han unido en los últimos años en un campo de investigación interdisciplinar que ha alcanzado un importante desarrollo en las investigaciones centradas en la simulación por medios informáticos de aquellos aspectos del comportamiento humano en los que interviene el lenguaje natural. Las investigaciones conjuntas de la informática y la lingüística no han girado en torno al establecimiento de modelos teóricos de la comunicación, es decir, modelos teóricos de la producción y reconocimiento de frases por parte de un hablante oyente ideal, sino que han trabajado, fundamentalmente, en desarrollar descripciones gramaticales informatizadas, así como lenguajes formales, que puedan tener aplicaciones reales en el tratamiento informatizado de las lenguas naturales.

La realización de gramáticas y diccionarios electrónicos constituye uno de los campos en los que la investigación coordinada de la informática y la lingüística ha conseguido resultados más importantes. Dichos resultados no se han materializado únicamente en aplicaciones tecnológicas concretas dentro de lo que se ha empezado a llamar la "industria de la lengua"(2), es decir, la tecnología de la documentación automática, la generación o el análisis automático de textos, etc., sino que han puesto de manifiesto, a su vez, que las propiedades formales de las lenguas naturales son mucho más complejas de lo que se había supuesto inicialmente en la lingüística generativa. En este sentido, la constatación de la complejidad formal de las lenguas naturales ha llevado incluso a plantear una revisión del concepto de regla dentro de la lingüística transformacional (cf. Harris, 1982 y Gross, 1975).

La elaboración de un diccionario electrónico del español constituye una contribución dentro de esta nueva línea de investigación lingüística, la cual abrirá nuevas perspectivas teóricas y metodológicas en la lingüística española.

2. El diccionario electrónico consiste en una lista de entradas, que pretende constituir un léxico exhaustivo del español. Cada una de las entradas de dicha lista va acompañada de un código alfanumérico que especifica las siguientes propiedades morfológicas:

- a. la categoría gramatical;

- b. las propiedades de flexión de los nombres, verbos y adjetivos, que son las clases de palabras que poseen propiedades de flexión en español.

El código alfanumérico que acompaña a los verbos, los nombres y los adjetivos remite a unos ficheros de algoritmos de la flexión morfológica del español. De este modo, los programas de generación, que forman parte del diccionario electrónico, utilizan la información codificada de las propiedades morfológicas de flexión de las entradas correspondientes para generar todas las formas flexivas. La obtención de las formas flexivas mediante programas de generación y no mediante listados tiene una importancia crucial en un diccionario electrónico del español, debido a su compleja morfología flexiva; en efecto, la abundancia de formas flexivas del español hace que el número de formas de base del diccionario electrónico se multiplique por un factor 10 tras la aplicación de los programas de generación morfológica.

La selección de las formas de base de un diccionario electrónico no se puede llevar a cabo tomando --o, mucho menos aun, copiando--, las entradas de un único diccionario, aunque se trate del que tradicionalmente, se ha consierado el más "prestigiso", como el Diccionario de la Real Academia Española. En efecto, los diccionarios tradicionales están pensados para que puedan ser utilizados por lectores humanos, los cuales poseen una capacidad, o una competencia lingüística, que les permite suplir fácilmente toda aquella información que en los mencionados diccionarios o bien se encuentra de forma implícita, o bien no se especifica en absoluto. Por ello, para determinar las propiedades morfológicas de cada una de las entradas del diccionario electrónico del español, se ha tenido que añadir, en primer lugar, toda la información morfológica que no se especifica en los diccionarios tradicionales y, en segundo lugar, se ha tenido que sistematizar toda la información que en dichos diccionarios se especifica de forma más o menos dispersa.

No obstante, nuestra labor no se ha limitado únicamente a añadir, o a sistematizar, toda aquella información que como ya hemos señalado anteriormente, en los diccinarios usuales, o bien se encuentra de forma implícita, o bien no se especifica en absoluto: nuestro trabajo ha consistido, además, en el establecimiento de un nuevo criterio para determinar las entradas del diccionario electrónico. En efecto, la lexicografía tradicional mezcla criterios semánticos, etimológicos y formales para establecer los criterios que le permiten determinar una entrada desde el punto de vista lexicográfico; por el contrario, la determinación de las formas de base del diccionario electrónico del español se ha realizado utilizando un criterio de entrada basado en propiedades formales.

2. CARACTERISTICAS DE LOS PROGRAMAS DE GENERACION DE FORMAS FLEXIVAS(3)

El paquete de programas de generación de formas flexivas está formado por:

- a. un programa de generación de formas nominales y adjetivas;
- b. un programa de generación de formas verbales.

El programa de generación de formas nominales y adjetivas está formado por un fichero de algoritmos de la flexión nominal y adjetiva del español y por el programa de flexión propiamente dicho. Por otro lado, el paquete de programas de generación de formas verbales tiene una estructura más compleja, debido a las características de la flexión verbal del español, en la que partiendo de una única forma de base, que corresponde al infinitivo, es necesario generar un gran número de formas flexivas; el paquete de programas de generación de las formas verbales está formado dos ficheros de datos y por tres programas:

- a. un fichero con los algoritmos de la flexión de los verbos regulares e irregulares en español;
- b. un fichero con todas las desinencias verbales de la conjugación española, establecida y clasificada en función de los tiempos verbales a los que pertenecen;
- c. un programa de pretratamiento que actúa sobre el fichero de conjugaciones y genera un nuevo fichero, que es el que realmente utiliza el programa de generación de formas verbales;
- d. un programa de pretratamiento que actúa sobre el fichero de desinencias, y que, al igual que en el caso anterior, genera un nuevo fichero que es el que utiliza el programa de generación de formas verbales;
- e. el programa de generación de formas verbales propiamente dicho.

El funcionamiento de los distintos programas de generación morfológica del diccionario electrónico es esencialmente el mismo. Por un lado, el programa de generación de formas nominales y adjetivas consulta el código alfabético que acompaña a cada una de las entradas del diccionario; en el caso de que dicho código corresponda con el que se les ha asignado a los nombres y los adjetivos, es decir, N y A, el programa consulta el código numérico

de la entrada. Dicho código le remite al fichero de algoritmos de la flexión nominal y adjetiva, en el que se especifican las propiedades formales de flexión necesarias para que el programa pueda generar de forma automática todas las formas flexivas correspondientes.

El programa de flexión verbal consulta igualmente el código alfabético de las entradas del diccionario. Cuando dicho código corresponde con el que se ha asignado a los verbos, es decir, V, el programa consulta el código numérico de la entrada verbal. Dicho código le remite al fichero de algoritmos de la conjugación, en el que se dan los datos para que el programa pueda hacer el cálculo de las raíces verbales; a su vez, en dicho fichero se especifica, además, la combinación de dichas raíces con las desinencias verbales correspondientes. El programa, entonces, consulta el fichero de desinencias y genera todas las formas flexivas, correspondientes a la única forma de base que se encuentra en el diccionario electrónico, es decir, el infinitivo. En realidad, la codificación de los verbos es algo más compleja, ya que existen códigos especiales para determinar si un verbo admite una única forma para el participio, como por ejemplo, el verbo ser (cf. Figura 1), o bien varias, como por ejemplo, el verbo haber (cf. Figura 2). Asimismo, a lo largo del desarrollo de nuestro proyecto de investigación, habrá que marcar todos los verbos con respecto a sus posibilidades de combinación con uno o dos pronombres clíticos, para que los programas puedan generar, asimismo, todas las formas del imperativo, así como las del infinitivo y del gerundio, que llevan uno, o dos, pronombres ortográficamente unidos al verbo, como por ejemplo, háblame, pensárselo, callándotelo, etc.

El resultado de los programas de generación morfológica los podemos observar en las figuras 1, 2 y 3. Nótese que los programas no se limitan a generar únicamente las cadenas de caracteres, correspondientes a las formas flexivas de una determinada forma de base del diccionario, sino que además especifican:

- a. sus principales propiedades morfológicas;
- b. las relaciones formales que mantienen con la forma de base del diccionario electrónico.

En efecto, observemos, por ejemplo, la siguiente forma verbal, generada automáticamente (cf. la Figura 2):

habremos,*****-5er.V:Flp

Además de la cadena de caracteres habremos, el programa nos proporciona una especificación codificada de la categoría gramatical a la que pertenece dicha cadena, así como una especificación de sus correspondientes propiedades morfológicas de

tiempo, modo, persona y número; en efecto, v nos indica que se trata de un verbo, F indica que se trata un tiempo futuro del modo indicativo, 1 indica que se trata de una primera persona y, finalmente, p indica que se trata de una forma verbal en plural. Los programas de generación de formas flexivas de nombres y adjetivos también dan una especificación de la categoría gramatical, así como de las correspondientes propiedades morfológicas de género y numero de las cadenas de caracteres generadas; en efecto, en (cf. Figura 3):

actrices,*****-5or.Nfp

N nos indica que se trata de un nombre, f indica que se trata de un femenino y p, que se trata de un plural.

A su vez, tanto los programas de generación de formas verbales como los de generación de formas nominales y adjetivas especifican la relación que mantiene cada una de las formas generadas con la correspondiente forma de base del diccionario electrónico. Así, en:

partiésemos,*****-6r.V:Tlp

inmediatamente después de los asteriscos, que completan la línea hasta el trigésimo espacio, el programa indica, mediante un número precedido del signo "-", el número de caracteres que hay que eliminar de la cadena generada por el programa para obtener la cadena de caracteres común, tanto a la forma generada, como a la forma de base del diccionario. Así, en el ejemplo anterior, el signo "-" y el número 6, indican que hay que eliminar 6 caracteres a la cadena partiésemos, para obtener parti, que es la cadena de caracteres común a la forma generada por el programa, es decir, partiésemos, y a la forma de base del diccionario, es decir, partir. Asimismo, el programa da una indicación del carácter, o caracteres, que hay que añadir a la cadena común para obtener la forma de base. En nuestro anterior ejemplo, el programa nos indica que a la cadena que tienen en común la forma generada y el infinitivo, es decir, a parti, hay que añadirle el carácter r para obtener la forma de base partir.

Asimismo, en:

bombines,*****-4ín.Nmp

-4 indica que hay que eliminar 4 caracteres para obtener bomb, es decir, la cadena de caracteres común a la forma generada bombines y

a la forma de base bombín; asimismo, el programa indica que hay que añadir ín a la cadena común bomb para obtener la forma de base bombín.

Las indicaciones que proporciona el programa con respecto a las propiedades morfológicas de las formas generadas van a tener una importancia crucial en todas las investigaciones que utilicen el diccionario electrónico para estudiar aspectos de la estructura de la palabra en español, en relación con sus propiedades morfológicas.

3. CONCLUSION

La elaboración del diccionario electrónico del español va a permitir realizar tanto investigaciones de tipo teórico, como aplicaciones tecnológicas dentro del campo de la "industria de la lengua". Desde un punto de vista teórico, la utilización del diccionario electrónico va a permitir, por primera vez para el español, la realización de investigaciones sobre la estructura de la palabra, partiendo de la utilización de un léxico exhaustivo. La utilización de datos exhaustivos va a permitir llevar a cabo investigaciones morfológicas sobre el español, partiendo de la consideración de todo el léxico y no únicamente de sectores parciales y aislados de éste, como se ha hecho hasta ahora tanto en la gramática generativa como en el estructuralismo. La investigación de la estructura del léxico en su globalidad no solo contribuirá a aumentar nuestro conocimiento sobre su estructura, sino que permitirá, además, establecer las bases de una morfología científica, sólidamente basada en un conocimiento empírico del léxico. Asimismo, en la medida en la que el diccionario electrónico pretende ser un reflejo exhaustivo del léxico español, puede constituir una base de datos fiable con un porcentaje de error nulo en sus aplicaciones tecnológicas en el campo de la "industria de la lengua".

ser,*****.V:W00
 siendo,*****-5ser.V:G00
 sido,*****-3ser.V:Kms
 soy,*****-2ser.V:P1s
 eres,*****-4ser.V:P2s
 es,*****-2ser.V:P3s
 somos,*****-4er.V:P1p
 sois,*****-3ser.V:P2p
 son,*****-2er.V:P3p
 era,*****-3ser.V:I1s:I3s
 eras,*****-4ser.V:I2s
 éramos,*****-6ser.V:I1p
 erais,*****-5ser.V:I2p
 eran,*****-4ser.V:I3p
 fui,*****-3ser.V:J1s
 fuiste,*****-6ser.V:J2s
 fue,*****-3ser.V:J3s
 fuimos,*****-6ser.V:J1p
 fuisteis,*****-8ser.V:J2p
 fueron,*****-6ser.V:J3p
 seré,*****-1.V:F1s
 serás,*****-2.V:F2s
 será,*****-1.V:F3s
 seremos,*****-4.V:F1p
 seréis,*****-3.V:F2p
 serán,*****-2.V:F3p
 sea,*****-1r.V:S1s:S3s
 seas,*****-2r.V:S2s
 seamos,*****-4r.V:S1p
 seáis,*****-3r.V:S2p
 sean,*****-2r.V:S3p
 fuera,*****-5ser.V:T1s:T3s
 fueras,*****-6ser.V:T2s
 fuéramos,*****-8ser.V:T1p
 fuerais,*****-7ser.V:T2p
 fueran,*****-6ser.V:T3p
 fuese,*****-5ser.V:T1s:T3s
 fueses,*****-6ser.V:T2s
 fuésemos,*****-8ser.V:T1p
 fueseis,*****-7ser.V:T2p
 fuesen,*****-6ser.V:T1s
 sé,*****-1er.V:Y2p
 sed,*****-1r.V:Y2p
 sería,*****-2.V:C1s:C3s
 serías,*****-3.V:C2s
 seríamos,*****-5.V:C1p
 seríais,*****-4.V:C2p
 serían,*****-3.V:C3p

Figura 1. Conjugación automática del verbo ser

haber,*****.V:W00
 habiendo,*****-5er.V:G00
 habido,*****-3er.V:Kms
 habida,*****-3er.V:Kfs
 habidos,*****-4er.V:Kmp
 habidas,*****-4er.V:Kfp
 he,*****-1aber.V:P1s:Y2p
 has,*****-1ber.V:P2s
 ha,*****-0ber.V:P3s
 hemos,*****-4aber.V:P1p
 habéis,*****-3er.V:P2p
 han,*****-1ber.V:P3p
 había,*****-2er.V:I1s:I3s
 habías,*****-3er.V:I2s
 habíamos,*****-5er.V:I1p
 habíais,*****-4er.V:I2p
 habían,*****-3er.V:I3p
 hube,*****-3aber.V:J1s
 hubiste,*****-6aber.V:J2s
 hubo,*****-3aber.V:J3s
 hubimos,*****-6aber.V:J1p
 hubisteis,*****-8aber.V:J2p
 hubieron,*****-7aber.V:J3p
 habré,*****-2er.V:F1s
 habrás,*****-3er.V:F2s
 habrá,*****-2er.V:F3s
 habremos,*****-5er.V:F1p
 habréis,*****-4er.V:F2p
 habrán,*****-3er.V:F3p
 haya,*****-2ber.V:S1s:S3s
 hayas,*****-3ber.V:S2s
 hayamos,*****-5ber.V:S1p
 hayáis,*****-4ber.V:S2p
 hayan,*****-3ber.V:S3p
 hubiera,*****-6aber.V:T1s:T3s
 hubieras,*****-7aber.V:T2s
 hubiéramos,*****-9aber.V:T1p
 hubierais,*****-8aber.V:T2p
 hubieran,*****-7aber.V:T3p
 hubiese,*****-6aber.V:T1s:T3s
 hubiesen,*****-7aber.V:T2s
 hubiésemos,*****-9aber.V:T1p
 hubieseis,*****-8aber.V:T2p
 hubiesen,*****-7aber.V:T1s
 habed,*****-1r.V:Y2p
 habría,*****-3er.V:C1s:C3s
 habrías,*****-4er.V:C2s
 habríamos,*****-6er.V:C1p
 habríais,*****-5er.V:C2p
 habrían,*****-4er.V:C3p

Figura 2. Conjugación automática del verbo haber

andaluz,*****.Nms
andaluza,*****-1.Nfs
andaluces,*****-3z.Nmp
andaluzas,*****-2.Nfp
andaluz,*****.Ams
andaluza,*****-1.Afs
andaluces,*****-3z.Amp
andaluzas,*****-2.Afp
monje,*****.Nms
monja,*****-1e.Nfs
monjes,*****-1.Nmp
monjas,*****-2e.Nfp
abad,*****.Nms
abadesa,*****-3.Nfs
abades,*****-2.Nmp
abadesas,*****-4.Nfp
alemán,*****.Nms
alemana,*****-3án.Nfs
alemanes,*****-4án.Nmp
alemanas,*****-4án.Nfp
alemán,*****.Ams
alemana,*****-3án.Afs
alemanes,*****-4án.Amp
alemanas,*****-4án.Afp
zar,*****.Nms
zarina,*****-3.Nfs
zares,*****-2.Nmp
zarinas,*****-4.Nfp
rey,*****.Nms
reina,*****-3y.Nfs
reyes,*****-2.Nmp
reinas,*****-4y.Nfp
bailarín,*****.Nms
bailarina,*****-3ín.Nfs
bailarines,*****-4ín.Nmp
bailarinas,*****-4ín.Nfp
actor,*****.Nms
actriz,*****-3or.Nfs
actores,*****-2.Nmp
actrices,*****-5or.Nfp
emperador,*****.Nms
emperatriz,*****-4dor.Nfs
emperadores,*****-2.Nmp
emperatrices,*****-6dor.Nfp

Figura 3. Flexión automática de algunos nombres y adjetivos

NOTAS

1. Queremos dar las gracias a D.Bernabé, J.B.Carreño, Ll.Ferrer, S.Galera, I.Nuñez, T.Peral, M.Rodriguez y J.Torrecilla, así como a los restantes miembros del Centro de Cálculo de la Universidad Autónoma de Barcelona por toda la ayuda que nos han ofrecido. También queremos expresar nuestro agradecimiento a Annibale Elia, así como a todos los miembros de su equipo de investigación del Istituto di Linguistica de la Università di Salerno (Italia) por su importante colaboración. Queremos expresar, asimismo, nuestro agradecimiento a M.Geffriaud, A.Guillet, G.Klarsfeld, C.Leclère, M.Piot, Ph.Vasseux, así como a todos los restantes miembros del Laboratoire d'Automatique Documentaire et Linguistique de la Université Paris 7, por toda la ayuda científica y técnica que nos han ofrecido. La ayuda de Blandine Courtois 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) del Ministerio de Educación y Ciencia y una ayuda de Acción Integrada Hispano-Francesa (9/322).

2. Cf. Maurice Gross. 1986. "Une mise à plat du français écrit", Les industries de la langue, Université Paris 8, pp.51-55.
3. Los programas de generación de formas verbales han sido escritos por Blandine Courtois (cf. Programas utilizados en las Referencias); dichos programas han sido transferidos al Departamento de Filología Española de la Universidad Autónoma de Barcelona por el Laboratoire d'Automatique Documentaire et Linguistique de la Université Paris 7, dentro del acuerdo oficial de cooperación científica que existe entre ambas instituciones.

REFERENCIAS

1. PROGRAMAS UTILIZADOS

1. Programa de pretratamiento del fichero de desinencias verbales, Blandine Courtois, Laboratoire d'Automatique Documentaire et Linguistique (CNRS), Université Paris 7
2. Programa de pretratamiento del fichero de conjugaciones, Blandine Courtois, Laboratoire d'Automatique Documentaire et Linguistique (CNRS), Université Paris 7
3. Programa de generación de formas verbales, Blandine Courtois, Laboratoire d'Automatique Documentaire et Linguistique (CNRS), Université Paris 7
4. Programa de generación de formas nominales y adjetivas, Blandine Courtois, Laboratoire d'Automatique Documentaire et Linguistique (CNRS), Université Paris 7

2. REFERENCIAS BIBLIOGRAFICAS

- * COURTOIS, Blandine. 1984. DELAS: Dictionnaire Électronique du LADL. Mots simples, Rapport Technique du LADL No.12.
- * CUERVO, Rufino J. 1886 y 1893. Diccionario de construcción y régimen de la lengua castellana, 2 vols. Bogotá: Instituto Caro y Cuervo, 1953 y 1954.
- * DANLOS, Laurence. 1987. The Linguistic Basis of Text Generation. Cambridge: Cambridge University Press.
- * ELIA, Annibale. 1986. Le dictionnaire électronique de l'italien, Istituto di Linguistica, Università di Salerno.
- * GROSS, Maurice. 1975. Méthodes en syntaxe. Paris: Hermann.
- * GROSS, Maurice. 1968-1987. Grammaire française, 3 vols; Syntaxe du verbe, vol.I; Syntaxe du nom, vol.II; Syntaxe de l'adverbe, vol.III. Paris: Cantilène.

- * MOLINER, María. 1975. Diccionario de uso del español, 2 vols. Madrid: Gredos.
- * NEBRIJA, Antonio de. 1503. De ui ac potestate litterarum, trad. y ed. A. Quilis y P. Usábel. Madrid: SGEL, 1987.
- * SOPEÑA, Luis. 1987. Diccionario del castellano en ordenador para la composición y verificación de textos, Centro de investigación UAM-IBM, Madrid.
- * SUBIRATS RÜGGERBERG, C. 1987. Sentential Complementation in Spanish. Amsterdam/Philadelphia: John Benjamins.
- * SUBIRATS RÜGGERBERG, C. (en prensa). El diccionario electrónico del español, Actas del III Coloquio de la Sociedad Española de Procesamiento del Lenguaje Natural, Madrid.