REVISTA ELECTRÓNICA DE ESTUDIOS FILOLÓGICOS

La codificación de la información morfológica en los lexicones computacionales

Rafael Marín, Begoña Martínez, David Miramón
Departamento de Lingüística Computacional, Planeta Actimedia
{rmarin, bmartinez, dmiramon}@planeta-actimedia.es

1. Introducción

Por los resultados obtenidos hasta el momento, uno de los ámbitos en que el procesamiento del lenguaje natural ha demostrado más a las claras su potencial es el del análisis morfológico automático. Cabe señalar, no obstante, que en los últimos años la atención se ha centrado más bien en cuestiones tales como la desambiguación morfológica o el tratamiento de la morfología derivativa, dejando en cierta medida de lado la problemática que supone el desarrollo de bases de datos léxicas, el módulo sobre el que pivotan la mayoría de estos sistemas.

En este sentido, conviene destacar un aspecto que, a simple vista, suele pasarse por alto: para el desarrollo de lexicones computacionales, no basta con codificar de forma adecuada la información de los diccionarios al uso, que resulta en muchos casos poco sistemática, cuando no insuficiente; en el proceso de codificación deben tomarse a menudo determinadas decisiones lingüísticas (fundamentalmente morfológicas), en algunos casos nada triviales. En este punto concreto pretendemos incidir aquí.

Después de un somero repaso a la selección y estructura de las entradas de un diccionario electrónico en la sección 2, nos ocuparemos de algunas cuestiones relacionadas con la codificación de la información categorial y flexiva del léxico común en la sección 3; en la 4, analizaremos la problemática particular que presentan los nombres comunes; finalmente, en la 5, resumiremos las conclusiones generales que pueden extraerse a partir de los datos analizados.

2. Selección y estructura de las entradas

Una de las primeras decisiones que se deben tomar en la elaboración de cualquier diccionario pasa por seleccionar las entradas que deben aparecer en él y las que deben quedar fuera, y la manera en que se estructurarán.

En este sentido, conviene recordar, sin ir más lejos, que los nombres propios no tienen cabida en los diccionarios tradicionales. Por otro lado, la decisión de qué formas compuestas están lexicalizadas y cuáles no es una cuestión que, aunque cada vez menos, sigue suscitando una cierta controversia (Cabré 1992; Piera y Varela, 1999; Val Álvaro, 1999).

Por lo que respecta a la estructura de los datos, precisamente en lo que a las formas compuestas se refiere, los diccionarios tradicionales no suelen considerarlas como entradas independientes; así, *agua de borrajas* sería una de las muchas acepciones de *agua*. En los diccionarios electrónicos, por regla general, cualquier entrada compuesta se incluye en un lema independiente, de modo tal que *agua de borrajas* constituye un lema distinto al de *agua*, con el que, además, no es necesario que mantenga relación alguna.

Esta cuestión nos permite establecer ya una de las diferencias más significativas entre diccionarios en formato papel y diccionarios electrónicos: mientras que los primeros conjugan (no siempre de forma clara ni sistemática) criterios morfológicos y semánticos a la hora de decidir la estructura de los datos, los segundos siguen un criterio estrictamente morfológico.

En lo que atañe a la codificación de esta información morfológica, en los sistemas de análisis morfológico, una de las posibilidades más ampliamente utilizadas consiste en distinguir, dentro de una misma etiqueta, una primera parte en la que se codifica la categoría gramatical de cada palabra (e.g. adjetivo, verbo), y una segunda en la que se incluye información relativa a la morfología flexiva (e.g. femenino, plural). Así, por ejemplo, en una etiqueta como NCMP hay que diferenciar claramente entre N(ombre) C(omún), por una parte, y M(asculino) S(ingular), por otra.

3. Problemas de codificación sobre el léxico común

Si se estructuran los datos de esta manera, se puede establecer una distinción entre categorías que va a ser muy útil para lograr las dos funcionalidades que se esperan de un sistema de análisis morfológico (la lematización y la categorización). Por un lado, dispondremos de categorías aplicables a un lema (e.g. *perro*-NC) y, por otro, de categorías aplicables a las formas de un lema (e.g. *perros*-NCMP, perra-NCFS). Asimismo, podremos aislar los problemas relativos a la codificación de la categoría gramatical y a la morfología flexiva, como vamos a comprobar a continuación

3.1. La categoría gramatical

El tratamiento de la información categorial no presenta tantos problemas como el de la información flexiva, si bien no debemos dejar de mencionar algunos de ellos. Una estrategia muy común en el desarrollo de diccionarios electrónicos consiste en prohibir explícitamente la aparición de dos lemas con la misma forma y con idéntica categoría gramatical (entre otras cosas, para preservar la

consistencia de los datos).

Sin embargo, en ocasiones encontramos ejemplos como *mida* o *boqueta*, con dos entradas como N(ombre) C(omún), una masculina (n.m.) y otra femenina (n.f.); esto es, la misma forma y la misma categoría gramatical (la única diferencia estriba en las marcas de flexión).

Entrada	**Categoría**	**Definición**
mida	n.m.	Brugo.
mida	n.f.	Medida.
boqueta	n.m.	Persona labihendida.
boqueta	n.f.	Abertura para ventilación de las labores subterráneas.

Tabla 1. Lemas con idéntica forma superficial y con la misma categoría de lema.

El tratamiento de estos casos[1] resulta ciertamente problemático: si incluimos dos lemas separados, la coherencia de los datos se resiente sensiblemente; si introducimos todas las formas como variantes de un mismo lema, podemos falsear la estructura interna del diccionario ya que, estrictamente, pertenecen a lemas diferentes.

Pero hay más problemas relacionados con la codificación de la categoría gramatical. Es el caso, por ejemplo, de aquellas palabras cuya categoría de partida (por lo general, un adverbio) puede pasar a ser una locución prepositiva (al adjuntársele una preposición) o una locución conjuntiva (al combinarse con una conjunción), como en las tríadas *fuera, fuera de, fuera de que* o *además, además de, además de que*. En ocasiones, parece que la solución más adecuada consiste en incluir tres lemas independientes y asignarle a cada uno de ellos la categoría gramatical que le corresponde: *además* (adverbio), *además de* (preposición), *además de que* (conjunción). Este tipo de información, nuevamente, queda fuera de los diccionarios al uso.

3.2. Morfología flexiva

No siempre es sencillo decidir el plural de una palabra aunque ésta pueda flexionarse fácilmente siguiendo la normativa de la gramática de la lengua. Algunos de los casos más dudosos son aquellos que hacen referencia a entidades abstractas o nombres no contables como, por ejemplo, los elementos químicos. Cabría preguntarse si tienen plural palabras del tipo de *oxígeno* o *cocaína*.

Tampoco resulta fácil determinar el género de algunas palabras. Si nos centramos en el campo semántico de los oficios y las profesiones, dado que hoy día participan en el mundo laboral tanto hombres como mujeres, es perfectamente normal hablar de 'la *teniente* de alcalde' o 'el *modelo* depasarela'. En algunos casos, la flexión de género no se refleja en la superficie de la forma pero sí en la categoría: *conferenciante*-NCCS (nombre común, género común y número singular). A este

respecto, tampoco debemos olvidar los ejemplos de masculinos regresivos como *modista* > *modisto* o *comadrona* > *comadrón* (Santana *et al.* 1999).

Con todo, en un buen número de casos, lo más problemático es decidir la manera más adecuada en que debe codificarse el género y el número. Veamos algunos ejemplos.

3.2.1. Taxonomías

Por su particular comportamiento, el grupo de las taxonomías debe recibir un tratamiento especial dentro del diccionario electrónico. Las familias y órdenes relativas a la descripción del mundo animal y vegetal suelen disponer en muchos casos de tres acepciones: la primera es el adjetivo, la segunda corresponde al nombre del individuo y la tercera designa el nombre de la familia o género.

Examinemos un ejemplo para comprender mejor el comportamiento de este tipo de palabras. En un diccionario al uso, la voz *acantáceo*tiene una acepción como adjetivo y sustantivo femenino, otra como adjetivo, y una subacepción, *acantáceas*, como sustantivo femenino plural, según se observa en la tabla siguiente:[2]

Entrada	**Categoría**	**Definición**
acantáceo	adj. y n.f.	Dícese de las plantas pertenecientes a la familia acantáceas.
acantáceo	adj.	Relativo o parecido al acanto.
acantáceas	n.f.pl.	Familia de plantas herbáceas o subarbustivas, con flores de colores vivos, en espiga terminal o aislada...

Tabla 2. La entrada *acantáceo* en un diccionario y sus correspondientes acepciones.

Para adaptar esta información a un lexicón computacional, una posibilidad consiste en incluir un lema adjetival (*acantáceo*) con cuatro formas, otro (*acantácea*), sustantivo femenino, con una forma para el singular y otra para el plural, y un tercero (*acantáceas*), femenino plural:

Lema	**Forma**	**Etiqueta**	**Ejemplo**
acantáceo	acantáceo	ATPMS	El arbusto acantáceo
	acantácea	ATPFS	La planta acantáceo
	acantáceos	ATPMP	Los arbustos acantáceos
	acantáceas	ATPFP	Las plantas acantáceas
acantácea	acantácea	NCFS	La acantácea
	acantáceas	NCFP	Las acantáceas
acantáceas	acantáceas	NCFP	La familia de las acantáceas

Tabla 3. La codificación de *acantáceo* en un lexicón computacional.

Habría otras opciones. Una de ellas consistiría en considerar que no es necesario introducir el lema *acantácea*-NC, y tratar estas posibilidades como nominalizaciones. Otra consistiría en no incluir *acantáceas*-NC como lema independiente de *acantácea*-NC. Todo depende en última instancia de consideraciones que, en cierta medida, superan el ámbito de lo estrictamente morfológico.

3.2.2. Gentilicios

El campo semántico de los pueblos y etnias también presentan cierta complejidad y exige un tratamiento especial. El lema de entrada puede ser un adjetivo (*el niño inglés*) o bien un sustantivo (*los ingleses*). Por su frecuencia de uso, no resulta difícil encontrar el modelo flexivo para este gentilicio en concreto. Hay otros casos, como *pueblo*, cuyas posibilidades de flexión también nos son conocidas: una única forma invariable, tanto para el género como para el número: *el indio pueblo, la india pueblo, los indios pueblo y las indias pueblo*.

Sin embargo, existen muchos otros nombres que designan a etnias y pueblos sobre los que no tenemos un conocimiento totalmente fiable. Para estos casos se han establecido una serie de criterios que nos permiten recoger de forma regular las diferentes posibilidades de flexión. Veamos algunos ejemplos.[3]

Lema	**Forma**	**Etiqueta**	**Ejemplo**
abdalí	abdalí	ATPCS	El niño /la niña abdalí
	abdalíes	ATPCP	Los niños / las niñas abdalíes
aamu	aamu	ATPCI	El niño / la niña aamu Los niños / las niñas aamu
	aamus	ATPCP	Los niños / las niñas aamus
acauayo	acauayo	ATPCI	El niño / la niña acauayo Los niños / las niñas acauayo
	acauaya	ATPFS	La niña acauaya
	acauayos	ATPMP	Los niños acauayos
	acauayas	ATPFP	Las niñas acauayas

Tabla 4. La codificación de ciertos gentilicios en un lexicón computacional.

Como se puede observar en la tabla anterior, a pesar de no conocer con seguridad las propiedades flexivas de estas palabras, hemos hecho algunas suposiciones. En casos como el de *abdalí*, por ejemplo, asumimos que la forma plural es *abdalíes* (aunque quizás también fuera conveniente incluir *abdalís*); en el caso de *aamu*, hemos considerado oportuno incluir una forma singular que cubra las

cuatro posibilidades, y una forma plural, invariable respecto al género. Finalmente, en ejemplos del tipo de *acauayo*, hemos incluido una forma singular que cubra las cuatro posibilidades (ATPCI), una forma en femenino singular, *acauaya* y dos en plural, *acauayos* y *acauayas*. De esta forma, seremos capaces de reconocer tanto *las niñas acauayas* como *las niñas acauayo*.

3.2.3. Las formas compuestas

Dejando de lado el problema relativo a la selección de entradas, al cual ya hemos hecho alusión anteriormente, los compuestos plantean una serie de problemas de codificación en su mayoría relacionados con el número, aunque en algunos casos también con el género. La tabla siguiente, adaptada de Subirats (1994), ofrece un resumen de la mayoría de las posibilidades con respecto a la flexión de las formas compuestas nominales:

Tipo	**Características tipológicas y flexivas**	**Ejemplos**
NA	clase de dos zonas con flexión de N y A	*bomba atómica*
NDN	clase de tres zonas fijas, con flexión exclusiva del primer N	*libro de familia*
AN	clase de dos zonas fijas: se flexionan A y N	*nuevo rico*
NN	clase de dos zonas fijas; la flexión de los dos N es variable (1) flexión del 1er. N y del 2º N (2) flexión del 1er. N , pero no del 2º N (3) flexión del 2º N, pero no del 1er. N	*(1) analista programador* *(2) paquete bomba* *(3) ave maría*
NX	clase de zonas variables, encabezada por un N, seguida de una cadena constante, se flexiona exclusivamente el primer N	*olla a presión*
NAX	clase de zonas variables, encabezada por N y A, seguidos de una cadena constante; se flexiona exclusivamente N y A	*huevo pasado por agua*
NCA y NCN	clases de zonas variables, en las que se flexiona el 1er. N y el 3er. elemento, ya sea este, A o N	*punto y coma*
NAA	clase de zonas variables; se flexionan el N y los dos A	*ácido graso saturado*
NACA	clase de zonas variables; se flexionan el N y los dos A	*objeto volante no identificado*

Tabla 5. Posibilidades de flexión de las formas compuestas.[4]

Como es lógico, si pretendemos dar cuenta de la flexión de estas formas compuestas en un lexicón,

deberemos introducir un código que explicite qué parte del compuesto es la que se flexiona y qué parte permanece invariable. En este sentido, téngase en cuenta que en muchos casos el código de flexión de los diferentes elementos que integran un compuesto no tiene por qué ser el mismo (e.g. *guardia**s** civil**es***).

No obstante, la codificación de algunos compuestos resulta especialmente compleja, debido fundamentalmente a la vacilación constatable en su uso cotidiano. Cabría preguntarse, si es que la hay, cúal es la forma plural de *azul marino* o *verde turquesa*. En el segundo caso, los hablantes suelen coincidir en el plural *verdes turquesa*; en la primera, hay una mayor vacilación.

En el caso de las locuciones verbales, este escollo resulta algo menos difícil de superar, puesto que la única parte del compuesto que se flexiona es el verbo. No obstante, presentan problemas cuya solución es ciertamente compleja, ya que un buen número de ellas admiten la inserción de material lingüístico de diverso tipo, como adverbios (*dar (mucho) la lata*), pronombres u otros argumentos (*meterse (a alguien) en el bolsillo*), sin olvidar la posibilidad de cambios de orden superficial.

En la sección siguiente nos ocupamos de los problemas particulares de codificación que presentan los nombres propios que, como se verá, se alejan considerablemente de los analizados hasta ahora.

4. Los nombres propios

Debido a una serie de propiedades idiosincrásicas que los separan de otros dominios gramaticales (Fernández Leborans, 1999), el análisis de los nombres propios entraña un grado de complejidad ciertamente elevado que, por lo que respecta a su procesamiento computacional, cobra una especial relevancia.

En este sentido, cabe recordar, al margen de otras consideraciones como su naturaleza eminentemente referencial, el hecho de que los nombres propios constituyen un conjunto de elementos potencialmente infinito. En parte debido a ello, algunas aproximaciones intentan llevar a cabo un análisis sin contar con una base de datos de partida (Wacholder y Ravin, 1997), si bien, en otros casos, se parte de un lemario previamente adaptado a las necesidades de esta tarea (Hayes, 1994).

Si se elige la segunda de estas opciones, la primera decisión pasa, nuevamente, por la selección de las entradas. De la misma forma que el diccionario es la base del léxico común, para los nombres propios puede utilizarse una enciclopedia: los nombres propios de partida serán aquellos que figuran en las entradas biográficas, toponímicas, etc. (Casanova *et al.* 2001).

4.1. Generación de variantes

Los nombres propios en general, y especialmente los antropónimos, pueden aparecer de formas diversas en los textos. Un mismo antropónimo, como por ejemplo *John Fitzgerald Kennedy*, puede representarse en un texto como *John F. Kennedy*, *J. F. Kennedy*, *Kennedy* o *JFK*, además de su forma más extensa *John Fitzgerald Kennedy*.

Cabe señalar, no obstante, que un número considerable de variantes de nombres propios se deben a cuestiones relacionadas con la transcripción o transliteración de términos de otras lenguas y su eventual adaptación al español. Encontramos ejemplos en los que tan sólo se trata de un problema de escritura, como sucede con *Boris Yelstin* frente a *Baris Ielstin*, otros casos con mayor o menor implicación fonológica *Mao Tse-Tung* o *Mao Zedong,* y otros en los que la adaptación al español genera formas que hacen difícil decidir si se trata o no de la misma palabra: *Londres*, *London*, *Donostia*, *San Sebastián*.

Para poder reconocer las distintas ocurrencias de un nombre propio en un texto es necesario asociar sus distintas variantes a un único identificador. Este identificador viene asociado a un lema o título de entrada de diccionario que, por lo general, corresponde a la forma de cita del nombre. Por lo común, esa forma tiene la estructura <Apellidos Coma Nombre>, por ejemplo *Picasso, Pablo Ruiz*, pero cuando incluye apellidos compuestos con preposiciones, tratamientos, títulos o referencias a dominios territoriales, puede ser muy compleja y presentar una importante hetereogeneidad de formatos, como en el caso de *Sebastiao José de Carvalho y Melo, conde de Oeyras y marqués de Pombal* o de *Robert Stewart, vizconde de Castlereagh y 2º marqués de Londonderry*.

En nuestro caso, hemos desarrollado un procedimiento de generación de variantes compuesto por dos módulos. El primero es una gramática de cláusulas definidas que reconoce los distintos formatos de cita e identifica sus constituyentes. El segundo puede verse como un traductor que, a partir de un determinado tipo de representación ofrecido por la gramática, aplica a sus elementos todas las posibles operaciones de generación. Entre estas operaciones se cuentan distintas ordenaciones, supresión de elementos y paso a iniciales.

Las transformaciones generales para la gran mayoría de nombres propios generan tres variantes principales: a) sólo el nombre (nom), b) sólo el/los apellido/s (aps) y c) nombre y apellidos (comp). El procedimiento construye también una tabla auxiliar donde se deja constancia del tipo de generación que se ha aplicado al obtener cada variante, de modo que esta información esté disponible para aplicaciones que deban utilizar la asociación entre cadenas de variantes e identificadores. Puede ser, por ejemplo, que se considere conveniente despreciar las variantes generadas como sólo nombre cuando sean excesivamente polisémicas, pero no así las generadas como apellidos.

Veamos algunos ejemplos. A partir de una forma de cita como *Amundsen, Roald Engelbert* se crean las siguientes variantes:

Tabla 6. Ejemplo de generación de variantes.

Junto a este tratamiento general, deben tenerse en cuenta casos más concretos. Respecto a los apellidos, por ejemplo, en determinados casos, es posible referirse a una persona únicamente por el segundo apellido; sería el caso de *Lorca* o de *Galdós*. En el caso de los nombres de pila, dejando de lado el problema que plantean los nombres compuestos respecto a las iniciales, hay otras cuestiones que deben contemplarse, como el uso de diminutivos (e.g. *Jacky Kennedy*).

4.2. Estructura y gestión de los datos

La entrada al procedimiento de generación de variantes es un conjunto de pares <Forma de cita, Identificador>. Las variantes obtenidas de una forma de cita deberán relacionarse con la entidad correspondiente a su identificador (relación n: 1). Al mismo tiempo, desde distintas formas de cita puede obtenerse una misma variante de uso, que deberá relacionarse con distintos identificadores (relación 1: m).

Nos hallamos aquí con el problema general de la ambigüedad léxica, pero con dimensiones particularmente explosivas. Existen formas de cita para las cuales pueden generarse más de cuarenta variantes y, obviamente, existen variantes que se relacionan, con un número muy elevado de entidades.

Una posibilidad consistiría en asociar cada identificador con un lema del diccionario. Dado que *Kennedy* puede haberse obtenido tanto de 'Fitzgerald Kennedy, John', como de su hermano, 'Kennedy, George', como de su mujer, 'Kennedy, Jackeline', deberían incluirse tres lemas distintos con la palabra *Kennedy* en el diccionario para las tres acepciones correspondientes a las tres formas de cita. De este modo, tras el análisis morfológico y la desambiguación necesaria en muchos casos, obtendríamos ya una referencia a los nombres de las distintas entidades que aparecen con una u otras variante en el texto que se procese.

No obstante, esta posibilidad es inconsistente con la estructura de nuestro diccionario: existen entradas para lemas distintos cuando estos lemas representan conjuntos distintos de formas o tienen una oposición de categoría o rasgos morfosintácticos. Así, por ejemplo, la ambigüedad existente entre las distintas acepciones de *banco* es de carácter semántico y está representada por la correspondencia de un único lema con diferentes acepciones.

Desde el punto de vista de gestión de la información hay que observar también que esa posibilidad supondría sobrecargar el espacio morfosintáctico con una complejidad que no le es propia: ni para el análisis ni la corrección interesa conocer la referencia de una forma como *Kennedy.* Por consistencia con el tratamiento de las formas comunes y para la optimización de su gestión, se ha decidido agrupar todas las variantes de nombres propios coincidentes en forma sobre un único lema.

Para solucionar estos y otros problemas, agrupamos los lemas en conjuntos de lemas que denominamos *metaformas*. Inicialmente todas las variantes que se han generado a partir de una misma forma de cita se han asociado a una misma metaforma. No obstante, el sistema permite que la agrupación de lemas en metaformas pueda realizarse por cualquier otro criterio; es más, se permite incluso la posibilidad de que una misma forma pueda pertenecer a distintas metaformas al usar criterios de agrupación distintos.

Nótese que el criterio inicial de generación de metaformas supone que una misma variante obtenida de formas distintas aparece una sola vez como lema, pero pertenece a distintas metaformas, tal como puede observarse en la figura que aparece a continuación:

>INSERTAR<

Figura 1. Ejemplo de la estructura interna de una metaforma.

Como se ve, el concepto de metaforma resuelve el problema de que una misma entidad pueda tener distintas formas. Además, el hecho de poder asociar atributos tanto en los lemas como en las metaformas permite que haya información específica para cada forma y para el conjunto de variantes con el mismo referente.

5. Conclusiones

A partir de los datos analizados, podemos extraer, entre otras, las conclusiones que resumimos a continuación.

Como consideración general, hemos constatado que la codificación morfológica en los lexicones computacionales no se reduce al traspaso de la información de los diccionarios tradicionales: las

diferencias estructurales entre unos y otros obligan a menudo a tomar decisiones gramaticales de cierto calado teórico.

Asimismo, por lo que respecta concretamente a la categorización gramatical y al tratamiento de la morfología flexiva, hemos comprobado que la información contenida en los diccionarios tradicionales no siempre resulta suficiente. Una clara muestra de ello puede encontrarse en la codificación de gentilicios, taxonomías o formas compuestas constituyen una clara muestra de ello.

Por otro lado, si pretendemos reconocer también los nombres propios, en el diseño de nuestro lexicón computacional debemos tener en cuenta cuestiones relacionadas con la generación de variantes y la estructura de los datos, de modo que sea posible una adecuada gestión posterior.

Referencias

Cabré, M. T. (1992). *La terminologia. La teoria, els mètodes, les aplicacions*, Empúries, Barcelona.

Casanova, D., Lloré, X., Marín, R., Merenciano, J. M., Pérez, G. y Trotzig, D. (2001). "ANTRO: Un sistema de reconocimiento y gestión de antropónimos", *Actas del XVII Congreso de la SEPLN* (Revista nº 27), pp. 311-312. Disponible también en http://www.planeta-actimedia.es/esp/banco/ling.htm.

Fernández Leborans, M. J. (1999). "El nombre propio", en I. Bosque y V. Demonte, (eds.), *Nueva gramática descriptiva de la lengua española*, Espasa Calpe, Madrid, vol. 1, pp. 77-128.

Hayes, P. (1994). "NameFinder: Software that finds names in text", *Proceedings of RIAO 94*, New York, pp. 762-774.

Piera, C. y Varela, S. (1999). "Relaciones entre morfología y sintaxis", en I. Bosque y V. Demonte, (eds.), *Nueva gramática descriptiva de la lengua española*, Espasa Calpe, Madrid, vol. 3, pp. 4367-4421.

Santana, O., Pérez, J., Carreras, F., Duque, J., Hernández, Z. y Rodríguez, G. (1999). "FLANOM: Flexionador y lematizador automático de formas nominales", *Lingüística Española Actual,* vol. 21(2), pp. 253-297.

Subirats (1994). "La flexión nominal en el diccionario electrónico de formas compuestas del español", *Lingua Franca*, pp. 63-69.

Val, J. F. (1999). "La composición", en I. Bosque y V. Demonte, (eds.), *Nueva gramática descriptiva de la lengua española*, Espasa Calpe, Madrid, vol. 3, pp. 4757-4841.

Wacholder, N. y Ravin, D. (1997). *Extracting Names from Natural-Language Text*, IBM Research Report 20338.

---

[1] Otros ejemplos en los que el género implica un cambio de significado son *cólera*, *frente* o *maqui*. Obsérvese que aquellos casos en los que es el número el que implica un cambio de significado no afectan a la cuestión que nos concierne aquí, ya que también hay un cambio en el lema: *anal / anales, celo /celos*.

[2] El comportamiento de las taxonomías de sustantivos y adjetivos masculinos (*arácnidos*) no se aparta, en lo esencial, del de los femeninos (*acantáceas*).

[3] Las siglas con que se designan estas etiquetas se rigen por el siguiente patrón: ATP corresponde a adjetivo (A) postnuclear (T) en grado positivo, y NC a nombre (N) común (C); a continuación aparece la información relativa al género: masculino (M), femenino (F) y común (C) y, finalmente, la información relativa al número: singular (S), plural (P) o invariable (I). Así, ATPCS, debe interpretarse como adjetivo (A) postnuclear (T) en grado positivo (P), común (C), singular (S), o NCMP como nombre (N) común (C), masculino (M), plural (P).

[4] N: nombre; A: adjetivo; X: cadena constante; C: elemento constante, que no experimenta variación de forma en la flexión del compuesto.