CODIFICACIÓN DE LA ANOTACIÓN MORFOSINTÁCTICA DE CORPUS EN LENGUAJE SGML

Marta Pino, Mª. Paula Santalla
Instituto de Lexicografía
Real Academia Española
e-mail: mpino@crea.rae.es
paula@crea.rae.es

ABSTRACT

This communication presents a proposal for the encoding of morphosyntactic annotation of language corpora in SGML. This encoding system has been designed for CREA, the Spanish Reference Corpus of the Royal Academy, in order to introduce SGML codes with morphosyntactic annotation in the texts, instead of linking a morphosyntactic tag directly to each word. Our encoding system implies a further representation of the morphosyntactic tagset worked out for CREA. Since all CREA texts are SGML documents, almost completely TEI conformant, an SGML representation of linguistic analysis was needed.

1. INTRODUCCIÓN

El Instituto de Lexicografía de la Real Academia Española está construyendo el *Corpus de Referencia del Español Actual* (CREA), un proyecto que prevé reunir unos cien millones de palabras de textos actuales, codificados en SGML y etiquetados morfosintácticamente. En el trabajo que sigue, se estudia la anotación lingüística del corpus desde el punto de vista de su integración en un entorno SGML y se propone, en consecuencia, un modo de codificar la información morfosintáctica en los textos.

2. ANOTACIÓN MORFOSINTÁCTICA

2.1 FORMATO DE LA ETIQUETACIÓN

Las etiquetas, en consonancia con lo que se describe en Santalla 1995, se organizan físicamente de acuerdo con una gramática de dos niveles, las categorías constituyen el primer nivel, mientras que la identificación de los valores para cada uno de los rasgos que afecta a cada categoría constituye el segundo.

Físicamente: **CATEGORÍA(valor rasgo 1, valor rasgo 2, valor rasgo 3...)**

Cuando para determinado rasgo un ítem puede representar a más de un valor (pero no todos) de los posibles para tal rasgo, la notación lo representa asignando todos los valores implicados en el espacio correspondiente y separándolos por una barra vertical (es una forma de representar lo que podríamos llamar *port-manteau features*).

Físicamente: **CATEGORÍA(valor 1 rasgo 1|valor 2 rasgo 1, valor rasgo 2, valor rasgo 3...)**

Se han previsto, no obstante, algunos valores que representan para determinados rasgos (género, número y número del poseedor) la posibilidad del ítem considerado de referirse a cualquiera de los valores en que tal rasgo puede materializarse (**GEN**, **GENERO**, **NUM**, **NUMP**).

Por otra parte, cuando lingüísticamente está claro que un rasgo no se aplica a determinada subcategoría de una categoría, simplemente tal rasgo no tiene un hueco reservado en la etiqueta implicada (por ejemplo, no todas las etiquetas verbales tienen el mismo número de especificaciones).

Se prevé, en algunos casos -en general, aquellos cuya resolución requiere el manejo de información contextual no estrictamente morfosintáctica al modo de la contenida en el mismo sistema de etiquetas-, la asignación de etiquetas compuestas (*port-manteau tags*):

Físicamente: **CATEGORIA1(valor/es rasgo 1, valor/es rasgo 2...)|CATEGORIA2(valor/es rasgo 1, valor/es rasgo 2...)**

2.2 SUMARIO DE LA INFORMACIÓN CONTENIDA EN LAS ETIQUETAS

En el esquema que sigue quedan reflejadas las categorías gramaticales (en el sentido de clases de palabras) consideradas en el esquema de anotación morfosintáctica diseñado, la denominación tradicional para ellas, y las claves de los atributos que afectan a cada una. Los atributos representados por cifras arábigas se refieren a las categorías gramaticales de tipo flexivo (incluidas las que, en tanto que flexivas, han de ser estimadas residuales, 6) y a especificaciones sintácticas consideradas de interés (7), y los representados por números romanos a las subcategorías identificadas dentro de cada clase de palabras considerada.

ADJ, adjetivo, III, 6, 1, 2
ART, artículo, 1, 2
AV, adverbio (interrogativo, negativo, relativo, deíctico), IV, 6
AV, adverbio (ninguno de los anteriores), 6
CONJ, conjunción, V
CUANT, cuantificador, 1, 2
DEM, demostrativo, 1, 2, 7
ED, enlace discursivo, IX.
IND, indefinido, 1, 2, 7
INT, interrogativo, 1, 2, 7
IT, interjección
MIS, miscelánea, VII
NUM, numerales, VIII, 1, 2
PERS, pronombre personal, 5, 9, 1, 2, 8
POS, posesivo, 5, 9, 1, 2, 7
PR, preposición
PUN, puntuación, VI
REL, relativo, 1, 2, 7
SUST, sustantivo (comunes, meses del año, días de la semana), I, 1, 2
SUST, sustantivo (propios), I
TOT, totalizador, 1, 2
VRB, verbo (no personal excepto participios), II, 3, 4
VRB, verbo (participio), II, 3, 4, 1, 2

VRB, verbo (personal), II, 3, 4, 5, 2

En el esquema que sigue, para cada atributo se enumeran los valores que puede asumir y las claves con que aparecen representados en las etiquetas. En primer lugar, aparecen los atributos que se refieren a categorías flexivas, en segundo lugar los que se refieren a subcategorías dentro de las clases de palabras.

1, género, **m**, masculino, f, femenino, **fd**, femenino débil, **n**, neutro, **GEN**, masculino, femenino, **GENERO**, masculino, femenino o neutro
2, número, sg, singular, **pl**, plural, **NUM**, singular o plural
3, modo, **ind**, indicativo, **sub**, subjuntivo, **imp**, imperativo, **inf**, infinitivo, **ger**, gerundio, **ppio**, participio
4, tiempo, **pres**, presente (amo-ame-ama (tú)-amar-amando), **perf**, pretérito perfecto (he amado-haya amado-haber amado-habiendo amado-amado), **imp**, imperfecto (amaba-amara/amase), **plusp**, pretérito pluscuamperfecto (había amado-hubiera/hubiese amado), **pret**, pretérito (amé), **ant**, pretérito anterior (hube amado), **cond**, condicional (amaría), **condc**, condicional compuesto (habría amado), **fut**, futuro (amaré-amare), **futp**, futuro perfecto (habré amado-hubiere amado)
5, persona, **1**, primera, **2**, segunda, **3**, tercera
6, grado, **pos**, positivo, **cp**, comparativo, **sup**, superlativo
7, capacidad funcional, **d**, determinante, **p**, pronombre, **FUN**, determinante o pronombre
8, caso, **nom**, nominativo, **prep**, preposicional, **ac**, acusativo, **dat**, dativo, **at**, átono (acusativo, dativo y/o pronominal, impersonal, pasivo)
9, número del poseedor, s, singular, **p**, plural, **NUMP**, singular o plural

I, subcategorías del sustantivo, **com**, común, **prp**, propio, **sem**, día de la semana, **mes**, mes del año
II, subcategorías del verbo, **ppal**, principal, **saux**, semiauxiliar
III, subcategorías del adjetivo, **tl**, título, **adj**, adjetivo no título
IV, subcategorías del adverbio, **int**, interrogativo, **rel**, relativo, **neg**, negativo, **deict**, deíctico
V, subcategorías de la conjunción, **coor**, coordinada, **sub**, subordinada, (tipología en extensión).
VI, tipo de puntuación, ;, punto y coma, ,, coma, :, dos puntos, ., punto, - guión, ?, cierre de interrog fs20 ación, ¿, apertura de interrogación, !, cierre de admiración, ¡, apertura de admiración, ", comillas.
VII, subcategorías de la miscelánea, **int**, interrupciones, **ono**, onomatopeyas, **form**, fórmula, **simb**, símbolo, **ext**, palabra extranjera, **abr**, abreviatura, **incl**, inclasificable,
VIII, tipo de numeral, **card**, cardinal, **ord**, ordinal, **frac**, fraccionario, **mult**, multiplicativo, -, cifra con guión intermedio
IX, tipo de enlace discursivo, tipología en desarrollo.

2.3 LEMATIZACIÓN

El concepto de lema que subyace a la anotación es el siguiente: abstracción del conjunto de formas flexionadas en cuanto a género, número, modo, tiempo y/o grado superlativo sintético productivo (formado con el sufijo -
'edsimo, o sus variantes cultas, *-érrimo/-imo,* normalmente aplicadas sobre formas latinizantes de los adjetivos correspondientes).

Junto a la flexión referida en 1, el lema abstrae también la variación producida por la formación derivativa de diminutivos o aumentativos.

Las variaciones formales introducidas por apócope prenominal de adjetivos, determinantes o adverbios no implican, a efectos de la anotación, variación de lema.

Las formas con más de una representación ortográfica posible (en las mismas circunstancias, sin implicar especificación funcional de ningún tipo) pertenecen al mismo lema.

El representante canónico del lema para los lemas con formas distintas para alguno de estos rasgos es el siguiente:

a) Para los lemas con moción de género: la forma que represente al masculino.
b) Para los lemas con moción de número: la forma que represente al singular.
c) Para los lemas con moción de modo y tiempo: la forma que represente al infinitivo presente, lo que implica para los verbos la abstracción de los rasgos de persona y número de la persona.
d) Para los lemas con una forma superlativa en *-ísimo* o sus variantes, la forma positiva de la que se derivan, de acuerdo con lo expuesto en a, b, c y d.
e) Para los lemas a cuyas formas se puede aplicar un sufijo diminutivo o aumentativo, la forma primitiva de acuerdo con lo expuesto en a, b, c y d.
f) Para los lemas alguna de cuyas formas sufre apócope en posición prenominal, la forma plena, de acuerdo con lo expuesto en a, b, c y d.
g) El representante canónico de los lemas con variantes ortográficas es el primero de ellos en orden alfabético. Si las variantes ortográficas sólo se diferencian en la acentuación, la forma sin acentuar.

A efectos de la anotación, pues, las variaciones de forma introducidas por los rasgos persona, grado (excepto lo indicado en 1 respecto al superlativo), capacidad funcional, caso y número de la persona, implican variaciones de lema.

2.4 SITUACIONES PROBLEMÁTICAS

El formato de la etiquetación de las situaciones problemáticas está condicionado por la *linealidad* del texto etiquetado y de las etiquetas. Allí donde es necesario representar un orden estructural (anidamiento entre unidades multilexicales de algún tipo), ello se consigue a través de la adición correlativa de dígitos a las palabras (en el sentido gráfico del término) implicadas según la tradición iniciada en Lancaster (Garside et al. 1987). Una representación visualmente estructural puede conseguirse en SGML, como se verá en la segunda parte de este documento. Por ahora nos atenemos a la representación lineal.

2.4.1 *UNIDADES MULTILEXICALES*

Llamamos unidades multilexicales compuestas a aquellas unidades multilexicales que se extienden a lo largo de dos o más palabras, en sentido estrictamente gráfico, contiguas en el discurso. La etiquetación de una unidad multilexical compuesta implica necesariamente la adición de un par de dígitos a cada componente de la unidad: el primero de ellos representa el lugar que ocupa ordinalmente en el compuesto, y el segundo el número total de componentes de que consta la unidad en cuestión. La numeración debe considerarse la marca de la composición. Además, a la unidad multilexical compuesta se

le debe asignar una de las siguientes posibilidades: a) Una etiqueta individual en cada componente, que entonces precede inmediatamente a los dígitos de composición. b) Tan sólo una etiqueta correspondiente al compuesto en tanto que tal: la misma que se asignaría a un ítem individual con el que pueda compararse funcionalmente, precedida de FR, excepto en el caso de los tiempos compuestos verbales, a los que se sigue asignando la etiqueta VRB. Posicionalmente esta etiqueta sigue, separada por un guión, al lema del último término de la composición. c) Ambas etiquetas. Se asigna siempre lema a todos los componentes, excepto en el caso de los tiempos compuestos. Al compuesto en tanto que tal puede asignársele lema o no. Dentro de estas directrices, se decidirá independientemente cómo quiere etiquetarse cada compuesto en concreto. El resultado, no obstante, al enfrentarnos a unidades multilexicales concretas, nos ha llevado a observar, y posteriormente conservar en la medida de lo posible, ciertas regularidades:

a) En general, son los compuestos funcionalmente comparables a clases abiertas los que tendemos a etiquetar composicional e individualmente a la vez. Los llamamos *lexías*. Por ahora, se etiquetan también de este modo todos los nombres propios compuestos de cualquier tipo (antropónimos, títulos de libros...).También muchos de los funcionalmente comparables a clases cerradas -algunos dudosos, incluso en tanto que unidades compuestas- son etiquetados de este modo:

(1) manos_SUST(com, f, pl)14_mano a_PR24_a la_ART(f, sg)34_el obra_SUST(com, f, sg)44_obra_FRAV(pos)_manos a la obra
(2) echaba_VRB(ppal, ind, imp,1|3, sg)13_echar de_PR23_de menos_AV(cp)33_menos _FRVRB(ppal, ind, imp, 1|3, s)_echar de menos
(3) en_PR13_en nombre_SUST(com, m, sg)23_nombre de_PR33_de_FRPR_en nombre de

b) A una buena parte de los compuestos funcionalmente comparables a clases cerradas, los más indiscutibles en tanto que compuestos, se les asigna solamente una etiqueta conjunta. Los consideramos también *lexías*. Entran dentro de este grupo también *los tiempos compuestos* verbales:

(4) en_13_en cuanto_23_cuanto a_33_a_FRPR_en cuanto a
(5) por_13_por más_23_más que_33_que_FRCONJ(sub)_por más que
(6) sin_12_sin embargo_22_embargo_ED()_sin embargo
(7) ha_12 venido_22_VRB(ppal, ind, perf, 3, sg)_venir

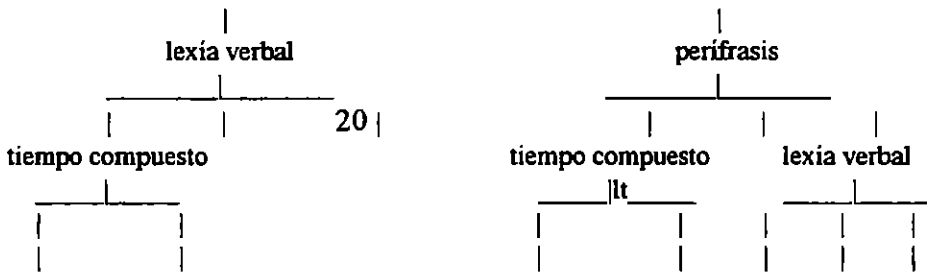
c) A aquellos compuestos en la frontera entre ser tales y ser *colocaciones*, y a las *perífrasis verbales*, se les asigna etiqueta individual pero no conjunta.

(8) arenas_SUST(com, f, pl)12_arena movedizas_ADJ(adj, pos, f, pl)22_movedizo
(9) reloj_SUST(com, m, sg)13_reloj de_PR23_de pulsera_SUST(com, f, sg)33_pulsera
(10) va_VRB(saux, ind, pres, 3, sg)_ir a_PR_a comprar_VRB(ppal, inf, pres)_comprar

La pronominalización verbal, incluso cuando no alterna con una forma no pronominal del verbo, no es considerada un caso de composición. En todo caso, si lo fuera, debería etiquetarse de acuerdo con el modelo (c), aunque podría defenderse también cualquiera de las otras posibilidades.

2.4.2 *UNIDADES MULTILEXICALES ANIDADAS*

Algunas unidades multilexicales compuestas pueden anidarse. En nuestra aproximación lingüística, se anidan los tiempos compuestos y las lexías en las perífrasis y éstas, también unas en otras recursivamente. Representamos el anidamiento de estructuras sin romper nunca las unidades tiempo compuesto y lexía verbal. Esta representación del anidamiento es coherente con la que hemos propuesto para la etiquetación de los compuestos implicados cuando no están anidados: las perífrasis son los compuestos verbales que no se etiquetan categorialmente en tanto que conjunto. Las perífrasis se anidan recursivamente hacia la izquierda:



Se da cuenta del anidamiento etiquetando el componente compuesto como tal y como le corresponde en el interior del compuesto en que se integra. La etiqueta general de la lexía verbal se considera el tercer componente de la perífrasis, por lo que la numeración que marca en estos casos la perífrasis queda interrumpida hasta que se etiqueta el último componente de la lexía anidada. Para las perífrasis verbales progresivamente más externas de una sucesión de perífrasis anidadas, y para las inmediatamente inferiores a ellas, el primer y el último componente de los conjuntos que constituyen es el mismo ítem. Por ello tal componente va seguido de cuatro dígitos: los dos primeros indican que es el último miembro de la perífrasis inferior o anidada, mientras que los dos últimos lo identifican como el primero de la superior o anidante.

(tiempo compuesto anidado en perífrasis)

(11) ha_12 habido_22_VRB(saux, ind, perf, 3, sg)13_haber que_CONJ(sub)23_que salir_VRB(ppal, inf, pres)33_salir

(tiempo compuesto anidado en lexía)

(12) ha_12 echado_22_VRB(ppal, ind, perf, pres, 3, sg)13_echar de_PR23_de menos_AV(cp)33_menos_FRVRB(ppal, ind, perf, pres, 3, sg)

(tiempo compuesto y lexía anidados en perífrasis)

(13) ha_12 acabado_22_VRB(saux, ind, perf, 3, sg)13_acabar por_PR23_por echar_VRB(ppal, inf, pres)13_echar de_PR23_de menos_ADV(cp)33_menos_FRVRB(ppal, inf, pres)33_echar de menos

(perífrasis anidada en perífrasis)

(14) empieza_VRB(saux, ind, pres, 1, sg)13_empezar a_PR23_a ir_VRB(saux, inf, pres)3312_ir comprendiendo_VRB(ppal, ger, pres)22_comprender

(lexía anidada en perífrasis)

(15) va_VRB(saux, ind, pres, 3, sg)12_ir mereciendo_VRB(ppal, ger, pres)13_merecer la_ART(f, sg)23_la pena_SUST(f, sg)33_pena_FRVRB(ppal, ger, pres)22_merecer la pena

(tiempo compuesto anidado en perífrasis anidada en perífrasis)

(16) ha_12 tenido_22_VRB(saux, ind, perf, 3, sg)13_tener que_CONJ(sub)23_que ir_VRB(saux, inf, pres)3312_ir haciendo_VRB(ppal, ger, pres)22_hacer

2.4.3 *DISCONTINUIDADES*

Sólo físicamente las discontinuidades representan una situación distinta de los compuestos. Aunque ello no sea trivial para el procesamiento, desde el punto de vista de la anotación, les es asignada una etiqueta exactamente igual que la que se asigna a los compuestos, salvo por una D (de *discontinuidad*) que sigue a los dígitos que marcan la composición como en un compuesto contiguo. Teóricamente podrían darse casos en los que quisiéramos etiquetar sólo componentes, sólo conjunto o ambos. En la práctica, para el español -otras aproximaciones podrían encontrar justificado asignar etiquetas de acuerdo con las otras dos posibilidades- nosotros estamos manejando tan sólo el segundo tipo de casos, anotación sólo del conjunto.

(17) o_12D_o ..... o_22D_o_FRCONJ(coor)22D_o...o

2.4.4 *UNIDADES MULTILEXICALES CIRCUNSTANCIALMENTE DISCONTINUAS*

Los compuestos que se presentan circunstancialmente como discontinuos, tales como perífrasis o tiempos compuestos en los que se introduce algún elemento externo, se etiquetan como compuestos no discontinuos.

(18) He_12 ya_AV(pos) insistido_22_VRB(ppal, ind, perf, 1, sg)_insistir

2.4.5 *CONTRACCIONES Y CLÍTICOS*

Unos y otros se etiquetan y lematizan como corresponde a cada uno de los integrantes de la contracción o la enclisis en el orden en que se integran en la contracción o la enclisis.

(19) al_PR_a_ART(m, sg)_el, del_PR_de_ART(m, sg)_el
(20) hazlo_VRB(ppal, imp, pres, 2, sg)_hacer_PERS(3, s, m, ac)_lo

Fenómenos como la contracción y la enclisis pueden ocurrir en compuestos. Si los elementos de la contracción o la enclisis pertenecen al compuesto se etiqueta el ítem contracto o enclítico como le corresponde en tanto que contracto o enclítico, es decir, añadiendo las etiquetas y lemas que corresponden a cada uno de los elementos de la contracción o enclisis en el orden en que aparecen en la misma. Ahora bien, dado que tales elementos, además, son integrantes de un compuesto a cada uno de ellos les serán asignados también los dígitos que los marcan como tales.

(21) María_SUST(prp, f, sg)14_María del_PR24_de_ART(m, sg)34_el Mar_SUST(prp, f, sg)44_Mar_FRSUST(prp, f, sg)_María del Mar

Si uno de los elementos contractos o enclíticos no pertenece al compuesto se procede exactamente del mismo modo, pero no se añaden los dígitos de la composición:

(22) en_PR13_en virtud_SUST(com, m, sg)23_virtud del_PR33_de_ART(m, sg)_el_FRPR_en virtud de
(23) haberlo_12_PERS(3, s, m, ac)_lo hecho_22_VRB(ppal, inf, perf)_hacer

3. REPRESENTACIÓN DEL ANÁLISIS MORFOSINTÁCTICO EN SGML

El esquema de codificación aplicado a los corpus (Pino 1996a) emplea SGML como lenguaje de marcación, y sigue las pautas de la TEI en la medida de lo posible. Por lo que se refiere a la información de tipo lingüístico, ha sido necesario elaborar una nueva propuesta, dado el escaso grado de desarrollo que presentan los modelos estándares de codificación en este terreno. En la TEI se sugieren varios sistemas de representación de datos lingüísticos, pero ninguno de ellos aparece tratado a fondo. En Ide & Véronis (1995), hay algunas aproximaciones nuevas, aunque siempre expuestas como trabajo en curso o como aplicaciones para fines muy específicos, de difícil estandarización. En consecuencia, ha sido preciso volver al punto de partida, el lenguaje SGML, y emprender, desde ahí, la búsqueda de un sistema de representación de datos morfosintácticos.

Los principales componentes del sistema de anotación morfosintáctica ya tenían, antes de comenzar esta tarea, algún modo de representación en la TEI. Tanto la palabra, unidad básica del análisis morfosintáctico, como su clasificación gramatical o su lema, se podían codificar de modos diversos. Las principales propuestas de la TEI eran las siguientes:

(a)Palabra: elemento <w>
Lema: elemento <lemma>
Morfema: elemento <m>
(b) Palabra: elemento <fs>
Análisis gramatical de la palabra: elementos <f>, que forman una estructura de rasgos dentro
del elemento <f>.
(c)Palabra: elemento <w>
Análisis gramatical de la palabra: atributo ***ana*** del elemento <w>: <w ana=>
Lema: atributo ***lemma*** del elemento <w>: <w lemma=>

La propuesta (c) parecía la más acertada porque evitaba recargar con demasiados elementos la codificación de tipo morfosintáctico y permitía, asimismo, no perder la linealidad del texto. Ahora bien, el procedimiento descrito en la TEI no resolvía numerosos casos de análisis más complejo. En el momento en que un grupo de palabras constituye una unidad, el sistema de la TEI obligaba a tratarlas separadamente. Dar cuenta de la relación entre varias palabras, ya estén continuas o discontinuas en el discurso, era algo no previsto. Sin embargo, para los propósitos lexicográficos de los corpus de la Academia, resultaba mucho más conveniente marcar el grupo y la palabra individual. A la inversa también surgían problemas, porque las palabras contractas sólo se

podían representar como un conjunto y no como palabras individuales. El sistema de representación de la TEI obligaba, como se ve, a una marcación enteramente condicionada por la linealidad del texto y basada en un concepto de palabra que se define como un conjunto de caracteres separados por espacios en blanco.

Antes todas estas restricciones, se optó por modificar los principios de codificación sin renunciar a los elementos y atributos básicos definidos por la TEI. Se elaboraron nuevas normas, v
'e1lidas para marcar todos los fenómenos morfosintácticos previstos en el sistema de anotación. En los siguientes epígrafes se proporcionan soluciones para cada caso especial, además de mostrar, con ejemplos, el sistema de representación de las palabras individuales.

3.1 CODIFICACIÓN DE LA INFORMACIÓN MORFOSINTÁCTICA EN CADA PALABRA

La información morfosintáctica asociada a cada palabra se representa por medio del elemento <w>, previsto en la TEI, y atributos que aportan datos relacionados con los siguientes parámetros:

PARÁMETRO	NOMBRE DEL ATRIBUTO
• Lema	*lemma*
• fs20	
• Análisis morfosintáctico (clase de palabra + atributos)	*ana*
• Tipo de palabra desde el punto de vista estructural:	
* relación constituto-constituyente	*type1*
* relación constituyente-constituto	*type2*

El atributo *lemma* indica el lema de la palabra. La información de análisis morfosintáctico, que coincide con la de la etiqueta diseñada en el esquema de anotación, se introduce como valor del atributo *ana*. Los atributos *type1* y *type2* caracterizan la palabra desde el punto de vista estructural. El primero de los dos atributos indica si la palabra es simple o si se trata de alguna de las variedades de unidad multilexical (lexías, colocaciones, tiempos compuestos, perífrasis, verbos con pronombre enclítico, contracciones o dicontinuidades). El segundo indica que la palabra forma parte de un determinado subtipo de unidad multilexical. En las palabras que no formen parte de unidades multilexicales, sólo se empleará el primero de los dos atributos *type*, puesto que sólo será necesario indicar sus componentes mosfosintácticos. En casos de construcciones formadas por varias palabras, que veremos en apartados sucesivos, es preciso mostrar también la relación existente entre cada palabra y el conjunto multilexical al que pertenece. Para ello se empleará el atributo *type2*. Los valores posibles de *type1* son estos:

Para palabras simples:
"sim" palabra simple
Para unidades multilexicales:
"lex" frases hechas o lexías

"col"	colocación de uso frecuente
"tco"	tiempo compuesto
"per"	perífrasis
"cli"	verbo con pronombre enclítico
"con"	contracciones
"dis"	discontinuidades

Los valores que admite *type2* son "lex", "col", "tco", "per" y "dis".

3.2 CODIFICACIÓN DE LA INFORMACIÓN MORFOSINTÁCTICA EN PALABRAS SIMPLES

Una palabra simple es aquella que no contiene ninguna palabra en su interior. Se representa como un elemento <w> sin subelementos.

(24) "bueno"
<w ana="ADJ(adj, pos, m, sg)" lemma="bueno" type1="sim">bueno</w>
(25) "amáis"
<w ana="VRB(ppal, ind, pres, 2, p)" lemma="amar" type1="sim">amáis</w>

No debe confundirse el concepto de "palabra simple" con el de "palabra individual". Que una palabra sea "simple" no significa que no pueda formar parte de una unidad multilexical. Las unidades multilexicales se componen de palabras simples, tal como se verá en el epígrafe siguiente. "Simple" se utiliza aquí, por tanto, con un sentido estructural, opuesto a "multilexical", mientras que "individual" puede entenderse como independiente, desde el punto de vista morfológico, de cualquier otra palabra presente en su mismo contexto. "Individual" se pone aquí a "morfológicamente vinculado". Claro está que las relaciones de concordancia quedan al margen de la consideración de la palabra como "individual" o "morfológicamente vinculada" a otras palabras.

3.3 CODIFICACIÓN DE LA INFORMACIÓN MORFOSINTÁCTICA EN UNIDADES MULTILEXICALES

La codificación de las unidades multilexicales es, en esencia, la misma de las palabras simples individuales, aunque presenta algunas dificultades que no aparecen en aquellas. El elemento que sigue marcando estas unidades es <w>, aunque aquí tiene también subelementos <w>. La unidad multilexical es un elemento <w> de nivel jerárquico más alto que el de sus componentes, que pueden ser, a su vez, más o menos complejos. Aparte de la posibilidad recursiva que caracteriza a estas unidades, que se trata en el apartado 3.3.8, existe una serie limitada de subtipos de formas multilexicales. Los siguientes epígrafes 3.3.1-3.3.7 muestran cómo se codifica en SGML cada subtipo.

3.3.1 *FRASES HECHAS O LEXÍAS*

(26) "manos a la obra"
<w ana="FRAV(pos)" lemma="manos a la obra" type1="lex">
<w ana="SUST(com, m, pl)" lemma="mano" type1="sim" type2="lex">manos</w>
<w ana="PR" lemma="a" type1="sim" type2="lex">a</w>
<w ana="ART(f, sg)" lemma="el" type1="sim" type2="lex">la</w>
<w ana="SUST(com, f, sg)" lemma="obra" type1="sim" type2="lex">obra</w>

```
</w>

(27)  "echaba de menos"
<w ana="FRVRB(ppal, ind, imp, 1|3, sg)" lemma="echar de menos" type1="lex">
   <w ana="VRB(ppal, ind, imp, 1|3, sg)" lemma="echar" type1="sim"
   type2="lex">echaba</w>
   <w ana="PR" lemma="de" type1="sim" type2="lex">de</w>
   <w ana="AV(cp)" lemma="menos" type1="sim" type2="lex">menos</w>
</w>
```

Debe observarse que cada unidad integrada en la lexía tiene valor en *type1* y en *type2*, mientras que el compuesto tiene sólo *type1*.

3.3.2 *COLOCACIONES DE USO FRECUENTE*

```
(28)  "reloj de pulsera"
<w type1="col">
   <w ana="SUST(com, m, sg)" lemma="reloj" type1="sim" type2="col">reloj</w>
   <w ana="PR" lemma="de" type1="sim" type2="col">de</w>
   <w ana="SUST(com, f, sg)" lemma="pulsera" type1="sim" type2="col">pulsera</w>
</w>
```

Es importante tener en cuenta que cuando el compuesto es una colocación de uso frecuente, no hay lema ni atributo ***ana*** para todo el compuesto, pero sí para cada componente.

3.3.3 *PERÍFRASIS*

```
(29)  "volver a hacer"
<w type1="per">
   <w ana="VRB(saux, inf, pres)" lemma="volver" type1="sim" type2="per">volver
   </w>
   <w ana="PR" lemma="a" type1="sim" type2="per">a</w>
   <w ana="VRB(ppal, inf, pres)" lemma="hacer" type1="sim" type2="per">hacer
   </w>
</w>
```

En las perífrasis no se especifica ni lema ni análisis morfosintáctico del compuesto, pero sí de cada componente, diferenciando el verbo semiauxiliar del verbo principal.

3.3.4 *TIEMPOS COMPUESTOS*

```
(30)  "he llegado"
<w ana="VRB(ppal, ind, perf, 1, sg)" lemma="llegar" type1="tco">
   <w type1="sim" type2="tco">he</w>
   <w type1="sim" type2="tco">llegado</w>
</w>
```

Los tiempos compuestos, sin embargo, no llevan indicación de lema ni de análisis morfosintáctico en cada uno de los componentes. Todos esos datos figuran una sola vez al comienzo del tiempo compuesto.

3.3.5 *VERBOS CON PRONOMBRE ENCLÍTICO*

```
(31)  "hacerlo"
<w type1="cli">hacerlo
    <w ana="VRB(ppal, inf, pres)" lemma="hacer" type1="sim"></w>
    <w ana="PERS(3, sg, m, ac)" lemma="lo" type1="cli"></w>
</w>
```

Los verbos con pronombre enclítico no llevan información de análisis morfosintáctico ni lema del compuesto, pero sí de cada componente. La forma, tal como aparece en el texto, se reproduce, en cambio, en el elemento <w> que abarca todo el compuesto.

3.3.6 *CONTRACCIONES*

```
(32)  "del"
<w type1="con">del
    <w ana="PR" lemma="de" type1="con"></w>
    <w ana="ART(f, sg)" lemma="el" type1="con"></w>
</w>
```

Las contracciones reciben un tratamiento semejante al de los verbos con pronombre enclítico.

3.3.7 *DISCONTINUIDADES*

```
(33)  "o ... o"
<w id="d1" ana="FRCONJ(coor)" lemma="o...o" type1="dis"></w>
<w target="d1" lemma="o" type1="sim" type2="dis">o</w>
(...)
<w target="d1" lemma="o" type1="sim" type2="dis">o</w>
```

Las discontinuidades se codifican como un elemento vacío que tiene, entre sus atributos, un identificador. Cada uno de los elementos discontinuos que componen la unidad multilexical envía una llamada o *link* al identificador del compuesto. De ese modo, es siempre posible localizar los segmentos que se correlacionan. Como puede verse en el ejemplo anterior, el tratamiento que se da al compuesto discontinuo es semejante al de las unidades multilexicales no discontinuas, puesto que existe especificación de lema tanto para el conjunto como para sus componentes.

3.3.8 *REPRESENTACIÓN DE FORMAS COMPUESTAS RECURSIVAS*

Como se mostró anteriormente (2.4.2), las formas multilexicales admiten recursividad. La codificación en SGML permite representar los compuestos anidados. El procedimiento básico es el mismo que ya se ha explicado para los casos no recursivos, y en él juegan un papel fundamental los atributos *type1* y *type2*, que fijan en todo momento la relación entre un constituto y sus constituyentes inmediatos. Veamos algunas muestras de la representación, en SGML, de unidades multilexicales anidadas de diverso tipo.

3.3.8.1 *Tiempo compuesto anidado en perífrasis*

(34) "ha habido que salir"

```
<w type1="per">
   <w ana="VRB(saux, ind, perf, 3, sg)" lemma="haber" type1="tco" type2="per">
      <w type1="sim" type2="tco">ha</w>
      <w type2="sim" type2="tco">habido</w>
   </w>
   <w ana="CONJ(sub)" lemma="que" type1="sim" type2="per">que</w>
   <w ana="VRB(ppal, inf, pres)" lemma="salir" type1="sim" type2="per">salir</w>
</w>
```

3.3.8.2 *Tiempo compuesto anidado en lexía*

(35) "ha echado de menos"

```
<w ana="FRVRB(ppal, ind, perf, 3, sg)" lemma="echar de menos" type1="lex">
   <w ana="VRB(ppal, ind, perf, 3, sg)" lemma="echar" type1="tco" type2="lex">
      <w type1="sim" type2="tco">he</w>
      <w type1="sim" type2="tco">echado</w>
   </w>
   <w ana="PR" lemma="de" type1="sim" type2="lex">de</w>
   <w ana="ADV(cp)" lemma="menos" type1="sim" type2="lex">menos</w>
</w>
```

3.3.8.3 *Tiempo compuesto y lexía anidados en perífrasis*

(36) "ha acabado por echar de menos"

```
<w type1="per">
   <w ana="VRB(saux, ind, perf, 3, sg)" lemma="acabar" type1="tco" type2="per">
      <w type1="sim" type2="tco">ha</w>
      <w type2="sim" type2="tco">acabado</w>
   </w>
   <w ana="PR" lemma="por" type1="sim" type2="per">por</w>
   <w ana="FRVRB(ppal, inf, pres)" lemma="echar de menos" type1="lex" type2="per">
      <w ana="VRB(ppal, inf, pres)" lemma="echar" type1="sim" type2="lex">echar
      </w>
      <w ana="PR" lemma="de" type1="sim" type2="lex">de</w>
      <w ana="ADV(cp)" lemma="menos" type1="sim" type2="lex">menos</w>
   </w>
</w>
```

3.3.8.4 *Perífrasis anidada en perífrasis*

(37) "empieza a ir comprendiendo"

```
<w type1="per">
   <w type1="per" type2="per">
      <w ana="VRB(saux, ind, pres, 3, sg)" lemma="empezar" type1="sim"
      type2="per">empieza
      </w>
      <w ana="PR" lemma="a" type1="sim" type2="per">a</w>
      <w ana="VRB(saux, inf, pres)" lemma="ir" type1="sim" type2="per">ir</w>
   </w>
   <w ana="VRB(ppal, ger, pres)" lemma="comprender" type1="sim"
   type2="per">comprendiendo</w>
   </w>
</w>
```

3.3.8.5 *Lexía anidada en perífrasis*

(38) "va mereciendo la pena"

```
<w type1="per">
   <w ana="VRB(saux, ind, pres, 3, sg)" lemma="ir" type1="sim" type2="per">va
   </w>
   <w ana="FRVRB(ppal, ger, pres)" lemma="merecer  la pena" type1="lex" type2="per">
      <w ana="VRB(ppal, ger, pres)" lemma="merecer" type1=sim" type2="lex">mereciendo
      </w>
      <w ana="ART(f, sg)" lemma="la" type1="sim" type2="lex">la</w>
      <w ana="SUST(f, sg)" lemma="pena" type1="sim" type2="lex">pena</w>
   </w>
</w>
```

3.3.8.6 *Tiempo compuesto anidado en una perífrasis anidada en perífrasis*

(39) "ha tenido que ir haciendo"

```
<w type1="per">
   < w type1="per" type2="per">
      <w ana="VRB(saux, ind, pres, 3, sg)" lemma="tener" type1="tco" type2="per">
         <w type1="sim" type2="tco">ha</w>
         <w type1="sim" type2="tco">tenido</w>
      </w>
      <w ana="CONJ(sub)" lemma="que" type1="sim" type2="per">que</w>
      <w ana="VRB(saux, inf, pres)" lemma="ir" type1="sim" type2="per">ir</w>
   <7w>
   <w ana="VRB(ppal, ger, pres)" lemma="hacer" type1="sim" type2="per">haciendo</w>
   </w>
</w>
```

3.3.8.7 *Contracciones y clíticos anidados en lexías*

(40) "María del Mar"

```
<w ana="FRSUST(prp, f, sg)" lemma="María del Mar" type1="lex">
   <w ana="SUST(prp, f, sg)" lemma="María" type1="sim" type2="lex">María</w>
   <w type1="con">del
      <w ana="PR" lemma="de" type1="con" type2="lex"></w>
```

```
        <w ana=ART(f, sg) lemma="el" type1="con" type2="lex""></w>
    </w>
    <w ana="SUST(prp, f, sg)" lemma="Mar" type1="sim" type2="lex">Mar</w>
</w>
```

(41) "en virtud del"

```
<w ana="FRPR" lemma="en virtud de" type1="lex">
    <w ana="PR" lemma="en" type1="sim" type2="lex">en</w>
    <w ana="SUST(com, f, sg)" lemma="virtud" type1="sim" type2="lex">virtud</w>
    <w type1="con">del
        <w ana="PR" lemma="de" type1="con" type2="lex"></w>
        <w ana=ART(f, sg) lemma="el" type1="con"></w>
    </w>
</w>
```

3.3.8.8 *Contracciones y clíticos anidados en tiempos compuestos*

(42) "haberlo hecho"

```
<w ana="VRB(ppal, inf, perf)" lemma="hacer" type1="tco">
    <w type1="cli">haberlo
        <w type1="sim" type2="tco"></w>
        <w ana="PERS(3, sg, m, ac) lemma="lo" type1="cli"></w>
    </w>
    <w type1="sim" type2="tco">hecho</w>
</w>
```

3.3.9 *UNIDADES MULTILEXICALES CIRCUNSTANCIALMENTE DISCONTINUAS*

Los compuestos que sólo con carácter circunstancial aparecen discontinuos, se representan del siguiente modo en SGML:

(43) "he ya insistido"

```
<w ana="VRB(ppal, ind, perf, 1, sg)" lemma="insistir" type1="tco">
    <w type1="sim" type2="tco">he</w>
    <w ana="AV(pos)" lemma="ya" type1="sim">ya</w>
    <w type1="sim" type2="tco">insistido</w>
</w>
```

4. CONCLUSIÓN

A lo largo de la exposición, se ha podido comprobar la aptitud del lenguaje SGML para la codficación lingüística de corpus. A partir de un sistema de etiquetas morfosintácticas que requieren, en principio, una representació3n lineal, directamente ligada a la palabra, se ha llegado a un modelo SGML de codificación. La propuesta retoma la etiqueta morfosintáctica prevista en el sistema de anotación, y la incorpora como valor de un atributo del elemento "palabra": <w ana="etiqueta">. Otro atributo del mismo elemento, ***lemma***, muestra el lema de la palabra. Se añaden, además, dos atributos de tipo y un identificador que caracterizan estructuralmente las palabras, al tiempo que relacionan los miembros de un compuesto, ya sea este continuo o discontinuo. La sintaxis SGML

permite dar cuenta de los diversos tipos de unidad multipalabra, así como de la recursividad.

5. BIBLIOGRAFÍA

Barnett, R. (1995), *Recommendations for forming a Spanish Tagset to conform with the EAGLES Guidelines*, Lancaster: Lancaster University, publicación interna.

Burnage, G., Dunlop, D. (1993), "Encoding the British National Corpus", in J. Aarts, P. de Haan and N. Oostdijk eds., *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam: Rodopi.

Burnard, L., Sperberg-McQueen, C. M. (1995), *TEI Lite: An Introduction to Text Encoding for Interchange.Document No: TEI U 5*, Groningen: Groningen University.

Garside, R., Leech, G., Sampson, G. eds. (1987), *The Computational Analysis of English: a Corpus-based Approach*, London: Longman.

Ide, N., Véronis, J. (1994), "Corpus Encoding", EAGLES Document EAG-CSG/IR-T2.1 en N. Calzolari, J. M. McNaught eds., *EAGLES Interim Report* EAG-EB-IR-2.

Ide, N., Véronis, J. eds. (1995), *The Text Encoding Initiative: Background and contexts. Computers and the Humanities*, 29, 1-3.

Leech, G, Wilson, A. (1994), "Morphosyntactic Annotation", Draft-Work In Progress, EAGLES Document EAG-CXG/IR.T3.1. en N. Calzolari, J. M. McNaught eds., *EAGLES Interim Report* EAG-EB-IR-2.

Nerc-1 Consortium: ILC-Pisa and University of Pisa (1994), "Linguistic Annotation of Texts: scientific and technical problems; guidelines for harmonization", en N. Calzolari, M. Baker, T. Kruyt eds., *NERC-, Network on European Reference Corpora. Final Report*, ILC-CNR-PISA, pp. 133-184.

Pino, M. (1996a), *Manual de codificación textual para los corpus CREA y CORDE. Normas de marcación en SGML según las recomendaciones de la TEI*, Versión 1.0., Madrid: RAE, publicaciones internas.

Pino, M. (1996b), "Encoding two large Spanish corpora with the TEI scheme: design and technical aspects of textual markup", Bethesda, Maryland: TEI Workshop at the ACM Digital Libraries '96. Accesible en URL: <http://WWW/cs.vassar.edu/~ide/DL96>.

Real Academia Española (1992), *Diccionario de la lengua española*,Madrid: Espasa-Calpe, Madrid.

Santalla del Río, Mª. P. (1995), *Anotación morfosintáctica*, Madrid: RAE, publicaciones internas.

Sánchez León, F. (1995), "Appendix I, Spanish Tagset currently used in the CRATER Project", en R. Barnett ed., *Recommendations for forming a Spanish Tagset to conform with the EAGLES Guidelines*, Lancaster: Lancaster University, publicación interna.

Sperberg-McQueen, C.M., Burnard, L. eds. (1994), *Guidelines for Electronic Text Encoding and Interchange. TEI-P3*, Chicago / Oxford: Text Encoding Initiative.