# Creación, etiquetación y desambiguación de un corpus de referencia del español

**Montserrat Civit Torruella**
TALP Research Center – Universitat Politècnica de Catalunya

**Irene Castellón Masalles** y **M. Antònia Martí Antonín**
Centre de Llenguatge i Computació – Departament de Lingüística
Universitat de Barcelona
civit@talp.upc.es, {castel,amarti}@fil.ub.es

**Resumen** En este artículo presentamos los criterios para la anotación y desambiguación morfosintáctica de un corpus de referencia del español que será de libre disposición, proponiendo respuestas concretas a casos problemáticos de ambas tareas. El objetivo final es disponer de una colección escrita de 1 millón de palabras desambiguadas manualmente a nivel tanto morfológico como sintáctico, que pueda utilizarse para el aprendizaje automático así como para la consulta lingüística. Discutimos detalladamente la categorización de las palabras del español así como los criterios lingüísticos de desambiguación.

## *1 Introducción*

En los últimos años se ha puesto de manifiesto la necesidad creciente de disponer de corpus lingüísticos etiquetados. Entre los trabajos más recientes realizados para el español cabe destacar [15], [14], [9] y [11], mientras que entre los trabajos más recientes en otras lenguas pueden citarse [3] y [4].

En este artículo presentamos los criterios para la anotación y desambiguación de un corpus para el español que será de libre disposición: el Corpus CLiC–TALP. Proponemos respuestas concretas a casos problemáticos de anotación y desambiguación que no se tratan tan en detalle en los trabajos realizados hasta el momento. El objetivo final es disponer de una colección escrita de 1 millón de palabras desambiguadas manualmente a nivel tanto morfológico como sintáctico. Los textos utilizados provienen de dos fuentes distintas: por un lado se han seleccionado 500.000 palabras procedentes de un corpus ya existente del español (LexEsp [18]) que constituyen una muestra representativa de la lengua, puesto que incluye muy diversa tipología textual (ensayo, literatura, texto científico, etc.); por otro, un conjunto de 500.000 palabras de estilo periodístico cedidas por el periódico *La Vanguardia*. Los fines de este corpus son fundamentalmente dos: por un lado proporcionar material desambiguado a mano para el entrenamiento de un tagger morfológico estadístico; por otro, constituir un corpus de referencia del español útil para la consulta e investigación lingüísticas.

La anotación se lleva a cabo automáticamente, mientras que el proceso de desambiguación es semiautomático (manual asistido).

En la sección 2 presentamos los criterios de anotación establecidos y en la 3 los criterios seguidos para la desambiguación.

## *2 Anotación*

El proceso de etiquetación puede dividirse en dos fases: anotación (asignación de todas las posibles etiquetas a cada palabra) y desambiguación (selección de la mejor etiqueta: etiqueta correcta). En esta sección nos ocuparemos de la primera.

La anotación se ha realizado de modo totalmente automático con la herramienta MACO ([7]). Dada una palabra, el analizador proporciona todas sus posibles etiquetas morfosintácticas junto con el lema de cada forma de palabra. El resultado de esta fase es la lematización y anotación (no desambiguada) de todas las palabras del corpus.

### 2.1 Categorías establecidas

Para el establecimiento del sistema de codificación de la información morfosintáctica se han tenido en cuenta las propuestas de estandarización realizadas por el grupo EAGLES ([12] y, especialmente, [13]). En este último caso ya se advierte (pp. 28-29.), que lo que proponen es un conjunto básico de rasgos que no admite una implementación directa, pues son el resultado de generalizar a partir de los

lexicones existentes en ese momento. Admiten, en el mismo sentido, la redundancia del sistema que proponen. Así, por ejemplo, se ofrecen dos formas de etiquetar los adjetivos demostrativos, como adjetivos (de tipo determinativo demostrativo) o como determinantes (demostrativos); lo mismo que ocurre con otras clases de palabras. Lo que aquí presentamos (y justificamos) es la adaptación de estas propuestas al español, con el objetivo de presentar un sistema de codificación completo y consistente.

No vamos a comentar aquí aquellos casos de adaptación directa de las propuestas a nuestro sistema, sino que sólo presentaremos aquellos casos que suponen cierta problemática, para discutir la solución adoptada.

Como regla general, anotamos todos los rasgos morfosintácticos que la palabra presenta. Cuando ello es posible, anotamos también información semántica de la palabra (cuando esta información puede obtenerse directamente de la forma de palabra, como por ejemplo, los diminutivos o la persona semántica de los posesivos). El objetivo es recoger toda la información que la palabra proporciona.

Las categorías con las que trabajamos son: *Adjetivo, Adverbio, Artículo, Determinante, Pronombre, Nombre, Verbo, Conjunción, Preposición, Interjección, Abreviatura, Signos de puntuación* y *Fechas*.

Las tres últimas no aparecen en el estándar (o se incluyen en la categoría *Residual*), sin embargo las utilizamos en esta adaptación porque el analizador morfológico contiene módulos que permiten su tratamiento. Por otra parte hemos integrado los numerales en los sistemas de pronombres y determinantes puesto que su distribución es, por lo general, la misma.

En lo referente a las demás categorías, hemos introducido dos variaciones. Siguiendo a [10] incluimos en la categoría *adjetivo* sólo los adjetivos calificativos. En la propuesta EAGLES se contemplaba la posibilidad de categorizar así también a los posesivos, indefinidos y numerales. En cambio, y tomando como referencia las propuestas teóricas mencionadas, hemos optado por separar ambas clases de palabras en dos categorías distintas: adjetivos y determinantes. Por otro lado, en la categoría *artículo* incluimos sólo el tradicionalmente llamado artículo definido. El indefinido se categoriza como determinante; la razón de esta decisión hay que buscarla en el hecho que es casi imposible, incluso para una persona, discriminar entre el uso de la palabra *un* como artículo o como determinante indefinido, por lo que hemos optado por categorizarlo sólo como determinante.

Los atributos para cada categoría se han adaptado a las características del español: sólo se utilizan aquellos atributos que son relevantes para esta lengua, entendiendo por relevantes aquellos que tienen un reflejo morfológico en la palabra. Por lo general se respeta el nivel 1 del estándar[1]. En las tablas 1, 2, 3 y 4 que aparecen al final del artículo, pueden observarse, respectivamente, las etiquetas para los adjetivos calificativos, los pronombres, los nombres y los verbos. Se dan algunos casos, especialmente en el verbo y en el pronombre, en que algunos atributos se implican mútuamente[2]. En lo referente a los atributos considerados para cada categoría, hemos hecho, como más representativas, las siguientes precisiones:

1. Adjetivo. No contemplamos el atributo de grado, ya que en español es un procedimiento fundamentalmente sintáctico[3]. Sin embargo, utilizamos este dígito para marcar los apreciativos (aumentativos, diminutivos y despectivos), puesto que es una información que la palabra posee y que consideramos útil para estudios lingüísticos.

   Hemos añadido al final un atributo booleano *Participio* que aparece marcado con el código *P* en los adjetivos de origen verbal: los participios. El objetivo de esta etiquetación es sintáctico, ya que los participios suelen mantener la estructura argumental verbal, y tenerlos marcados puede ayudar a la posterior desambiguación sintáctica de los sintagmas preposicionales.

2. Pronombre. No tratamos los pronombres reflexivos y recíprocos como tipos

[1] Recordemos que desde EAGLES se proponen tres niveles de anotación: *L0* que se corresponde con el nivel obligatorio; *L1* o nivel recomendado, y *L2* o nivel opcional.

[2] Por ejemplo, el género en el verbo sólo afecta a las formas de participio. En los demás casos el valor de este atributo será 0.

[3] Los comparativos sintéticos existen sólo para un grupo muy reducido de palabras.

específicos de pronombres, sino que quedan incluidos en el tipo personal, ya que no pueden distinguirse sin recurrir a la semántica de la oración.

La persona pronominal sólo se ha diferenciado en el caso de los pronombres personales, puesto que son los únicos que manifiestan concordancia en persona con el verbo. Para los demás pronombres, este atributo siempre tiene valor de tercera persona[4]. Para los posesivos se han establecido nuevos valores para el atributo *poseedor* con el fin de poder marcar la persona semántica a la que refieren.

En cuanto al atributo de caso, que sólo afecta a los pronombres personales, contemplamos la presencia del nominativo (sólo para las formas *yo, tú*), acusativo (sólo para *lo, la, los, las*), dativo (sólo para *le, les*) y oblicuo (para aquellas formas con preposición: *mí, ti, sí, conmigo, contigo, consigo*). El resto de pronombres personales no aparecen especificados para este atributo porque pueden aparecen en dos o más contextos distintos de caso[5].

Hemos utilizado el género neutro para las formas *esto, eso, aquello* y para un caso del pronombre acusativo *lo*. Tenemos dos etiquetas para esta forma: una masculina singular y otra neutra y de número invariable. La primera la utilizaremos para aquellos casos en que funciona como complemento directo (*Lo vio con sus propios ojos*), mientras que la segunda la utilizaremos para aquellos casos en que aparece en oraciones atributivas (*Lo ha sido desde siempre*).

Hemos utilizado el atributo *Politeness* para marcar las formas pronominales para el tratamiento de cortesía, a saber, *usted, ustedes, vos*.

3. Determinantes. Aparecen todos marcados con el valor 3 para la persona. Se ha utilizado el rasgo *poseedor* en los posesivos igual que con los pronombres.
4. Adverbio. Al igual que ocurría con el adjetivo, no hemos incluido el atributo de grado, porque el proceso morfológico afecta a muy pocas formas.
5. Preposiciones. Tratamos las contracciones (*al / del*) como formas de preposición.

Las locuciones se tratan como términos únicos y reciben la etiqueta correspondiente a su categoría gramatical (adverbio, conjunción o preposición).

## 2.2 Criterios para la categorización de las formas

En este apartado presentamos los criterios seguidos a la hora de adscribir las palabras a una categoría determinada. Comentamos también aquí el criterio seguido para la lematización de cada clase de palabras.

Consideramos *adjetivos* todas aquellas palabras pertenecientes a una clase abierta que tienen variación de género y/o número concordante ([6]): es decir, aquellas palabras cuyo género y número depende de otra (un sustantivo)[6]. El lema de los adjetivos es la forma masculina singular (o la singular si no hay variación de género).

Tratamos como *determinantes* aquellas palabras pertenecientes a clases cerradas que capacitan al nombre como expresión referencial ([10]: p. 133). Los tipos que consideramos son: demostrativos (las series de *este, ese, aquel* más la de *tal*); posesivos (*mi, tu, su; mío, tuyo, suyo,* etc.); interrogativos (*qué* y la serie de *cuánto*); indefinidos (las series correspondientes a *alguno, bastante, cada, cierto, cualquiera, cuanto, demasiado, diferente, distinto, mismo, mucho, ninguno, otro, poco, sendos, tanto, todo, un, varios*); cardinales (la serie de los números naturales más las formas de los distributivos, multiplicativos y partitivos); y, por último, los ordinales (*primero, segundo, ... penúltimo, último*). Igual que con los adjetivos, el lema asignado a los determinantes es la forma masculina singular plena o apocopada.

Consideramos *nombres* todas aquellas palabras con género inherente y número

[4] Si bien es cierto que algunos indefinidos de tercera persona pueden aparecen como sujetos de verbos en primera, como por ejemplo *Algunos recordamos a aquel ministro...*, hemos optado por no inespecificar este valor y resolver este problema en la sintaxis.

[5] Por ejemplo, la forma *nosotros* puede aparecen en contextos nominativos *Nosotros la contemplamos desde fuera del relato*; oblicuos *Las ratas conviven con nosotros*, dativos *No nos lo permitirían a nosotros*, etc.

[6] Como ya se ha mencionado anteriormente, establecemos la distinción entre adjetivos calificativos (aquí *adjetivos*) y determinativos (aquí *determinantes*).

semántico ([6]). Y establecemos dos grandes tipos, los comunes y los propios[7]. Además del género y el número también marcamos, como en el caso del adjetivo, las formas apreciativas. El lema del nombre es la forma singular, ya sea masculina o femenina. La justificación de esta decisión está en el hecho de que el género de los sustantivos es inherente, por lo que no se puede considerar que *niña* sea el femenino de *niño*[8].

Las palabras categorizadas como *pronombres* se corresponden fundamentalmente con las de los determinantes. La diferencia es que consideramos pronombres aquellas palabras que aparecen como núcleo de un constituyente. Esto significa que algunas palabras reciben una doble etiquetación (determinante y pronombre), mientras que otras (personales y relativos) sólo aparecen como pronombres. Los pronombres personales tienen tres lemas: *yo, tú, él* [9]

Las formas *verbales* son las que presentan los morfemas de tiempo, modo, persona y número (el género aparece sólo en las formas del participio). Distinguimos tres tipos: *auxiliar*, *semiauxiliar* y *principal*, que se corresponden, respectivamente, con *haber, ser* y el resto de formas verbales. Etiquetamos separadamente los tiempos compuestos de la conjugación así como las perífrasis verbales. Se trata de no perder información, cosa que sucedería en una etiquetación conjunta de las dos formas de los tiempos compuestos en que no quedaría reflejado el auxiliar. Sin embargo, el tratamiento unitario de estas formas se lleva a cabo en la sintaxis.

Consideramos *adverbios* aquellas palabras dotadas de contenido léxico que son invariables y que tienen como función primordial la modificación verbal. Reciben esta etiqueta un total de 97 palabras (además de los acabados en *-mente* que se tratan en un postproceso).

Las *preposiciones* son elementos de relación que enlazan constituyentes sintagmáticos. Las *conjunciones* son elementos de relación que enlazan constituyentes oracionales. Las *interjecciones* son aquellas expresiones de la lengua que pueden utilizarse en un contexto determinado para expresar la función emotiva o apelativa del lenguaje.

Los lemas de estas cuatro últimas categorías son las propias palabras, puesto que ninguna de ellas presenta ningún tipo de variación morfológica.

Finalmente, también se han establecido códigos para marcar los diferentes *signos de puntuación.*

## *3 Desambiguación*

Comentamos en esta sección las principales ambigüedades resultantes tras la anotación automática del corpus, así como los criterios seguidos para su desambiguación. Antes de entrar en el detalle de este problema, comentaremos brevemente cómo se lleva a cabo este proceso. La desambiguación se hace manualmente, con un editor que muestra todas las posibles etiquetas de la palabra, pero que proporciona en primer lugar la que el tagger ([16]) considera más adecuada, de modo que, si es la acertada, el corrector no tiene más que validarla. En caso contrario, se puede seleccionar otra de las etiquetas propuestas para esa palabra, o incluso, es posible seleccionar una de entre toda la colección de etiquetas disponibles.

A grandes rasgos podemos establecer dos tipos de ambigüedades resultantes de esta etiquetación. Por un lado puede que una forma de palabra reciba etiquetas correspondientes a dos categorías distintas. Así, por ejemplo, la palabra *joven* puede interpretarse como adjetivo o como nombre[10]. Llamaremos a este tipo *ambigüedad categorial.* Por otra lado, la ambigüedad puede darse en diferentes valores de uno o más atributos en el seno de una misma categoría, lo que ocurre, por ejemplo en el nombre *cometa* que puede ser masculino o femenino, o en *cantamos* que es una forma verbal en tiempo presente o pasado. Llamaremos a este segundo tipo *ambigüedad intracategorial.* Esta última, puede resolverse en algunos casos de modo automático[11],

[7] En la actualidad se está desarrollando un módulo para etiquetar semánticamente los nombres propios.

[8] Del mismo modo que *puerta* no es el femenino de *puerto.*

[9] Otra opción hubiera sido tratarlos como el nombre en cuanto al género (cf. supra) pero ello implicaba casos de indeterminación: por ejemplo, ¿cómo decidir cuál es el lema de las formas *le, se, nos*? Además, puesto que en la primera y segunda persona del singular no hay distinción de género, tampoco podría hacerse para las formas *nosotros/nosotras, vosotros/vosotras.*

[10] Respectivamente en *Es una mujer joven* o *Vino una joven.*

[11] Por ejemplo, en la asignación de género al sustantivo *cometa* la primera opción que propone el editor es la que tiene en cuenta el género del determinante

aunque no siempre es así. La resolución de la ambigüedad temporal sólo puede hacerla el corrector humano. Nos ocuparemos en adelante de la ambigüedad categorial.

**Ambigüedades categoriales.** Las principales (que no las únicas) ambigüedades categoriales son: adjetivo / participio; nombre / adjetivo; nombre / verbo, y relativo / conjunción. Un caso que comentamos aparte, es la ambigüedad de la palabra *se*. Vamos a verlas con más detalle.

*Adjetivo / participio.* Este tipo de ambigüedad afecta a un gran número de palabras[12]. Por ejemplo, la forma *cansado* es adjetivo en *Viene cansado* puesto que alterna con *Viene contento*[13], mientras que es verbo participio en *Se ha cansado tras la compra.* Reservamos la etiqueta de forma verbal participia (*vmp00..*) para aquellas formas que aparecen en los tiempos compuestos de la conjugación (siempre precedidas de formas del verbo *haber*) y para las formas pasivas del verbo (precedidas del verbo *ser*). En el resto de los casos, la etiqueta que reciben es *aq0..p.* Esto puede expresarse para el desambiguador con una regla como la siguiente[14]:

```
1.0 (<VMP*>)
       (0 (<AQ*>))
       (-1 (<VA*>));
```

que puede parafrasearse del siguiente modo: si una palabra tiene la etiqueta correspondiente a participio (VMP*) y a la vez tiene la etiqueta correspondiente a adjetivo (AQ*) será participio si la palabra anterior tiene la etiqueta correspondiente al verbo *haber* (VA*). En el caso de las pasivas, aparece `(-1 (<VS*>))` en la última línea de la regla.

*Nombre / adjetivo.* También muchas palabras, como se vio anteriormente con el ejemplo de *joven*, funcionan en el discurso como nombres o adjetivos. Según Ignacio Bosque ([5]) los nombres categorizan o determinan clases de objetos mientras que los adjetivos describen propiedades que no constituyen clases. Ahora bien, hay propiedades, sobre todo de las personas, que son lo bastante representativas como para formar clases de individuos, como *ciego, hablante, salvaje, industrial, turco.* Como señala este autor, la conversión de propiedades en clases depende de factores extralingüísticos, por lo que este proceso no puede preverse a nivel léxico. Los criterios para la desambiguación son los siguientes: (i) dadas dos palabras contiguas que presenten esta doble posibilidad de etiquetación la de la izquierda será nombre y la de la derecha adjetivo; (ii) dada una palabra con doble etiquetación, será adjetivo si la siguiente o la anterior son sustantivos, y será nombre en caso contrario; (iii) dada la situación de (ii), la palabra con doble etiqueta será nombre si la palabra anterior es el indefinido *un* y la siguiente no es un nombre. El último caso puede expresarse con la regla:

```
1.0 (<NC*>)
       (0 (<AQ*>))
       (-1 (''un'' <DI*>))
       (NOT 1 (<NC*>));
```

Este tipo de ambigüedad afecta a un total de 40275 formas de palabra.

*Nombre / verbo.* Esta clase de ambigüedad esconde dos problemas distintos. Por una parte está al ambigüedad nombre / infinitivo, que aquí no tratamos extensamente. Sólo están etiquetadas doblemente aquellas formas de infinitivos que ya son sustantivos de la lengua, es decir, que ya han adoptado las características morfológicas propias del sustantivo, como por ejemplo *haber, deber, andar,* etc. En estos casos la presencia de un determinante o adjetivo previos es decisiva para la interpretación nominal de la forma. Los demás casos aparecerán siempre con la etiqueta verbal[15] El otro tipo de ambigüedad es el que afecta a formas como *cometa* que puede ser o bien sustantivo o bien una forma subjuntiva del verbo *cometer.* Aquí la palabra inmediatamente anterior suele resolver el problema, aunque en una caso como *la cometa* se necesita más información, como por ejemplo, saber si en el seno de la misma oración hay otra forma verbal (sin barreras como las conjunciones o los relativos).

*Relativo / conjunción.* Esta ambigüedad afecta a las palabras *que, como, donde,* y

que le precede.

[12] Nuestro diccionario dispone actualmente de unas 15.000 formas flexionadas que reciben esta doble etiquetación.

[13] Ejemplos adaptados de [17]: p. 494

[14] Para el formalismo de expresión de las reglas, basado en *Constraint Grammars*, cf. [16]

[15] Sin embargo, este es un caso en que el corrector deberá introducir la etiqueta de nombre común para aquellos casos en que el infinitivo se comporta como un nombre exigiendo complementos de tipo nominal *el andar de María*, frente a los casos de *el andar María* donde la forma es plenamente verbal [5].

*cuando*. Como relativos son anafóricas, mientras que como conjunción no. Éste es uno de los casos más complejos de desambiguación y debe hacerse totalmente a mano. Por lo general el tagger propone como primera opción la de relativo si la palabra anterior es un nombre (cadena mucho más frecuente que la de nombre–conjunción), pero no siempre es una previsión acertada. Por ejemplo, en la frase *Le dijo a su amigo que no llegara tarde* la palabra *que* es una conjunción y no un relativo.

Por último comentamos la ambigüedad que afecta a la palabra *se*. El etiquetador la categoriza siempre como PP3CN000 y con lema *él*. Sin embargo, aquí el corrector puede introducir dos nuevas etiquetas: P0300000 con lema *él* y P03CN000 con lema *se*. La primera etiqueta (PP3CN000) está reservada a aquellos casos en que esta forma es un pronombre personal con una función sintáctica determinada. La segunda (P0300000), es la que debe acompañar a los verbos pronominales puros (como *quejarse, atreverse*) o alternantes (como *acordarse, irse, casarse*). La última, debe aparecer sólo en los contextos en que *se* es una marca oracional de impersonalidad o de pasiva refleja. Así el corpus podrá ser utilizado para consultas de tipo lingüístico y en un aprendizaje posterior el tagger podrá inducir las diversas interpretaciones de esta palabra en función del contexto, pero teniendo en cuenta ya no la etiqueta de la palabra siguiente sino la propia palabra[16].

Para establecer explícitamente los criterios de desambiguación manual se ha elaborado una guía de anotación ([8]), al estilo de [1] y de [2] que recoge todos los casos más o menos conflictivos con abundantes ejemplos. Los criterios que se han seguido son fundamentalmente morfosintácticos.

## 4 Conclusión

Tomando como punto de partida el sistema de anotación morfosintáctica de EAGLES, hemos presentado una propuesta concreta de codificación de rasgos para el español. Asimismo, hemos presentado los criterios utilizados para la anotación y desambiguación de corpus, haciendo un especial hincapié en los problemas que presenta, por una parte, la adscripción de las palabras a una categoría gramatical determinada, y, por otra, el establecimiento de criterios de base lingüística para la desambiguación. Además, también ponemos a disposición de la comunidad lingüística un sistema de etiquetación completo siguiendo los estándares establecidos que puede utilizarse como punto de referencia para trabajos posteriores.

[16] Puesto que el tagger realiza la desambiguación a partir de las etiquetas morfosintácticas, no de las formas de palabra.

## 5 Agradecimientos

Este trabajo ha sido parcialmente financiado por una beca FPU (AP98-39864555), por la CICYT TIC98-0423-C06 y por X-Tract (PB98-1226). Agradecemos asimismo al periódico *La Vanguardia* la cesión de los textos que constituirán el corpus.

## Referencias

[1] A. Abeillé and L. Clément. Désambiguation morpho-syntaxique. guide pour les annotateurs - mots composés. corpus le monde. available http://talana.linguist.jussieu.fr/ lionel/corpus/convention.html.

[2] A. Abeillé and L. Clément. Désambiguation morpho-syntaxique. guide pour les annotateurs - mots simples. corpus le monde. available http://talana.linguist.jussieu.fr/ lionel/corpus/convention.html.

[3] A. Abeillé, L. Clément, and A. Kinyon. Building a treebank for French. In *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 87–94, Athens, Greece, 2000.

[4] C. Bosco, V. Lombardo, D. Vassallo, and L. Lesmo. Building a treebank for Italian: a Data-driven Annotation Schema. In *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 99–105, Athens, Greece, 2000.

[5] I. Bosque. *Las categorías gramaticales*. Number 11 in Textos de Apoyo. Lingüística. Ed. Síntesis, 1991 (tercera reimpresión).

[6] I. Bosque. El nombre común. In I. Bosque y V. Demonte, editor, *Gramática Descriptiva de la Lengua Española*, pages 3–75. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, 1999.

[7] J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, pages 915–922, Granada, 1998.

[8] M. Civit. Guía para la anotación morfológica de corpus. Technical Report X-Tract WP-00/06, Universitat de Barcelona, 2000.

[9] A. Martín de Santa Olalla Sánchez. *Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española*, volume 3. Estudios de Lingüística Española (ELiEs), 1999. available: http://elies.rediris.es/elies3/.

[10] V. Demonte. El Adjetivo. Clases y usos. La posición del adjetivo en el sintagma nominal. In I. Bosque y V. Demonte, editor, *Gramática Descriptiva de la Lengua Española*, chapter 3, pages 129–215. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, 1999.

[11] J.L. Sancho A. Nieto A. Ballester A. Fernández J. Gómez L. Gómez E. Raigal R. Ruiz F. Sánchez, J. Porta. La anotación de los corpus CREA y CORDE. *Revista de la Sociedad Española de Procesamiento de Lenguaje Natural*, (25):175–182, Septiembre 1999.

[12] G. Leech and A. Wilson. Recommendations for the Morphosyntactic Annotation of Corpora. Technical report, EAGLES, Mar 1996. available: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

[13] M. Monachini and N. Calzolari. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. Technical report, EAGLES, 1996. available: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

[14] A. Moreno, R. Grishman, S. López, F. Sánchez, and S. Sekine. A Treebank of Spanish and its Application to Parsing. In *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 107–111, Athens, Greece, 2000.

[15] A. Moreno and S. López. Developing a Spanish TreeBank. Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999. available: http://talana.linguist.jussieu.fr/treebanks99/.

[16] Lluís Padró. *A Hybrib Environment for Syntax-Semantic Tagging.* PhD thesis, Software Department (LSI). Technical University of Catalonia (UPC), 1997.

[17] RAE. *Esbozo de una nueva gramática de la lengua española.* RAE, 1973.

[18] N. Sebastián, M.A. Martí, M.F. Carreiras, and F. Cuetos. *LEXESP: Léxico Informatizado del Español.* Edicions de la Universitat de Barcelona, 2000.

## *A Ejemplo de corpus etiquetado y desambiguado*

No no RG000
quiero querer VMIP1S0
decir decir VMN0000
que que CS00
lo él PP3CNA00
sea ser VAMP3S0
, , Fc
cínico cínico AQ0MS00
o o CC00
divertido divertir AQ0MS0P
, , Fc
sino_que sino_que CS00
ante ante SPS00
un un DI3MS00
mazo mazo NCMS000
de de SPS00
hojas hoja NCFP000
grabadas grabada AQ0FP0P
coloca colocar VMIP3S0
un un DI3MS00
cristal cristal NCMS000
bien bien RG000
tallado tallado AQ0MS0P
y y CC00
lo él PP3CSA00
hace hacer VMIP3S0
girar girar VMN0000
. . Fp

## *B Las etiquetas utilizadas*

| Atributo | Valor | Código |
|---|---|---|
| Categoría | Adjetivo | A |
| Tipo | Calificativo | Q |
| Apreciativo | Sí | A |
| Género | Masculino | M |
| | Femenino | F |
| | Común | C |
| Número | Singular | S |
| | Plural | P |
| | Invariable | N |
| Participio | Sí | P |

Tabla 1: Etiquetas para los calificativos

| Atributo | Valor | Código |
|---|---|---|
| Categoría | Pronombre | P |
| Tipo | Personal | P |
| | Demostrativo | D |
| | Posesivo | P |
| | Interrogativo | T |
| | Relativo | R |
| | Indefinido | I |
| | Cardinal | C |
| | Ordinal | O |
| Persona | Primera | 1 |
| | Segunda | 2 |
| | Tercera | 3 |
| Género | Masculino | M |
| | Femenino | F |
| | Neutro | N |
| | Común | C |
| Número | Singular | S |
| | Plural | P |
| | Invariable | N |
| Caso | Nominativo | N |
| | Acusativo | A |
| | Dativo | D |
| | Oblicuo | O |
| Persona semántica | Primera-sg | 1 |
| | Segunda-sg | 2 |
| | Tercera | 0 |
| | Primera-pl | 4 |
| | Segunda-pl | 5 |
| *Politeness* | Sí | P |

Tabla 2: Etiquetas para el pronombre

| Atributo | Valor | Código |
|---|---|---|
| Categoría | Nombre | N |
| Tipo | Común | C |
| | Propio | P |
| Género | Masculino | M |
| | Femenino | F |
| | Común | C |
| Número | Singular | S |
| | Plural | P |
| | Invariable | N |
| – | – | 0 |
| – | – | 0 |
| Apreciativo | Sí | A |

Tabla 3: Etiquetas para el nombre

| Atributo | Valor | Código |
|---|---|---|
| Categoría | Verbo | V |
| Tipo | Principal | M |
| | Semiauxiliar | S |
| | Auxiliar | A |
| Modo | Indicativo | I |
| | Subjuntivo | S |
| | Imperativo | M |
| | Infinitivo | N |
| | Gerundio | G |
| | Participio | P |
| Tiempo | Presente | P |
| | Imperfecto | I |
| | Condicional | C |
| | Futuro | F |
| | Pasado | S |
| Persona | Primera | 1 |
| | Segunda | 2 |
| | Tercera | 3 |
| Número | Singular | S |
| | Plural | P |
| Género | Masculino | M |
| | Femenino | F |

Tabla 4: Etiquetas para el verbo