Análisis cualitativo y cuantitativo del acuerdo entre anotadores en el desarrollo de corpus interpretados lingüísticamente*

M. Civit†, A. Ageno‡, B. Navarro†, N. Bufí†, M.A. Martí†
†CLiC Centre de Llenguatge i Computació
Adolf Florensa s/n (Torre Florensa) 08028 Barcelona
{civit, nuria}@clic.fil.ub.es; amarti@fil.ub.es
‡TALP Research Centre (UPC)
Jordi Girona nº 3 08034 Barcelona
ageno@lsi.upc.es
† Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante Campus de San Vicente del Raspeig
Apartado 99. 03080 Alicante
borja@dlsi.ua.es

Resumen: El objetivo de este trabajo es presentar un análisis cualitativo y cuantitativo de las discrepancias entre anotadores en el etiquetado sintáctico del corpus Cast3LB. Para ello se ha definido un corpus de prueba de mil oraciones que ha sido etiquetado paralelamente por cinco anotadores. Se han realizado sucesivas evaluaciones de los resultados que han dado lugar a otras tantas mejoras de la guía de anotación hasta su versión definitiva. En una última fase, se analizan cualitativamente y se clasifican las discrepancias entre anotadores.
Palabras clave: Anotación sintáctica, corpus, acuerdo entre anotadores.

Abstract: The main goal of this work is to present a qualitative and quantitative analysis of disagreements among annotators during the syntactic labeling of the Cast3LB corpus. To do so, a one-thousand-sentence corpus has been established and it has been annotated by five annotators. Consecutive evaluations of the results have been done and have led to succesive improvements of the guidelines. In the last phase, we present the qualitative analysis and the classification of the differences among annotators.
Keywords: Syntactic annotation, corpus, annotators' agreement.

1. Introducción

El objetivo de este trabajo es presentar un análisis cualitativo y cuantitativo de las discrepancias entre anotadores en el etiquetado sintáctico del corpus Cast3LB. Este corpus, que actualmente está en fase de desarrollo dentro del proyecto general 3LB, consta de 100 000 palabras en español, de las cuales se ha etiquetado ya a nivel sintáctico más de un 25 por ciento.

El desarrollo de corpus interpretados lingüísticamente (*Treebanks*) va ligado a tres elementos: el desarrollo de sistemas automáticos para el análisis sintáctico; la especificación de esquemas de anotación, dado que debe imponerse un análisis consistente de los datos; y la creación de métricas para cuantificar la precisión en el análisis. Por lo general, estas métricas, cuyas primeras definiciones aparecieron en los *workshops* Parseval, cuantifican el grado de precisión de un cierto análisis respecto de un *gold-estandard* preestablecido y se han utilizado principalmente para comparar distintos sistemas de análisis sobre un mismo corpus de referencia. El objetivo de estas métricas es proporcionar datos sobre la similitud entre los análisis, pero en ningún caso proporcionan información sobre la localización de los desacuerdos ni sobre la naturaleza de los mismos.

La problemática del etiquetado de corpus es compleja. Por una parte, porque las expresiones lingüísticas que aparecen en los corpus (reflejo del uso de la lengua) plantean problemas que muchas veces no se ven reflejados en las gramáticas o que aparecen tratados de modo parcial. Por otra parte, una misma estructura lingüística es susceptible de

* Este trabajo ha sido parcialmente financiado por los proyectos PROFIT (FIT-15 0500-2002-244) y X-TRACT-II (BFF2002-04226-C03-03)

diferentes interpretaciones, todas ellas correctas. Por último, cada individuo tiene su propia concepción del lenguaje y lo interpreta de una forma determinada.

Dado que la labor de anotación de corpus es un trabajo de equipo es importante poder evaluar el grado de consistencia en los análisis proporcionados por distintos anotadores. La consistencia en la anotación es necesaria para que aumente la calidad del corpus así como su utilidad tanto para el entrenamiento y test de sistemas de análisis automático, como para la investigación lingüística.

Por otra parte, hasta ahora, se ha prestado poca atención a otro elemento que interviene en el desarrollo de los *Treebanks* y que hace referencia a la capacidad de los anotadores para el análisis sintáctico: *how precisely can human beings analyse language structure?* (Sampson y Babarczy, 2003).

Hasta la fecha, no hay estudios en profundidad sobre la consistencia entre anotadores a nivel sintáctico[1] , aunque puede mencionarse el trabajo de Brants (2000) sobre el acuerdo entre anotadores en el proyecto NEGRA (Brants, Skut, y Uszkoreit, 2003). En la actualidad G. Sampson y A. Babarczy están llevando a cabo un experimento para valorar cualitativemente las discrepancias entre anotadores, sobre un fragmento del BNC, utilizando el esquema de anotación diseñado para el corpus SUSANNE (véase, para más detalles, Sampson y Babarczy (2003)).

Como indican Sampson y Babarczy (2003), existe un límite en la capacidad humana para analizar con precisión su propia lengua y, por consiguiente, existe un límite en la precisión de la anotación humana. Así, en la construcción de un corpus anotado lingüísticamente, el objetivo es minimizar este margen de desacuerdo entre anotadores y acercarse a este límite todo lo posible. No debe olvidarse que el modo en que los humanos resolvemos estas cuestiones constituye el criterio de referencia para el análisis automático del lenguaje.

En este trabajo se presenta un estudio sobre el acuerdo entre los anotadores del corpus Cast3LB a nivel sintáctico.

Para poder llevar a cabo este estudio, se han analizado 1000 oraciones por parte de cinco personas distintas en fases sucesivas (cf. sección 3). Se ha desarrollado en paralelo una guía de anotación (Civit, 2002) que incluye una descripción del etiquetario utilizado así como una detallada casuística de los problemas que pueden surgir y de las soluciones que deben adoptarse en cada caso.

El primer objetivo de este trabajo es estudiar, desde un punto de vista cuantitativo, el acuerdo que existe entre anotadores en el etiquetado sintáctico a nivel de constituyentes siguiendo la aplicación de la guía de anotación. Con ello se obtienen medidas que cuantifican el grado de acuerdo y, por tanto, la consistencia en la anotación. Consideramos que unos resultados superiores al 90 % son aceptables para considerar que la anotación que proporcionamos es una anotación consistente y que, por tanto, la utilidad del corpus queda consolidada (véase la sección 3).

El segundo objetivo de este trabajo es estudiar el acuerdo entre anotadores desde un punto de vista cualitativo, con la finalidad de clasificar y analizar los casos concretos de discrepancias. Seguimos para ello la propuesta de Sampson y Babarczy (2003) (véase la sección 4).

Antes de entrar en las comparaciones y de presentar los resultados, presentamos brevemente el proyecto en el que se enmarca el trabajo así como los criterios básicos de anotación (sección 2).

2. *El proyecto Cast3LB*

El proyecto Cast3LB forma parte de un proyecto más amplio, 3LB, cuyo objetivo es construir tres corpus anotados con información lingüística, uno para el euskera (Eus3LB), otro para el catalán (Cat3LB) y otro para el castellano (Cast3LB).

Por lo que respecta a Cast3LB, la información que se está etiquetando corresponde a los siguientes cuatro niveles de descripción lingüística[2]:

- Nivel de forma sintáctica, en el que se parentizan y etiquetan los constituyentes sintácticos;
- Nivel de función sintáctica, en el que se etiqueta la función de los principales constituyentes de cada oración;

[1] Sí hay trabajos sobre la consistencia en la anotación semántica (Véronis, 2000) y la morfológica (Babarczy, Carroll, y Sampson, 2001).

[2] Más detalles sobre la anotación de este corpus aparecen en Navarro et al. (2003)

- Nivel semántico, en el que se etiqueta el sentido desambiguado de las palabras (nombres, adjetivos, verbos y algunos adverbios) a partir de EuroWordNet;
- Nivel pragmático, en el que se etiquetan las principales anáforas y elementos correferenciales del corpus, así como sus antecedentes (cadenas de correferencia).

El corpus de Cast3LB está formado por 75.000 palabras extraídas del corpus CLIC-TALP (Civit, Castellón, y Martí, 2001) –que, a su vez, es un fragmento del corpus LexEsp (Sebastián et al., 2000)–, y 25.000 palabras procedentes de un corpus de noticias cedido por la Agencia Efe.

El fragmento correspondiente al corpus CLiC-TALP consta de textos de procedencia muy heterogénea (periodísticos, literarios, científicos, etc.), extraídos de diferentes zonas de habla hispana (tanto de España como de Hispanoamérica), lo que lo convierte en un corpus representativo de la situación actual del español. Este corpus, además, está anotado con información morfológica (PoS)[3] y ha sido validado manualmente, lo que nos permite partir ya de un análisis lingüístico correcto. No ocurre lo mismo con las noticias de la Agencia Efe, porque el proceso de desambiguación morfológica ha sido automático y no se ha validado. Esta parte del corpus es comparable, en cuanto al contenido, con los corpus que se están utilizando en el proyecto para el catalán y el euskera.

2.1. Anotación sintáctica de Cast3LB

La anotación sintáctica del corpus Cast3LB se lleva a cabo en dos fases: en la primera se parentizan y etiquetan los principales constituyentes de la oración, mientras que en la segunda se asigna a cada uno de los constituyentes principales una etiqueta de función sintáctica.

Los principios básicos para la primera fase de esta anotación son los siguientes[4]:

- Sólo se etiquetan los elementos explícitos de las oraciones. Sin embargo, y puesto que en el futuro está prevista la anotación de cadenas de correferencia[5] hemos optado por introducir un nodo especial para los sujetos elípticos de las oraciones finitas. En lo referente a la elipsis verbal, la marcamos añadiendo un sufijo (*****) a las etiquetas de las oraciones.
- El orden de aparición de los elementos en la oración no se altera. El español es una lengua de orden libre por lo que respecta a los constituyentes de la oración, de tal manera que el orden específico en el que aparecen en la oración responde a motivaciones diversas de carácter funcional o comunicativo. Por ello, alterar el orden de los elementos significaría la pérdida de esta información.
- Se ha seguido un esquema de anotación basado en constituyentes, frente a la opción de anotación de dependencias.
- Se ha tratado que el etiquetado sea lo más neutro posible, sin seguir ninguna teoría lingüística ni ningún marco teórico concreto. Esta decisión se tomó con el propósito de que el corpus anotado fuera apto para llevar a cabo investigaciones lingüísticas y computacionales sin ningún tipo de restricción.

Para realizar la anotación sintáctica partimos del corpus previamente analizado y desambiguado morfológicamente y con anotación de *chunks* , de modo que el trabajo de los anotadores se ha centrado en la construcción de los constituyentes de las oraciones (parentización) y en la asignación de la etiqueta sintáctica correcta. Para facilitar esta labor, utilizamos una interfaz de anotación (Cotton y Bird, 2000) que permite la adición y borrado de nodos, el cambio de etiquetas, nuevos niveles de anidamiento, etc.

El número de etiquetas que utilizamos para los constituyentes es de 91, algunas de las cuales aparecen en el anexo 1 del artículo.

3. Comparación: análisis cuantitativo

No existiendo medidas específicas para la comparación cuantitativa del acuerdo entre anotadores, se ha decidido usar alguna de las métricas utilizadas para la evaluacion de gramáticas y/o métodos de

[3] Véase (Civit, Castellón, y Martí, 2001)

[4] Estos principios generales de anotación sintáctica aparecen con detalle en (Civit y Martí, 2002) y (Civit et al., 2003) y son los mismos que se aplican al corpus del catalán Cat3LB.

[5] Véase página 3.

análisis. La necesidad de una evaluación rigurosa a la hora de desarrollar analizadores de amplia cobertura es claramente reconocida. Queda fuera del alcance de este artículo entrar a describir en detalle los diferentes sistemas de evaluación (se pueden consultar por ejemplo dos excelentes revisiones de los diferentes métodos definidos a partir de 1991, (Carroll, Briscoe, y Sanfilippo, 1998) y (Bangalore et al., 1998)). En nuestro caso, se ha decidido utilizar las que se pueden considerar las primeras medidas objetivas, las definidas en los workshops Parseval (Black et al., 1991), para evaluar sintácticamente analizadores de amplia cobertura para el inglés. Aún no siendo exclusivo, su uso está bastante estandarizado para la evaluación de gramáticas y/o métodos de análisis, comparando la similitud de los resultados obtenidos con los árboles de análisis de referencia (los previamente considerados *correctos*, que en inglés se conocen como *gold standard*). Estas medidas de similitud se basan en la comparación de los constituyentes de ambos árboles de análisis, tanto en lo que se refiere a sus límites (punto de inicio y final en la frase), como a su etiqueta. Las medidas concretas que se han utilizado se definen a continuación:

- **Ratio de Precisión Etiquetada** (*Labelled Precision Rate*): Número de constituyentes del árbol de análisis evaluado que coinciden completamente (tanto sus límites como su etiqueta) con algún constituyente del árbol de análisis de referencia, dividido por el número total de constituyentes del árbol de análisis evaluado.
- **Ratio de Precisión Parentizada** (*Bracketed Precision Rate*): Número de constituyentes del árbol de análisis evaluado cuyos límites coinciden con los de algún constituyente del árbol de análisis de referencia, dividido por el número total de constituyentes del árbol de análisis evaluado.
- **Ratio de Cobertura Etiquetada** (*Labelled Recall Rate*): Número de constituyentes del árbol de análisis evaluado que coinciden completamente (tanto sus límites como su etiqueta) con algún constituyente del árbol de análisis de referencia, dividido por el número total de constituyentes del árbol de análisis de referencia.
- **Ratio de Cobertura Parentizada** (*Bracketed Recall Rate*): Número de constituyentes del árbol de análisis evaluado cuyos límites coinciden con los de algún constituyente del árbol de análisis de referencia, dividido por el número total de constituyentes del árbol de análisis de referencia.
- **Ratio de Cobertura de Paréntesis Consistentes** (*Consistent Brackets Recall Rate*): Número de constituyentes del árbol de análisis evaluado cuyos límites no se cruzan con los límites de ninguno de los constituyentes del árbol de análisis de referencia, dividido por el número total de constituyentes del árbol de análisis de referencia. Se considera que un constituyente con límites $[i, j]$ se cruza con otro constituyente con límites $[i', j']$ si $i < i' <= j < j'$, es decir, si los límites se solapan pero ningún constituyente está incluido completamente en el otro.

En otras palabras, la cobertura indica la proporción de constituyentes correctos que son planteados como hipótesis, mientras que la precisión evalúa la proporción de constituyentes planteados como hipótesis que son correctos. A su vez, las dos medidas parentizadas son menos estrictas, pues consideran únicamente las palabras de la frase que abarcan los constituyentes, ignorando la etiqueta que tienen asignada. En cuanto a la cobertura de paréntesis consistentes, es aún menos estricta, pues tiene en cuenta sólo la proporción de constituyentes del árbol evaluado que son inconsistentes con el árbol de referencia, es decir, que nunca podrían estar en el mismo árbol de análisis.

Se ha de tener en cuenta que, en nuestro caso, no estamos evaluando la anotación proporcionada por un cierto método de análisis, sino comparando las anotaciones realizadas por dos lingüistas. Por lo tanto, ninguno de los dos análisis que se comparan se pueden considerar de referencia, no existe un *gold standard*. Por ello hemos decidido comparar los análisis en los dos sentidos (análisis del primer lingüista con análisis del segundo, y viceversa), y considerar ambas medidas a la hora de calcular las medias. Teniendo en cuenta las definiciones de las

medidas descritas anteriormente, esto provoca que, de alguna forma, los conceptos de precisión y cobertura dejen de tener sentido, y se unifiquen en una sola medida de comparación, que denominaremos indistintamente precisión etiquetada o parentizada.

La evaluación cuantitativa del acuerdo se ha efectuado durante cinco fases, a lo largo de las cuales se han ido resolviendo los problemas de desacuerdo descritos en la sección 4:

1. En la primera fase se anotaron 100 oraciones y se establecieron los principios básicos de la anotación.
2. En la segunda fase se anotaron otras 220 oraciones. De las discusiones sobre el esquema de anotación surgió una primera versión de la guía de anotación que ya presentaba más detalles sobre el sistema adoptado.
3. En la tercera fase se revisaron y compararon todas las anotaciones anteriores con el objetivo de comprobar que la guía no presentaba ambigüedades y que los anotadores se habían familiarizado ya con el esquema de trabajo.
4. En la cuarta fase se anotaron 670 oraciones.
5. La quinta fase corresponde a los resultados del experimento de evaluación de la anotación descrito en la sección 4.

La figura 1 muestra la evolución de las medidas a lo largo de estas cinco fases. Lógicamente el incremento de todas las métricas es menos acusado a medida que se avanza en las fases, exceptuando significativamente el paso de la cuarta a la última fase. Se observa además que la precisión etiquetada llega a mejorar cerca de un 27 % desde la fase inicial a la final, la precisión parentizada en más de un 20 %, y la consistencia en el parentizado en casi un 15 % (obviamente, cuanto menos estricta es la medida, menor ha de ser la mejora posible).

Una de las principales discrepancias entre anotadores que apareció en las primeras fases de análisis fue la consideración como locuciones o no de estructuras complejas del tipo *desde que, dar lugar a, a lo largo de*, etc., lo cual afectaba a la longitud de las frases[6]. Como nuestras medidas toman en

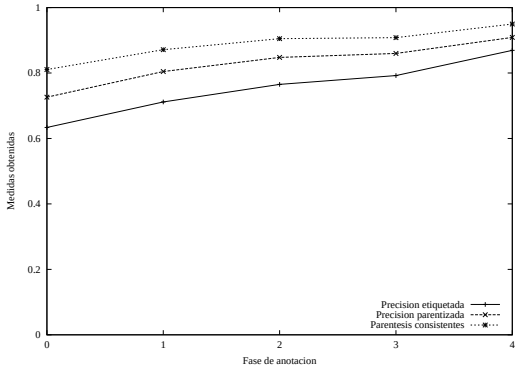

Figura 1: Evolución de las medidas

consideración los puntos de comienzo y finalización de cada constituyente, el hecho de que la longitud de la frase varíe implica un descenso substancial de las medidas (más acusado cuanto más próximas al principio de la frase esté(n) la(s) palabra(s) considerada(s) diferentemente). Por ello, hemos querido evaluar también las medidas de acuerdo obtenidas sólo para aquellas frases cuyas longitudes son iguales. El cuadro 1 muestra todos los resultados obtenidos, incluyendo la evaluación de las medidas para el subconjunto mencionado (sólo a partir de la la 3ª fase, en la cual se ha detectado el desacuerdo en cuestión). Considerando sólo este subconjunto de árboles de análisis, la precisión etiquetada llega a mejorar por encima del 30 %, la precisión parentizada casi un 23 %, y el ratio de paréntesis consistentes en más de un 16 %. Además, todos los valores finales superan holgadamente el 90 % de acuerdo, aproximándonos quizá a ese límite en la precisión de la anotación del que hablábamos en la introducción.

	P. e.	**P. p.**	**P. c.**
Fase 1	0.63359	0.72611	0.81072
Fase 2	0.71166	0.80454	0.87124
Fase 3	0.76537	0.84762	0.90487
Fase 4	0.79222	0.85979	0.90821
Fase 5	0.86927	0.90889	0.94958
Frases de igual longitud			
Fase 3	0.85672	0.91683	0.95485
Fase 4	0.90155	0.93323	0.96034
Fase 5	0.91529	0.94036	0.96985

Cuadro 1: Resultados de la comparación

[6]Si tales expresiones se consideraban locuciones había menos terminales (palabras) en la oración que si se consideraban por separado.

4. Evaluación cualitativa de las discrepancias

4.1. Tipología de desavenencias

Para estudiar y evaluar las discrepancias producidas entre los anotadores, hemos seguido la tipología que presentan Sampson y Babarczy (2003) según los cuales éstas pueden deberse a cuatro motivos:

1. Desacuerdos producidos por la propia ambigüedad o vaguedad del lenguaje. En este tipo se incluyen las ambigüedades de anidamiento de los sintagmas preposicionales y de las relativas. Un ejemplo es el que aparece en el siguiente sintagma, en el que el adjetivo puede complementar al segundo nombre o a los dos: *fibrillas o partículas metálicas*. La guía de anotación proporciona un criterio para etiquetar estas estructuras que es el de anidarlas en el nodo más alto a la izquierda. El desacuerdo surge cuando uno o más anotadores no ven esta ambigüedad, cuando no interpretan esta secuencia como una secuencia ambigua.

2. Desacuerdos producidos por aspectos vagos, por contradicciones o carencias de la guía de anotación. En este caso, la estructura lingüística está clara pero la guía de anotación no indica cómo debe etiquetarse. Un ejemplo de problema que no estaba incluido en la guía es el del tratamiento de expresiones como: *(fig. 2)*; otro era la ubicación de los signos de puntuación que preceden y siguen a expresiones como *es_decir, esto_es*.

3. Desacuerdos producidos por aspectos vagos, contradicciones o simples carencias de la guía de anotación, pero que no se pueden subsanar a priori en la guía. Este aspecto se refiere a fenómenos particulares que aparecen con determinados tipos de textos, con determinadas estructuras que por ser poco frecuentes o muy específicas no pueden aparecer en las guías de anotación más que al final, cuando el proceso de anotación ya ha finalizado. Un ejemplo lo proporcionan las fórmulas matemáticas o algunas convenciones dependientes del dominio al que pertenece el texto.

4. Desacuerdos producidos por un error del anotador en la aplicación de la guía de anotación. Los errores de los anotadores pueden ir desde el olvido de una etiqueta hasta la interpretación errónea de una estructura sintáctica. Un ejemplo del primer caso es el olvido del sufijo **.co** para la etiqueta de un constituyente sintagmático coordinado; un ejemplo del segundo, la interpretación errónea de estructuras similares, como las completivas y las relativas.

Esta tipología de errores tiene también que ver con la segmentación y la etiquetación de los constituyentes, ya que mientras el error debido a la ambigüedad de la lengua está estrechamente relacionado con la parentización, los tres casos restantes están relacionados con la etiquetación.

4.2. Resultados

Para llevar a cabo el estudio cualitativo de las discrepancias entre anotadores se pidió a los cinco anotadores del proyecto que anotaran 33 frases correspondientes a 1038 palabras (31.45 palabras/frase) de un texto de dominio científico y que constituyen el material con que se ha evaluado la quinta fase [7]. Las frases anotadas se compararon de modo manual de dos en dos para proceder a una clasificación de las discrepacias halladas. El cuadro 2 muestra los resultados:

Tipo 1	**Tipo 2**	**Tipo 3**	**Tipo 4**
25.74 %	12.17 %	2.39 %	59.86 %

Cuadro 2: Clasificación de las discrepancias

Como se puede observar, el mayor número de discrepancias son debidas a errores de algún anotador al aplicar los criterios de la guía (tipo 4) o bien son debidas a ambigüedades propias de la lengua (tipo 1). Ambos casos son los más difíciles de controlar. Si bien siempre se puede intentar minimizar estos errores, nunca se podrá llegar a una anotación en la que no exista error humano alguno, y mucho menos habrá un corpus de lengua real sin oraciones ambiguas. El error humano, el de mayor porcentaje, está determinado por gran cantidad de factores externos: estado de ánimo del anotador, cansancio, metodología, etc.

[7] Consideramos que las oraciones restantes constituyeron por así decirlo la fase de entrenamiento.

A diferencia de los dos casos anteriores, las discrepancias del tipo 2, debidas a un error, omisión o contradicción en la guía de anotación, son fácilmente subsanables, ya que la guía de anotación de va enriqueciendo constantemente.

Por último, existen pocos desacuerdos tipo 3, debidos a errores de la guía difíciles de subsanar. Su porcentaje es bajo porque se dan pocos casos en el corpus.

5. Conclusiones

En este artículo se ha presentado una definición de pautas o modelos tanto de procedimiento como de contenido para la anotación sintáctica de corpus. Se han evaluado los resultados del proceso de etiquetación tanto desde un punto de vista cuantitativo como cualitativo. Este último aspecto es especialmente relevante, ya que hasta ahora no ha recibido una atención especial y sin embargo resulta esencial para garantizar la consistencia en la anotación, que es lo que proporciona calidad al corpus anotado.

Bibliografía

Babarczy, A., J. Carroll, y G. Sampson. 2001. Annotator error rates for part-of-speech tagging. En *LINC2001, at 34th SLE*, Leuven.

Bangalore, S., A. Sarkar, C. Doran, y B.A. Hockey. 1998. Grammar & Parser Evaluation in the XTAG Project. En *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, Granada.

Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, y T. Strzalkowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. En *Proceedings of the Speech and Natural Language Workshop*, páginas 306–311, Pacific Grove, CA. DARPA.

Brants, T. 2000. Inter_Annotator Agreement for a German Newspaper Corpus. En *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

Brants, T., W. Skut, y H. Uszkoreit. 2003. Syntactic Annotation of a German Newspaper Corpus. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Carroll, J., T. Briscoe, y A. Sanfilippo. 1998. Parser Evaluation: a Survey and a New proposal. En *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, páginas 447–454, Granada.

Civit, M. 2002. Guía para la anotación sintáctica de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática. Informe Técnico X-Tract-II WP-02/01, 3LB WP 02-01, Universitat de Barcelona. disponible: http://www.lsi.upc.es/ civit/publicacions.html.

Civit, M., I. Castellón, y M.A. Martí. 2001. Creación, etiquetación y desambiguación de un corpus de referencia del español. *Procesamiento del Lenguaje Natural*, (27):21–28, Septiembre. disponible: http://www.lsi.upc.es/ civit/publicacions.html.

Civit, M. y M.A. Martí. 2002. Design Principles for a Spanish Treebank. En *Proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002)*, páginas 61–77, September.

Civit, M., M.A. Martí, B.Navarro, N. Bufí, B. Fernández, y R. Marcos. 2003. Issues in the Syntactic Annotation of Cast3LB. En *Proceedings of the LINC03 Workshop*, Budapest.

Cotton, S. y S. Bird. 2000. An integrated Framework for Treebanks and Multilayer Annotations. En *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

Navarro, B., M. Civit, M.A. Martí, B. Fernández, y R. Marcos. 2003. Syntactic, semantic and pragmatic annotation in Cast3LB. En *Proceedings of the Corpus Linguistics*, Lancaster.

Sampson, G. y A. Babarczy. 2003. Limits to annotation precision. En *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC03),*

Workshop of the 10th EACL Conference, páginas 61–68, Budapest.

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

Véronis, J. 2000. Sense Tagging: don't look for the meaning but for the use. En *Computational Lexicography and Multimedia Dictionaries, COMLEX*, páginas 1–9, Kato Achia, Greece. disponible: http://www.up.univ-mrs.fr/veronis/.

A. Anexo 1: Etiquetas para los constituyentes

oración	S
subord. completiva	S.F.C
subord. adjetiva	S.F.R
subord. adverbial	S.F.A
subord. adv. comparativa	S.F.AComp
subord. adv. condicional	S.F.ACond
subord. adv. concesiva	S.F.AConc
subord. adv. consecutiva	S.F.ACons

Cuadro 3: Etiquetas para las oraciones finitas

subor. completiva	S.NF.C
subor. adjetiva	S.NF.P
subor. absoluta	S.NF.PA
subor. relativa	S.NF.R
subor. adverbial	S.NF.A

Cuadro 4: Etiquetas para las oraciones no finitas

Las etiquetas para las oraciones puede llevar además los sufijos ***** si tienen el verbo elíptico y **.co** si son estructuras coordinadas.

sn	sintagma nominal
gv	grupo verbal
sp	sintagma preposicional
sadv	sintagma adverbial
sa	sintagma adjetivo
conj.subord	conjunción subordinante
coord	conjunción coordinante
infinitiu	verbo en infinitivo
gerundi	verbo en gerundio
interjeccio	interjección
neg	adverbio de negación
morfema.verbal	SE (impers./pasivo)
morf.pron	otros usos

Cuadro 5: Etiquetas para los principales constituyentes oracionales

Las cinco primeras etiquetas del cuadro 5 pueden aparecer también con el sufijo **.co** si son estrucutras coordinadas.