Procesamiento del lenguaje Natural, nº 41 (2008), pp. 233-240 recibido 7-05-2008; aceptado 16-06-2008

Algunos problemas concretos en la anotación de papeles semánticos. Breve estudio comparativo a partir de los datos de AnCorA, SenSem y ADESSE

Gael Vaamonde
Grupo de Investigación Gramática y Léxico
(GIGRALEX)
Departamento de Tradución e Lingüística
Universidade de Vigo
E-36200 Vigo, España
gaelv@uvigo.es

Resumen: La etiquetación de papeles semánticos se ha convertido en un reto importante tanto en el campo de la lingüística de corpus como en el procesamiento del lenguaje natural. Sin embargo, se trata de una tarea compleja en la que debemos afrontar ciertos problemas de anotación y en la que diferentes grupos de trabajo a menudo adoptan soluciones dispares, independientemente del marco teórico que sustente el análisis. En este artículo se describen algunos de estos problemas a la vez que se comparan las distintas soluciones adoptadas por tres proyectos de investigación que han abordado el análisis sintáctico-semántico de un corpus en español.
Palabras clave: Papeles semánticos, anotación de corpus, clasificación de verbos, estructura argumental.

Abstract: The labelling of semantic roles has become an important challenge both in the field of corpus linguistics and in the natural language processing. However, it is a hard task in which we have to deal with certain problems of annotation and in which different groups often take different solutions, regardless of the theoretical framework which supports the analysis. This paper outlines some of these problems and simultaneously compares the different solutions adopted by three research projects that have dealt with the syntactic-semantic analysis of a Spanish corpus.
Keywords: Semantic roles, corpus annotation, verbal classification, argument structure.

1 Introducción

El proceso de anotación de un corpus suele ser modular, es decir, suele obedecer a distintos niveles de análisis lingüístico (morfología, sintaxis, semántica, pragmática). En este sentido, el trabajo en corpus no escapa a algunos de los problemas que ha tenido que tratar la lingüística teórica. El objeto de estudio en cada nivel de análisis se va haciendo cada vez más "escurridizo", menos sistemático, y cada salto de nivel parece implicar una mayor reticencia a la descripción lingüística en términos de unidades discretas, claramente definidas y de fácil aplicación.

A esta complejidad progresiva hay que añadir, de forma paralela, un acuerdo decreciente a efectos de anotación. Frente al relativo consenso que encontramos en el enriquecimiento morfosintáctico de corpus diferentes (siempre que se haga abstracción de teorías sintácticas concretas), la etiquetación semántica puede variar significativamente entre unos anotadores y otros, tanto en lo metodológico como en lo descriptivo, y llevar a soluciones de análisis diferentes para un mismo ejemplo concreto.

Este trabajo pretende sacar a la luz algunos de los problemas con los que se encuentra el anotador al añadir información semántica a un corpus, en concreto al afrontar cuestiones relativas a la etiquetación de papeles semánticos.

ISSN: 1135-5948

Para ello, se han tomado como referencia tres proyectos de investigación que han abordado esta tarea en el ámbito del español: AnCora (*Annotated Corpora*)[1], SenSem (*Sentence Semantics: Creación de una base de datos de Semántica Oracional*)[2] y ADESSE (*Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*)[3].

El trabajo se estructura del modo siguiente. El apartado 2 está dedicado a explicar brevemente los proyectos que serán objeto de estudio. En el apartado 3 se apuntan algunas consideraciones previas que deben ser tenidas en cuenta a la hora de realizar un estudio comparativo entre dichos proyectos. El apartado 4 se centra en tres problemas concretos que ilustran algunas dificultades en la anotación de papeles semánticos. El trabajo finalizará con algunas conclusiones generales en lo que concierne a la etiquetación de papeles en corpus.

2 Los recursos lingüísticos utilizados

2.1 Ancora

El proyecto AnCora, llevado a cabo por el Centre de Llenguatge i Computació (CLiC) de la Universidad de Barcelona, presenta dos corpus de 500.000 palabras cada uno, uno para el catalán (AnCora-CAT) y otro para el español (AnCora-ESP), aunque en este trabajo sólo se tendrán en cuenta los datos de AnCora-ESP. Dicho corpus está compuesto por 400.000 palabras extraídas de distintas fuentes periodísticas y 100.000 palabras provenientes del corpus 3LB-ESP (Civit y Martí, 2004).

La anotación semántica de AnCora parte de una clasificación verbal basada en la conocida tipología de Vendler (1967), posteriormente desarrollada en Dowty (1979), que diferencia cuatro tipos de eventos en función de la Aktionsart: estados, actividades, logros y realizaciones. Además, AnCora adopta la descomposición léxica como método de análisis (Levin y Rappaport, 1995; Rappaport y Levin, 1998), de tal forma que cada tipo de evento es asociado a una Estructura Léxico-Semántica, esto es, una combinación de variables, constantes y predicados primitivos que representan la estructura lógica del evento. Estas cuatro clases generales son a su vez divididas en diferentes subclases en función de la estructura argumental, los papeles semánticos y las alternancia de diátesis, dando lugar a un total de 13 clases semánticas.

La asignación de papeles semánticos a cada argumento del verbo dependerá de la clase semántica asociada a ese verbo (sentido verbal), más concretamente de la estructura léxico-semántica y las alternancias de diátesis en las que aparece (cf. Martí et al., 2007:27 y ss.)

2.2 SenSem

El proyecto SenSem, desarrollado por el Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL) de Cataluña, ofrece información sintáctico-semántica de los que considera los 250 verbos más frecuentes del español. Partiendo de un corpus de aproximadamente 13 millones de palabras, creado íntegramente a partir de las versiones online de "El Periódico de Catalunya", en SenSem se ha optado por seleccionar 25.000 oraciones, 100 por cada verbo, que posteriormente han sido anotadas con información sintáctica y semántica.

El proceso de anotación en SenSem responde básicamente a tres niveles: la unidad léxica, los constituyentes y la oración en sí. Para cada participante se ha señalado su estatus argumental (argumentos frente a adjuntos) y se ha añadido información sintáctica relevante (categoría y función). Además, cada argumento es asociado a un rol semántico determinado.

A nivel oracional, para cada sentido verbal se ha incluido información acerca del tipo de evento designado (evento, proceso o estado) y cada esquema sintáctico se asocia con una etiqueta que resume su significado construccional (anticausativa, antiagentiva, reflexiva, habitual, …), algo que, como se apunta en Castellón et al. (2006), distingue a SenSem de otros proyectos similares.

2.3 ADESSE

ADESSE (Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español) es un proyecto que se está desarrollando en la Universidad de Vigo y que, a partir de la anotación sintáctico-semántica de un corpus del español, pretende ofrecer una base de datos para el estudio empírico de la interacción entre verbos y construcciones.

Toda la información sintáctica de ADESSE es una herencia directa de la Base de Datos

[1] http://clic.ub.edu/ancora/.
HUM2006-27378-E. TIN2006-15265-C06-06
[2] http://grial.uab.es/fproj.php?id=1.
BFF2003-06456
[3] http://webs.uvigo.es/adesse.
HUM2005-01573

Sintácticos del Español Actual (BDS)[4], que contiene el análisis sintáctico e información sobre los elementos valenciales de las casi 160.000 cláusulas que conforman la parte contemporánea del corpus ARTHUS[5]. Este corpus de aproximadamente 1,5 millones de palabras está compuesto por una variada naturaleza de textos (narrativos, teatrales, ensayísticos, periodísticos y orales) procedentes de España e Hispanoamérica.

El proyecto ADESSE basa su razón de ser en el enriquecimiento semántico de los datos aportados por la BDS y este enriquecimiento se orienta fundamentalmente hacia tres objetivos claros: diferenciación de acepciones, clasificación semántica y etiquetación de papeles.

En ADESSE, cada sentido verbal es asociado a una clase semántica determinada (o a varias). Para cada clase semántica se ha previsto una serie de papeles prototípicos del dominio cognitivo descrito. A su vez, cada sentido verbal incuye un conjunto de papeles semánticos para el total de los participantes posibles con ese verbo (potencial valencial). En general, el verbo hereda por defecto los papeles de la(s) clase(s) en que se integra, y se añaden aquellos que se consideran necesarios para dar cuenta de todas las posibilidad construccionales con ese verbo (cf. García-Miguel y Albertuz, 2005)

2.4 Algunas consideraciones previas

Antes de realizar cualquier tipo de comparación entre los proyectos citados, conviene apuntar algunos de los aspectos que los individualizan y que deben ser tenidos en cuenta como paso previo al estudio comparativo que se pretende.

Uno de los problemas más comunes, no sólo en la anotación de corpus sino de manera general en el estudio de la interfaz sintáctico-semántica, es el de la delimitación entre argumentos y adjuntos.

En la tarea de etiquetar los participantes verbales esta delimitación juega un papel relevante, desde el momento en que la anotación de roles semánticos suele estar asociada de manera exclusiva a aquellos elementos que se consideran exigidos por el predicado. Como se aprecia en los ejemplos siguientes, tan sólo AnCora, (1), incluye los adjuntos entre los participantes que llevan etiqueta semántica. En SenSem, (2), los elementos que no son considerados argumentales prescinden de descripción semántica, mientras que ADESSE, (3), asume el trabajo anterior de la BDS y persigue únicamente la anotación de los elementos que fueron tratados como valenciales en dicha base de datos.

(1) […] asistirá a la XII Cumbre de Jefes de Estados Andinos que (**Arg1-PAC**) se celebrará en Lima (**ArgM-LOC**) el 9 y 10 de junio (**ArgM-TMP**)

(2) En Juriba, ciudad del interior marroquí, cada verano se celebra el mercado de los italianos (**Tema**)

(3) […] el tema de estos cursos que (**A2 Actividad**) se celebrarán la semana próxima en el Área de Cultura de Caixa Galicia

Haremos notar, también, que ninguno de los tres proyectos mencionados adopta como único recurso de anotación semántica el inventario de papeles. En los tres casos se aprecia una clasificación semántica de los verbos, bien sea de tipo aspectual (AnCora y SenSem) como nocional (ADESSE). Además, AnCora recurre a la estructura léxico-semántica como método previo a la delimitación y asignación de papeles, mientras que en ADESSE la clase semántica a la que corresponde casa sentido verbal determina en gran medida el conjunto de etiquetas utilizadas para describir su potencial valencial.

Las dos tablas siguientes resumen las características principales de cada proyecto, tanto en lo que se refiere a datos del corpus como en lo tocante al proceso de anotación semántica:

	Corpus		
	palabras	*cláusulas*	*lemas*
AnCora	500.000[6]	6.009	1.895
SenSem	700.000	25.000	250
ADESSE	1.450.000	160.000	3.436

Tabla 1: Relación del número de palabras, cláusulas y lemas verbales en cada corpus

	Anotación de papeles		
	método	*cobertura*	*etiquetas*
AnCora	semiautomático	parcial	20
SenSem	manual	total	24
ADESSE	manual	total	143[7]

Tabla 2: Relación del tipo de método, grado de cobertura y número de etiquetas

[4] http://www.bds.usc.es/
[5] http://www.bds.usc.es/corpus.html

[6] En el momento de redactar estas páginas, la anotación semántica de Ancora-ESP todavía no se ha finalizado (188.513 palabras de un total de 500.000).
[7] Esta lista está actulmente en proceso de revisión

Por último, deben tenerse en cuenta también los objetivos fundamentales de cada proyecto. SenSem y ADESSE son recursos lingüísticos primordialmente descriptivos que proporcionan un sistema de consulta de los datos analizados en cada corpus[8]; AnCora, en cambio, tiene una clara finalidad computacional como fuente de aplicaciones y herramientas relacionadas con el procesamiento del lenguaje natural.

Estos aspectos, que condicionan en muchos casos las soluciones de análisis adoptadas, no eximen, sin embargo, de un estudio como el presente, en el que se busca contrastar algunos problemas concretos en la etiquetación de papeles semánticos en tres proyectos de investigación que comparten el uso dichos papeles como herramienta descriptiva para anotar corpus del español.

3 Algumos problemas de anotación

3.1 La anotación de dativos y CINDs

Son numerosos los trabajos que han mostrado interés por el CIND en español. Para el presente estudio, tomaremos como referencia a Gutiérrez Ordoñez (1999), donde se establece una distinción entre CINDs argumentales (CIND1), ejemplificados en (4a-b) y que aparecen prototípicamente con verbos de transferencia, CINDs no argumentales o incorporados (CIND2), ejemplificados en (4c-d), y que suelen aparecen con verbos de creación, destrucción o preparación, y dativos superfluos, ejemplificados en (4e-f) y diferenciados de los CINDs fundamentalmente en su presentación exclusivamente pronominal y en la posibilidad de coaparición con cualquier otra función sintáctica (cf. Gutiérrez Ordoñez, 1999:1909 y ss.).

(4) a. Le envió una postal a su hermano
b. No nos corresponden esos lujos
c. Te arreglé las tijeras
d. Le arañó la cara
e. Nos tememos lo peor
f. No te me acalores

Partiendo de dicha tipología, veamos cómo trata cada proyecto las funciones citadas.

En Ancora, tanto los CIND1 como los CIND2 llevan de manera general la etiqueta de Beneficiario (BEN), como se aprecia en los ejemplos (7a-b) y (7c-d), respectivamente. En cuanto a los dativos que cita Gutiérrez Ordóñez, parecen quedar fuera del proceso de anotación semántica para este proyecto, a la luz de los ejemplos recogidos en (9e-f):

(5) a. […] dar a los fabricantes de ordenadores (**Arg2-BEN**) mayor flexibilidad
b. […] uno de los dos puestos que le corresponden a España **(Arg2-BEN)**
c. […] que abrirá a este país (**Arg2-BEN**) los mercados chinos
d. […] para arreglarle la jaima a la Caballé (**Arg2-BEN**)
e. Un solo visón se (ø) comió 87 huevos
f. Se (ø) llevó una bolsa de 200.000 dólares

Quizás lo que más llama la atención es que en AnCora son tratados como argumentales tanto los Beneficiarios que funcionan como CIND1 como los que funcionan como CIND2, por lo que, a efectos de anotación, no parece haber ningún aspecto diferenciador, ni sintáctico ni semántico, entre uno y otro caso.

En lo que concierne a SenSem, los CIND1 son etiquetados como Destino (Dest), como se ve en (6a-b), mientras que los CIND2 ofrecen una solución dispar. Generalmente, no son etiquetados semánticamente (6d), en consonancia con la idea de reservar esta información tan sólo para los elementos argumentales del verbo. En cualquier caso, la determinación de la argumentalidad es una cuestión compleja y sujeta a diferentes interpretaciones, por lo que encontramos casos clasificados por Gutiérrez Ordoñez como CIND2 que vienen acompañados por un papel semántico en SenSem (6c). Los dativos superfluos, obviamente, carecen de etiquetación semántica en este proyecto (6e-f)

(6) a. Daremos una respuesta positiva a las personas que trabajan en las casas (**Dest**)
b. […] de los que 4.133 [trabajadores] corresponden a España (**Dest**)
c. Si nos (**Dest**) crean una nueva barrera, que nos quiten otra
d. […] pidiendo que le (ø) arreglen la Casa dels Canonges
e. […] como las ovejas no son suyas sale corriendo, y el lobo se (ø) las come
f. Se (ø) llevó la mano derecha a la boca

Conviene señalar que en SenSem algunas cláusulas han sido anotadas a nivel oracional con la etiqueta "Dativo de interés", donde se incluyen tanto dativos posesivos (7a), que en Gutiérrez Ordoñez (1999) son tratados dentro del grupo CIND2, como dativos claramente

[8] Para una comparación entre ambos proyectos, véase Cuadros Muñoz, 2005:126 y ss.

superfluos (7b). Sin embargo, casos como los de (7c-d), que también parecen claros dativos posesivos, no están tratados como "Dativo de interés" y tan sólo uno de ellos aparece acompañado de etiqueta semántica, por lo que no parece haber una solución sistemática para la anotación de este tipo de dativos en SenSem:

(7) a. […] me (ø) he reducido el estómago
b. Se nos (ø) va Julia de TV-3
c. Se le (**Dest**) ve demasiado el truco
d. […] como ciego, en Telecupón, llega a tocarle el culo a Belinda Washington (ø)

Por último, en ADESSE tanto los CIND1 como los CIND2 llevan etiqueta semántica. En el primer caso esta etiqueta vendrá determinada por la clase semántica asociada al verbo en cuestión (Poseedor-final con verbos de transferencia, Entidad2 con verbos de atribución, …) y en el segundo caso será habitualmente un Beneficiario o un Poseedor, etiquetas generales no asociadas a ninguna clase concreta (AG). Por su parte, la mayor parte de los denominados dativos superfluos carecen de etiquetación y se interpretan como marca de voz media.

(8) a. […] ya se venció el plazo que le dimos a la gerencia (**A1 POS-FIN**)
b. […] las mayores subidas han correspondido a Madrid (**A2 ENT2**)
c. Ya de paso que nos (**AG POS-A1**) arreglé la cocina
d. A la mañana siguiente no quiso abrirme (**AG BEN**) la puerta
e. ¿Y sabe mi señora qué haría después? ¡Me (ø) comería los cocodrilos!
f. Se (ø) llevaron a mi padre, y mi madre lo veía en sueños

La tabla siguiente resume las diferentes soluciones de anotación en cada proyecto para los CINDs diferenciados:

	CIND1		**CIND2**		**Dat**
			dat. pos.	otros	
AnCora	BEN	BEN	BEN	BEN	ø
SenSem	Dest	Dest	ø/dat interés	Dest/ø	ø
ADESSE	Papel de la clase		POS	BEN	ø

Tabla 3: Papeles semánticos frecuentemente asociados a CINDs y dativos en AnCora, SenSem y ADESSE

La diferencia principal entre unos proyectos y otros estriba en cómo tratar los CINDs incorporados (CIND2). En AnCora se ha optado por unificar todos los participantes que son codificados mediante CIND y que presentan cierto grado de afectación bajo la etiqueta general de Beneficiario.

SenSem parece dar un paso más allá en el tratamiento de estos constituyentes y, aunque en términos generales sigue la misma línea de análisis que AnCora, en esta caso tomando como papel unificador el de Destino, reconoce una solución específica para los conocidos dativos posesivos. Sin embargo, esta solución se presenta a nivel oracional, no mediante la adopción de un papel diferente, y de forma asistemática, como prueban los ejemplos de (7).

Por último, ADESSE entiende que los CIND2, al no ser claramente argumentales, no heredan un papel de la clase correspondiente, como sí lo hacen los CIND1, y por eso deben ser etiquetados con papeles generales. Además, entre esos últimos se establece una distinción, al menos en el nivel más específico del análisis, entre Beneficiarios y Poseedores. Esto otorga mayor granularidad al análisis que presenta ADESSE, aunque como contrapartida pueden darse aparentes incoherencias como las de (9), fruto de la etiquetación de casos ambiguos que suponen un problema adicional respecto de SenSem y AnCora:

(9) a. Le mira las manos (**POS**)
b. Le inmoviliza los brazos (**BEN**)
c. Se les ha detectado un virus (**POS**)
d. Se le designó un abogado (**BEN**)
e. Le golpeaba en el estómago (**POS**)
f. Le soplaba en la boca (**BEN**)

3.2 Las alternancias con participantes adicionales

Otro de los problemas con lo que debe lidiar todo proceso de anotación semántica en corpus tiene que ver con las conocidas alternancias construccionales que puede presentar un mismo núcleo verbal. De entre ellas, nos centraremos en aquellas que son consecuencia de añadir un participante adicional en el evento descrito, como se ilustra en los esquemas de (10) y (11):

(10) a. Alguien imita algo
b. Alguien imita a alguien
c. Alguien le imita algo a alguien
d. Alguien imita a alguien en algo

(11) a. Alguien sorprende a alguien
b. Algo sorprende a alguien
c. Alguien sorprende a alguien con algo
d. Algo sorprende a alguien de alguien

En general, para casos como estos se hace necesario el uso de tres etiquetas semánticas, uno por cada constituyente de las cláusulas triactanciales correspondientes (10c-d) y (11c-d). El problema radica en cómo se recoge la relación entre los diferentes esquemas que conforman la alternancia a través de los papeles semánticos seleccionados y en cómo aplicar estos papeles en función del carácter animado o inanimado del participante en cuestión.

Las dos tablas siguientes ilustran, a partir de la observación de diferentes ejemplos, las distintas soluciones de anotación adoptadas:

	(SUJ) Alguien	**algo**	**a alguien**	**en algo**
An	Agt	Pat		
	Agt		Pat	
	Agt	Pat	Ben	
	Agt		Pat	Adv
Sen	No registrado			
AD	Act	Obj		
	Act		Obj	
	Act	Obj	Ref	
	Act		Obj	Ámb

Tabla 4: Soluciones de anotación para *imitar* y similares[9]

	(SUJ)		**a alguien**	**de alguien**	**con algo**
	alguien	**algo**			
An	Cau		Pat		
		Cau	Pat		
	Cau		Pat		Adv
		No registrado			
Sen	Agt		Exp		
		Cau	Exp		
	Agt		Exp		Cau
		Cau	Exp	ø	
AD	Est		Exp		
		Est	Exp		
	Est		Exp		Med
		Est	Exp	Ref	

Tabla 4: Soluciones de anotación para *sorprender* y similares

[9] Act (Actor), Adv (Adverbial), Agt (Agente), Ámb (Ámbito), Ben (Beneficiario), Cau (Causa), Est (Estímulo), Exp (Experimentador), Med (Medio), Obj (Objeto), Ref (Referencia)

En líneas generales, nos encontramos con dos vías de etiquetación para estos casos. La primera de ellas consiste en utilizar etiquetas diferentes para un mismo esquema sintáctico en función del carácter animado o inanimado de sus constituyentes. Es lo que sucede en SenSem para verbos como *sorprender* y similares, en los que se establece una diferencia a efectos de anotación entre Agentes (entidades animadas) y Causas (entidades inanimadas). Por tanto, a un mismo esquema sintáctico (SUJ-CDIR) le corresponden dos esquemas semánticos diferentes (Agente/Causa -Exp) y en las construcciones con tres participantes, el Experimentador se mantiene inalterable y los papeles Agente y Causa cubren el resto de posibilidades, si es que se consideran argumentales.

La otra vía pasa por obviar la animación de los participantes en estos casos y usar el mismo papel para el objeto (*imitar* y similares) o el sujeto (*sorprender* y similares), sea o no animado, en los esquemas transitivos. Mediante esta vía, en las construcciones con tres constituyentes se hace necesario recurrir a papeles específicos para anotar el tercer participante en cuestión. Como se aprecia en las dos tablas anteriores, esta opción es la adoptada en AnCora y en ADESSE.

Aunque las dos soluciones son válidas, debe tenerse en cuenta que implican una diferencia importante. Tomando como ejemplo el caso de *sorprender*, en el primer caso la relación semántica del constituyente en función de SUJ varía como consecuencia de la animicidad, pero se entiende que la relación semántica que mantienen con el verbo tanto el SUJ inanimado como el CPREP(con) es la misma. En el segundo caso, por el contrario, el carácter animado o inanimado del participante no supone un cambio de función semántica y, sin embargo, al CPREP(con) sí se le asocia una función semántica concreta, distinta de la del SUJ inanimado.

Dicho de otro modo, el análisis de SenSem refleja una asociación directa entre referentes y relaciones semánticas, independientemente de la función sintáctica que los codifique, mientras que el análisis adoptado por ADESSE y AnCora entiende que la identidad de referentes no implica identidad de papeles semánticos, sino que es la alternancia construccional la que conlleva un cambio de relaciones semánticas con el verbo.

3.3 Casos fronterizos y de difícil asignación

Ya apuntamos con anterioridad que los límites entre papeles como Beneficiario y Poseedor no son fáciles de establecer. Pero este era un problema específico de ADESSE, que opta por esta distincion de manera recurrente. Hay, sin embargo, otros casos fronterizos que suponen un problema común a los tres corpus anotados.

Tal es el caso de ciertas etiquetas semánticas utilizadas para anotar participantes que no están directamente implicados en el evento descrito, sino que suelen designar significados generales y hasta cierto punto opcionales. Me refiero a papeles semánticos como Manera, Instrumento, Finalidad o Estado final.

Una prueba evidente del carácter fronterizo y ambiguo que representan estas etiquetas es el hecho de que muchas veces éstas no presentan el mismo valor extensional en cada proyecto.

Así, la misma construcción con un verbo como *cerrar* ofrece soluciones diferentes en AnCora (12a), SenSem (12b) y ADESSE (12c). En (13) se ilustra un caso similar con el verbo *conducir*:

(12) a. El IBEX cierra otro mal mes con una caída acumulada del 6,8 % (**Manera**)
b. Con un 25 % de cuota de pantalla (**Instrumento**), Telecinco cierra su mejor mes
c. Este año espera cerrar el ejercicio con una facturación de 15 millones (**ø**)

(13) a. […] transformaciones que conduzcan a disminuir las desigualdades (**Estado final**)
b. […] diseñó una planificación que conducía a lograr un estado de forma óptimo (**Finalidad**)
c. […] distraer recursos en cuestiones que no conducen de forma inminente a desterrar su endemia (**Dirección**)

Incluso es posible que dentro de un mismo corpus ejemplos similares reciben una anotación diferentes como consecuencia de una aplicación vacilante de algunos de los papeles mecionados. Es lo que ocurre, por ejemplo, con el papel Manera en AnCora, que puede presentar vacilaciones con el Instrumento o el Estado final, entre otros:

(14) a. […] limpiándose los dientes con un trozo de abeto (**Instrumento**)
b. Un hombre que era capaz de decapitar una rata con los dientes (**Manera**)
c. […] anunció que doblaría a cinco dólares (**Estado final**) el salario mínimo
d. […] otro planteamiento que dividirá a la empresa en tres compañías (Manera)

Para solventar este problema, en AnCora muchas veces se opta por aplicar una misma etiqueta. Es lo que sucede con el papel Beneficiario, usado de manera general para todos los casos de CINDs ya mencionados. La desventaja obvia que esto implica es una relativa carencia de poder descriptivo, puesto que el análisis se detiene en un nivel a veces demasiado superficial. En este sentido, me parecen reveladores ejemplos como los de (15), donde todos los constituyentes subrayados han sido tratados como Manera en dicho proyecto:

(15) a. En la reanudación, el marroquí Yunes el Aynaui se impuso finalmente a Ferrero por 6-7, 3-7, 6-4…
b. […] forzó la tercera y última [manga] al imponerse en el segundo set
c. No pudo desarrollar el tenis con el que se impuso al croata Goran Ivanisevic

Quizás el caso opuesto en este sentido lo encontramos en ADESSE, que ofrece una alta granularidad en su anotación. El precio que debe pagar por ello es el de tener que lidiar con un mayor número de casos fronterizos. Así sucede con papelse semánticos como Beneficiario y Poseedor, Finalidad y Rol, Manera y Estado final, Asunto y Ámbito o Causa y Referencia, entre otros

4 Conclusiones

Desde los conocidos trabajos de Gruber (1965) y Fillmore (1968), no son pocos los autores que han mostrado su escepticismo sobre la noción misma de papel semántico, al menos en el sentido más tradicional y reduccionista del término. Sin embargo, en un corpus lingüístico encontramos una variedad de ejemplos enorme, que responden a muestras de uso de la lengua y que necesitan ser descritos de forma práctica y sencilla. De ahí que el inventario de papeles semánticos resulta un método ampliamente aceptado en lingüística de corpus.

Pero se debe asumir igualmente que el significado es muchas veces reacio a una descripción en términos discretos y que, como consecuencia de ello, el proceso de etiquetación no está exento de problemas. En este trabajo se han querido mostrar algunos de esos problemas a partir de la

comparación de tres proyectos de investigación que etiquetan corpus del español.

En lo que se refiere al tratamiento de los CINDs, la complejidad intrínseca de esta función obliga a elegir entre dos vías de análisis. AnCora y SemSem aplican una etiqueta general para la mayor parte de los casos, ya sea (BEN), ya sea (Dest), aunque SenSem adicionalmente informa de ciertos casos de dativo de interés a nivel oracional. ADESSE opta por un análisis más específico y, al lado de las etiquetas propias de cada clase semántica, propone distinguir entre Beneficiarios y Poseedores, aunque ello lleve a encarar casos ambiguos de difícil asignación.

Respecto a las alternancias de diátesis comentadas, hemos visto que surgen también dos vías de anotación diferentes. En la primera, adoptada en SenSem, un mismo esquema puede ser anotada con papeles diferentes en función del carácter animado o inanimado de los referentes, lo que refleja una asociación directa entre referentes y relaciones semánticas. En la segunda vía, adoptada en AnCora y ADESSE, un mismo esquema del verbo recibe una única anotación, con lo que la animación de los participantes se vuelve secundaria. Es la alternancia construccional la que conlleva un cambio de relaciones semánticas con el verbo, añadiéndose una etiqueta específica para los esquemas triactanciales de la alternancia en cuestión.

Por último, el problema de los casos fronterizos responde una vez más a dos estrategias diferentes. La adopción de etiquetas generales reduce el número de casos ambiguos, pero puede llevar a una superficialidad en el análisis. Por el contrario, un análisis más exhaustivo de los datos, multiplica el número de ambiguëdades, por lo que se corre el riego de perder sistematicidad en la anotación

El reto principal en la etiquetación semántica de corpus estriba, de hecho, en conseguir ese equilibrio entre ambas condiciones: facilidad de aplicación, que se traduce en una consistencia interna de los datos, y calidad de la anotación, que se traduce en una mayor granularidad en el análisis. En la relación inversamente proporcional de ambos factores, SenSem y sobre todo AnCora (por su finalidad computacional), parecen decantarse por una mayor sistematicidad y coherencia internas, mientras que ADESSE, también por las características y objetivos del proyecto, persigue un mayor poder descriptivo en el tratamiento de los datos.

Bibliografía

Castellón, I., A. Fernández, G. Vázquez, L. Alonso y J. Capilla. 2006. The SenSem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. *Fifth International Conference on Language Resources and Evaluation*, páginas 355-359

Cuadros Muñoz, R. 2005. La complementación verbal. Viejos y nuevos enfoques. *Language Design*, 7:105-136.

Civit, M. y M. A. Martí. 2004. Building Cast3LB: a Spanish Treebank. En *Research on Language & Computation* 2(4):549-574

Dowty, D. R. 1979. *Word Meaning and Montague Grammar*. Reidel, Dordrecht

Fillmore, Ch. 1968. The Case for Case. En E. Bach y R. T. Harms (eds.). *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, páginas 1-88.

García-Miguel, J. M. y F. Albertuz. 2005. Verb, semantic classes and semantic roles in the ADESSE project. En K. Erk, A. Melinger y S. Walde (eds.). *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbrüken, páginas 50-55.

Gruber, J. S. 1965. *Studies in Lexical Relation*, Tesis doctoral. The MIT Press, Cambridge, Massachusetts.

Gutiérrez Ordoñez, S. 1999. Los dativos. En I. Bosque y V. Demonte. *Gramática descriptiva de la lengua española*. RAE/Espasa Calpe, Madrid, (vol. 2), páginas 1855-1930

Levin, B. y M. Rappaport-Hovav. 1995. *Unaccusativity. At the Syntax-Lexical Semantics Interface*. The MIT Press, Cambridge, Massachusetts.

Martí, M. A., M. Taulé, M. Bertrán y L. Márquez. 2007. *AnCora: Multilingual and Multilevel Annotated Corpora*. Draft version. [http://clic.ub.edu/ancora/ancora-corpus.pdf]

Rappaport-Hovav, M. y B. Levin. 1998. Building Verb Meanings. En M. Butt y W. Geuder (eds.). *The Projection of Arguments: Lexical and Compositional Factors*. CSLI Publications, Standford, páginas 97-134.

Vendler, Z. 1967. *Linguistics in Philosophy*. Cornell University Press, New York