ANÁLISIS CUANTITATIVO DEL CORPUS SENSEM

ANA FERNÁNDEZ MONTRAVETA
Universitat Autònoma de Barcelona

GLORIA VÁZQUEZ
Universitat de Lleida

RESUMEN

En este artículo presentamos un análisis cuantitativo de algunos datos extraídos del corpus SenSem del español. Nos hemos centrado en la descripción de la tiplogía de los argumentos de las oraciones anotadas, los patrones de subcategorización y las construcciones. Asimismo, aportamos también algunos datos complementarios sobre la semántica oracional en torno a la aspectualidad, la modalidad y la polaridad de las oraciones.

Palabras clave: corpus, argumentos, subcategorización, construcciones.

ABSTRACT

In this paper we present a quantitative analysis of data extracted from the SenSem corpus for Spanish. Our description of information focuses on argument typology, subcategorization patterns and constructions. Furthermore, we provide some additional data regarding other semantic related considerations such as aspectuality, modality and polarity.

Keywords: corpus, arguments, subcategorization, construccions.

1. EL PROYECTO SENSEM

En el proyecto SenSem[1] se ha construido un banco de datos (BD) de verbos del español, compuesto por un léxico verbal (250 lemas desglosados en aproximadamente 1.000 sentidos)[2] y un corpus anotado asociado al léxico[3].

Los objetivos son, por un lado, definir el comportamiento sintáctico-semántico de los verbos utilizando la información del corpus y, por otro, dar cuenta de la semántica global de las oraciones teniendo en cuenta el tipo de construcciones (pasivas, anticausativas, etc.), la aspectualidad, la modalidad y la polaridad y siendo en este campo pioneros en la lingüística de corpus del español[4].

En el momento de redacción de este artículo, el corpus está formado por un conjunto de 30.231 oraciones, de las cuales 25.044 pertenecen al registro periodístico (82,84%) y 5.187 al literario (17,16%). Aunque no se ha finalizado la anotación de todos los niveles ni la revisión de toda la anotación, estamos en la fase final del proyecto y los datos disponibles nos han permitido extraer ya conclusiones que se presentan en este trabajo.

En el apartado 2 presentamos la caracterización de los argumentos reflejados en las oraciones y en el apartado 3 los datos relativos a la semántica oracional.

2. SUBCATEGORIZACIÓN VERBAL

En este apartado nos centraremos en la descripción de la argumentalidad siguiendo las directrices de Vázquez y Fernández 2008. En 2.1 presentaremos una caracterización global de los argumentos y en 2.2 los datos sobre los esquemas de subcategorización extraídos.

2.1 Argumentos

Para describir los argumentos damos cuenta de sus categorías, funciones sintácticas y roles semánticos. Aportamos datos de 29.196 frases (96,6%): 24.347 del registro periodístico y 4.857 del literario.

En la tabla 1 se presenta la distribución de las categorías básicas[5]. Se observa que la categoría más habitual es el sintagma nominal (SN), seguido del preposicional (SPREP) y del pronominal (SPRON). Cabe decir que 1.618 de los 9.818 SPREP están constituidos por una preposición seguida de una oración. La presencia de sintagmas adjetivales (SADJ) y adverbiales (SADV) en el corpus es muy baja.

Tabla 1. Categorías sintagmáticas

Categoría	Nº frases	Porcentaje
SN	19.525	66,9%
SPREP	9.818	33,6%
SPRON	9.325	31,9%
O	4.923	16,9%
SADJ	658	2,3%
SADV	495	1,7%

En cuanto a las funciones sintácticas (tabla 2), el sujeto y el objeto directo aparecen representados con gran diferencia respecto a todos los demás casos y con datos similares. Respecto al sujeto, destacamos que es más habitual su expresión en el discurso periodístico (71,7%) que en el literario (53,3%). Además, hay 9.192 oraciones cuyo sujeto es elíptico (30,4%)[6], de las cuales 6.810 pertenecen al registro periodístico (el 27,2% del total del periodístico) y 2.379 al literario (el 45,90% del total del literario). La explicación parece encontrarse en el tipo de discurso.

Tabla 2. Funciones sintácticas

Función sintáctica	Nº frases	Porcentaje
Sujeto	20.707	70,9%
Objeto directo	17.111	58,6%
Objeto preposicional	7.824	26,8%
Objeto indirecto	2.116	7,2%
Circunstancial	1.132	3,9%
Predicativo	559	1,9%
Atributo	527	1,8%
Complemento agente	28	0,1%

Es destacable también el número de casos de objetos preposicionales. En cuanto a la función circunstancial, recordamos que estamos considerando sólo los casos de argumentalidad. También puede observarse la prácticamente ausencia de complementos agentes[7], lo que corrobora que las pasivas se suelen usar con el fin de generalizar sobre el sujeto lógico.

Observemos los datos de la tabla 3, donde se presentan las funciones semánticas de los argumentos.

Tabla 3. Roles semánticos

Rol semántico	Nº frases	Porcentaje
Tema	16.692	57,17%
Agente	9.833	33,7%
Tema desplazado	4.826	16,5%
Destino	3.382	11,6%
Iniciador	2.221	7,6%
Tema afectado	1.948	6,7%
Localización	1.694	5,8%
Experimentador	1.246	4,3%
Causa	1.181	4%
Finalidad	1.037	3,6%
Manera	896	3,1%
Cualidad	750	2,6%
Origen	587	2%
Cantidad	483	1,7%
Instrumento	422	1,4%
Tiempo	401	1,4%
Medio	160	0,1%
Perceptor	123	0,4%
Circunstancia	81	0,3%
Ruta	58	0,2%
Sustituto	42	0,1%
Compañía	8	0,02%

El rol tema (no incluimos los objetos desplazados ni los afectados) está poco definido y es usado a veces por defecto. Sí que es relevante la presencia de agentes en segundo lugar. También cabe incidir en la representación destacable del tema desplazado y destino, hecho que puede estar relacionado con el registro periodístico, ya que los verbos de comunicación han sido etiquetados como verbos de desplazamiento (Langacker 1987). Cabe subrayar que el rol iniciador, carente de tradición, tiene un papel destacado, superior a otros muy estudiados, como el tema afectado (paciente) o el experimentador. Lo mismo podemos decir si comparamos el rol cualidad con el de instrumento.

Los últimos roles de la lista tienen realmente una presencia muy poco representativa y ello puede llevarnos a revisar su uso.

2.2 Patrones de subcategorización

En este apartado se presentan los datos relativos a las estructuras de subcateogrización independientemente de su interpretación semántica (v. apartado 3.1)[8]. El total de frases que se usado para extraer la información en este apartado es de 29.450.

En primer lugar, presentamos una comparativa entre patrones impersonales vs. no impersonales y pronominales[9] vs. no pronominales (tabla 4). Remarcamos la escasa presencia de oraciones pronominales y, sobre todo, de impersonales.

Tabla 4. Impersonales y pronominales

	Impersonales	%	No impersonales	%
	318	1,08%	29.132	98,9%
No pronominales	179	0,6%	26.924	91,4%
Pronominales	139	0,5%	2.208	7,5%

En segundo lugar, aportamos datos sobre los patrones a partir de la relevancia del primer objeto (tabla 5) teniendo en cuenta, no su lugar de aparición en la frase, sino su prominencia según la función sintáctica, aplicando la siguiente jerarquía: directo, indirecto,

preposicional, atributo, predicativo, complemento agente y circunstancial.[10] Respecto a las oraciones no pronominales cabe subrayar que más del 50% subcategorizan "primordialmente" un SN, seguidas por aquellas que subcategorizan un SP, aunque con diferencia remarcable. En el caso de las pronominales, obviamente, se invierte la casuística.

Por otro lado, los SADV y los SADJ tienen un papel muy poco destacado en ambos tipos de oraciones.

Tabla 5. Subcategorización según el objeto

	No pronominal				Pronominal			
	No impers.		Impers.		No impers.		Impers.	
SN	15.725	53,4%	131	0,44%	110	0,37%	27	0,09%
SP	6.820	23,1%	29	0,10%	590	2%	99	0,35%
SADV	269	0,9%	2	0,007%	29	0,09%	2	0,007%
SADJ	421	1,4%	1	0,003%	28	0,07%	-	-
Sin objeto	3.822	13%	16	0,05%	1318	4,5%	11	0,04%

Por último, vamos a presentar los datos por lo que respecta a las valencias o número de actantes de las estructuras (tabla 6). Por lo que se refiere a las oraciones no impersonales no pronominales, la valencia más frecuente es la de 2 actantes, seguida a gran distancia por la de 1 y 3 actantes. En cuanto a las no impersonales pronominales el orden es por número de actantes: 1, 2 y 3. En los patrones de oraciones no impersonales pronominales e impersonales serán ahora los de 1 actante los más habituales, seguidos a gran distancia también de los de 0 y 2 actantes.

Tabla 6. Valencias

	No impersonales				Impersonales			
	No pron.		Pron.		No pron.		Pron.	
0 actantes	-				16	0,05%	11	0,04 %
1 actante	3.541	12,03%	1.303	4,4%	124	0,4%	100	0,34 %
2 actantes	19.922	67,6%	717	2,4%	40	0,13%	27	0,09
3 actantes	3.575	12,1%	23	0,08%	0	-	1	0,003 %
4 actantes	49	0,17%	1	0,003%	0	-	0	-

3. SEMÁNTICA DE LA ORACIÓN

En este apartado aportamos datos sobre aspectos relacionados con los significados de las oraciones. En 3.1, comentaremos las construcciones tradicionalmente conocidas como pasivas, anticausativas, reflexivas o recíprocas, entre otras. En 3.2 aportaremos datos sobre otros aspectos cruciales para configurar el significado oracional, como la aspectualidad o la modalidad.

3.1 Construcciones

Por coherencia en la anotación, se optó por distinguir básicamente entre dos grandes tipos de construcciones: aquellas en que el sujeto lógico coincidía con el sintáctico y aquellas en las que no quedaba expresado, bien por generalización o desconocimiento. Elegir las etiquetas para denominar estos fenómenos ha sido complicado. Por el momento, se optado por denominar estos fenómenos como casos de topicalización y destopicalización del sujeto lógico, respectivamente[11]. Así, por ejemplo, si se requieren datos sobre pasividad al consultar el corpus, se podrá determinar el concepto al que el usuario se refiere decidiendo si sólo quiere abarcar los casos de generalización de agente, o también de otros, como experimentadores.

Seguidamente, vamos a aportar los datos correspondientes al estudio realizado a partir de 20.955 oraciones (tabla 7). Como se observa, el número de las oraciones en que el sujeto lógico es el tópico

es mayor que en las que este elemento está destopicalizado. A diferencia de lo que se podría esperar según los datos de la tabla 2, el agente es el participante más topicalizado y más destopicalizado. De ello se desprende que el tema, el rol más usado en la anotación, ocupa un lugar relevante en la posición de objeto lógico.

Otro rol que tiene un papel destacado en la posición de sujeto lógico es la causa. Cabe decir, sin embargo, que el rol iniciador ocupa el segundo lugar en las oraciones de topicalización, por delante de la causa y, en cambio, es poco relevante en las oraciones de destopicalización, al contrario de la causa.

Respecto al experimentador, se presenta proporcionalmente en ambos tipos de construcciones y por debajo de los roles ya mencionados. Por último, el rol destino aparece en oraciones en que ocupa la posición de sujeto sintáctico y no se dan casos de destopicalización.

Tabla 7. (Des)topicalización del sujeto lógico

	Topicalización del sujeto lógico	%	Destopicalización del sujeto lógico	%
	17.927	85,5%	3.028	14,5%
Agente	9.701	54,1%	2.052	68%
Tema	3.935	22%	20	0,7%
Iniciador	2.002	11,2%	38	1,3%
Causa	1.006	5,6%	753	25%
Experimentador	966	5,4%	150	5%
Destino	195	1%	0	0%
Perceptor	122	0,7%	15	0%
Otros	0	0%	0	0%

Otro tipo de construcciones que hemos identificado dentro del corpus son las reflexivas y recíprocas. Cabe decir que su presencia es muy poco significativa: 143 y 106 respectivamente.

3.2 Información aspectual

La descripción de la información aspectual contenida en las oraciones se ha llevado a cabo teniendo en cuenta las aproximaciones de Smith 1997 y Xiao y McEnery 2004. Así, se ha considerado, además del tipo eventual léxico, el significado sobre aspectualidad que pueden aportar los diferentes elementos que comparecen en una determinada oración, ya sean auxiliares, desinencias verbales o la presencia y tipología de determinados argumentos o adjuntos.

Si observamos la tabla 8, podemos concluir que el tipo aspectual más habitual es el de los eventos, seguido de los procesos y, finalmente, con un porcentaje considerablemente más bajo, los estados, de los cuales alrededor de un 90% son de tipo permanente.

Tabla 8 – Tipos aspectuales

	Total corpus
Eventos	14.273 – 53,6%
Procesos	8.740 – 32,8%
Estados	3.630 – 13,6%

Por lo que respecta a la perfectividad, la distribución entre oraciones imperfectivas y perfectivas es prácticamente equivalente. Por otro lado, algunas oraciones con verbos usados con tiempo imperfectivo pueden llevar asociada una lectura habitual, que se da, por ejemplo, cuando se usan adverbios como siempre, a menudo, etc., o auxiliares como soler. No obstante, se han cuantificado muy pocos casos de oraciones con interpretación habitual (alrededor de un 2%).

3.3 Modalidad y polaridad

En el corpus SenSem, hasta el momento, se ha anotado la modalidad y polaridad a nivel muy básico, indicando los valores *asertiva* vs. *no asertiva* y *positiva* vs. *negativa* respectivamente. Respecto a la modalidad (tabla 9) lo más significativo es que se establece una clara diferencia entre el número de oraciones asertivas según el tipo de subcorpus, lo cual se explica por la diferente función comunicativa en cada tipo de registro. En cuanto a la polaridad (tabla 10), cabe remarcar la poca presencia de oraciones negativas.

Tabla 9. Modalidad

	Asertividad	Porcentaje dentro del subcorpus	No asertividad	Porcentaje dentro del subcorpus
Corpus periodístico	17.505	70,7%	7.228	29,3%
Corpus literario	1.438	44,1%	1.825	55,9%

Tabla 10. Polaridad

Polaridad positiva	Porcentaje	Polaridad negativa	Porcentaje
18.350	93,6%	1.264	6,4%

4. CONCLUSIONES

Hemos presentado datos cuantitativos que describen las oraciones del corpus SenSem desde el punto de vista sintáctico-semántico. Aunque dicho corpus contiene sobre todo oraciones pertenecientes al registro periodístico, hemos establecido algunas comparaciones con los datos obtenidos de las oraciones extraídas de obras literarias, cuando los datos han sido reveladores, como en el caso de la modalidad.

Nos gustaría señalar que a partir de la información presentada se pone de manifiesto que en ocasiones los fenómenos más tratados en la bibliografía no son necesariamente los más relevantes cuantitavivamente en los corpus. Así, por ejemplo, la presencia de construcciones como medias, reflexivas o recíprocas es muy baja. Por el contrario, se revelan como aspectos que cabría analizar con más detalle otros que a penas han sido considerados, como ocurre con el protagonismo del rol que hemos denominado iniciador.

El estudio presentado se ha centrado en la argumentalidad. Cabe decir que en el total de frases analizadas, se han identificado que alrededor del 50% contienen adjuntos. Consideramos que como trabajo futuro debería estudiarse para cada verbo qué tipo de adjuntos

suelen aparecer en las oraciones con mayor frecuencia para replantear, si fuera necesario, los límites de la argumentalidad en cada caso.

NOTAS

[1] Ministerio de Educación y Ciencia HUM2007-65267.
[2] http://grial.uab.es/tools/lexico
[3] http://grial.uab.es/tools/buscador
[4] Para una descripción teórica de los criterios lingüísticos aplicados en la anotación remitimos al lector a Fernández y Vázquez en prensa.
[5] No damos cuenta de los casos de frases en que hay más de una categoría ni función sintáctica o semántica del mismo tipo por limitaciones de la herramienta de búsqueda.
[6] Además, hay oraciones sin sujeto (impersonales) (v. ap. 2.2).
[7] Se usa este término siguiendo las convenciones de la gramática tradicional, pero existen casos en el rol semántico asociado al sujeto lógico no es agente, sino causa, por ejemplo.
[8] Estos datos se han extraído desde http://grial.uab.es/tools/esquemas/main.
[9] Los patrones de verbos pronominales léxicos están incluidos dentro de patrones no pronominales.
[10] Mostrar los datos a partir del "primer objeto" nos permite sintetizar y extraer generalizaciones, aunque se pierde información. No se presentan los patrones completos por razones de espacio.
[11] En las oraciones con orden neutro en español el tópico (tema) coincide con el sujeto lógico y, por tanto, aplicar el término *topicalización* a estas construcciones puede confundir, ya que el sufijo *–ción* implica que hay un proceso por el cual algo que no era tópico pasa a serlo, lo cual no es así en las oraciones mencionadas. Por otro lado, dentro de la *destopicalización* no se han tenido en cuenta los casos de cambio de orden sin cambio de función del tipo "A Luis le han dejado el coche sus padres".

REFERENCIAS BIBLIOGRÁFICAS

Fernández A., Vázquez G. En prensa. "Interfaz de consulta del corpus SenSem", *Actas del XXXIX Simposio de la Sociedad Española de Lingüística*.

Langacker R. 1987. *Foundations of Cognitive Grammar*. Stanford: Stanford University Press.

Smith C. 1997. *The parameter of aspect*. Dordrecht: Kluwer Academic Publishers.

Vázquez G., Fernández A. 2008. "Annotation de corpus: Sur la délimitation des arguments et des adjoints", *SKY Journal of Linguistics*, 2: 244-269.
Xiao R., McEnery T. 2004. *Aspect in Mandarin Chinese: A corpus-based study*. Amsterdam: John Benjamins.