El banco de datos SenSem: corpus anotado y léxico para el español y el catalán

Ana Fernández-Montraveta[1], Irene Castellón[2], Glòria Vázquez[3]
[1]Universitat Autònoma de Barcelona
[2]Universitat de Barcelona
[3]Universitat de Lleida

Abstract. En este artículo vamos a presentar un breve resumen de cómo se han anotado los corpus, los datos más relevantes referentes a dicha anotación por lo que se refiere a las construcciones, que es la información más novedosa de estos recursos, y cómo se han usado dichos corpus en distintas aplicaciones de procesamiento de lenguaje natural (PLN).
Keywords: anotación de corpus, construcciones, subcategorización, roles semánticos

1 Introducción

El banco de datos SenSem[1] está compuesto por una serie de recursos lingüísticos para las lenguas española y catalana [6], [9] y [10]. Se trata de dos corpus anotados y dos léxicos verbales, uno para cada lengua, que se pueden consultar en http://grial.uab.es/sensem. En el léxico se recoge el uso de 1243 sentidos verbales, pertenecientes a los 250 verbos más frecuentes con unas 120 frases por lema. El total de frases recogidas en los corpus es de 30.274 y 25.075 para el español y el catalán, respectivamente. Los textos que componen el corpus español pertenecen al registro periodístico y literario y en el caso del catalán sólo al periodístico.

2 Anotación del corpus

Las frases se han anotado a varios niveles lingüísticos: morfológico, léxico, sintagmático y oracional. La anotación del recurso ha sido manual aunque se desarrolló una versión del corpus español anotado a nivel morfológico automáticamente mediante Freeling [3]. A nivel léxico, se han desambiguado tanto los verbos como los núcleos argumentales nominales. En primer lugar, los verbos tienen asociado un sentido del léxico que, a su vez, ha sido asignado un *synset* de WN1.6 de forma manual y su correspondencia con WN3.0 [4] de forma automática. Además, se ha desambiguado un 82% de los núcleos argumentales nominales con el mismo recurso (23.307 formas correspondientes a 3.693 lemas).

A nivel sintagmático, se han categorizado los constituyentes (argumentos y adjuntos) con información sobre su categoría sintagmática y su función sintáctica y, en el caso de los argumentos, también se les ha asignado el rol semántico correspondiente (v. Fig. 1, color

[1] Estos recursos, y los generados a partir de los mismos, se han creado gracias a la financiación recibida en diversos proyectos a lo largo de más de 10 años. Los proyectos a través de los cuales se financia en la actualidad son FFI2011-27774 y TIN2012-38584-C06-06).

amarillo). A nivel oracional se anotan las construcciones teniendo en cuenta la información anterior juntamente con el orden de expresión de los participantes, información sobre la modalidad, la aspectualidad y la polaridad [5] (v. Fig. 1, color verde).

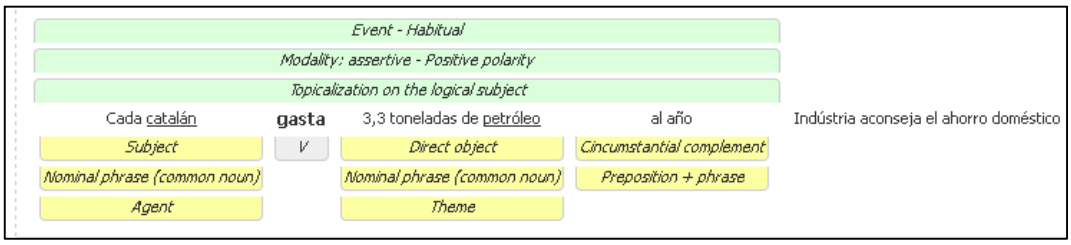

Fig. 1 Ejemplo de anotación

La anotación a nivel léxico de los verbos principales proviene de la entrada léxica (v. Fig. 2), de la que hereda el sentido de WordNet y la información aspectual (*aspectual class*).

Definition:	Hacer uso de dinero u otro bien o producto que pueda consumirse.
Semantic roles:	Agent, Theme, Purpose
Aspectual class:	Event
Wordnet:	00661955v

Fig. 2 Entrada léxica del sentido *gastar*_1

Durante los años en los que se ha trabajado en el proyecto han participado en las diferentes fases 10 lingüistas entrenados que se han ocupado de la anotación manual del corpus. Una vez establecidos los criterios, se creó un grupo de anotadores coordinados por dos personas con mayor formación. Se establecieron pruebas de acuerdo entre anotadores al inicio del proyecto. Se han realizado reuniones de coordinación, modificación de criterios y catas sobre el trabajo de los anotadores durante todo el proyecto. Por ejemplo, en el caso de la anotación semántica el acuerdo entre anotadores después del entrenamiento fue del 84.32%. El acuerdo en el corpus general fue similar variando entre el 68% y el 100% según el nivel anotado [1]. Los corpus, el léxico y todos los recursos derivados de la investigación del grupo a partir del corpus se pueden consultar y/o descargar desde http://grial.uab.es/descarregues.

3 Las construcciones del español: datos extraídos de SenSem

La anotación resultante nos permite observar los datos de la frecuencia de las distintas construcciones. En la tabla 1 es interesante observar la poca relevancia de determinadas construcciones y las distintas tendencias en algunos casos según los registros, como en la elisión de sujeto. Respecto a los esquemas de subcategorización (tabla 2) es destacable que un 22,86%

de los esquemas son ambiguos pero el resto no. No obstante, en la tabla 3 se muestra que el 90% de los esquemas quedan reflejados en 5 casos. En las tabla 4 y 5 se presentan los datos sobre modalidad/factualidad y aspectualidad [xx], respectivamente.

Tabla 1. Construcciones

Constructiones a nivel argumental	**Frecuencia**	**Periodístico**	**Literatio**
Topicalización	27,042 (89.03 %)	22,117 (88.20%)	4,948 (93.38%)
Destopicalización	3,332 (10.97%)	2,960 (11.80%)	351 (6.62%)
Elisión de sujeto	8,984 (28.58%)	6,614 (26.38%)	2,370 (44.73%)
Reflexivas	143 (0.47%)	113 (0.45%)	30 (0.57%)
Reciprocas	107 (0.35%)	90 (0.36%)	17 (0.32%)
Construcciones de dativo	230 (0.76%)	153 (0.61%)	77 (1.45%)

Tabla 2. Esquemas de subcategorización con respecto a la semántica

Número de esquemas	**70**
Sólo esquemas de topicalización	23 (32.86%)
Sólo esquemas de destopicalización	31 (44.28%)
Esquemas de topicalización y destopicalización	16 (22.86%)

Tabla 3. Esquemas sintácticos más frecuentes

1.	SN V SN	13,054 (42.98%)
2.	SN V SP	6,424 (21.15%)
3.	SN V	3,101 (10.21%)
4.	SN V SN SP	2,901 (9.55%)
5.	SN PRON V	1,873 (6.17%)
Total		90.06%

Tabla 4. Modalidad/Factualidad

Oraciones asertivas	23,451 (77.21%)
Oraciones no asertivas	6,923 (22.79%)
Futuro	4,521 (65.30%)
Pasado y presente	1,070 (15.46%)
Imposibilidad	65 (0.93%)
Epistemicidad	97 (1.40%)

Tabla 5. Aspectualidad

Estados	4,149 (13.66%)
Estados temporales	525 (12.65%)
Estados permanentes	3,624 (87.35%)
Eventos	17,037 (56.09%)
Procesos	9,188 (30.25%)
Perfectividad	10,360 (48.74%)
Imperfectividad	10,894 (51.25%)
Habitualidad	1,121 (10.29%)

4 Aplicación computacional de SenSem

Desde el punto de vista computacional, SenSem proporciona información sintáctica y semántica de diferentes niveles que es útil para procesos de PLN. En primer lugar, ha constituido un recurso central para el desarrollo de la gramática de dependencias del español *EsTxala*, gramática integrada en Freeling [8]. EsTxala utiliza la información de subcategorización extraída a partir de SenSem, además se abordó la adjunción de los sintagmas preposicionales con

información del corpus superando en 17 puntos el resultado de EsTxala para esta tarea [2]. SenSem también nos ha permitido generar un listado de las diferentes configuraciones de orden de los argumentos verbales integrado en un test suite *ParTes* orientado a la evaluación del análisis sintáctico.

Por último, en la actualidad se está llevando a cabo un proyecto sobre extracción de información y etiquetaje automático de papeles semánticos a través de reglas heurísticas basado en el concepto de similitud verbal. Este etiquetador usa SenSem como recurso que proporciona información sintáctico-semántica de los predicados verbales. La primera versión mostró resultados prometedores [6]. En el momento actual se están usando las técnicas de *clustering* para establecer clases verbales.

Referencias

1. Alonso, L., Capilla, J.A., Castellón, I., Fernández, A., Vázquez, G.: The SenSem Project: Syntactico-Semantic Annotation of Sentences in Spanish. En: Nikolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory 292, pp. 89-98. John Benjamins Publishing Co, Boroets, Bulgaria (2007)
2. Aguilar, N., Alonso, L., Lloberes, M., Castellón, I.: Resolving prepositional phrase attachment ambiguities in Spanish with a classifier. Revista de la SEPLN, 46, 75-82. Alicante
3. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3: Syntactic and semantic services in an opensource NLP library. En: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), 48-55 (2006)
4. Atserias J., Villarejo, L. Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P.: The MEANING Multilingual Central Repository. En: Proceedings of the Second International Global WordNet Conference (GWC'04)), Brno, Czech Republic (2004)
5. Fernández, A., Vázquez, G.: The SenSem Corpus: an annotated corpus for Spanish and Catalan constructions with information about aspectuality, modality, polarity and factuality. Corpus Linguistics and Linguistic Theory (in press)
6. Fernández, A., Vázquez, G., Castellón, I.: "SenSem: a Databank for Spanish Verbs". En: *Proceedings of the X Ibero-American Workshop on Artificial Intelligence, IBERAMIA*, Ribeirão Preto, Brasil (2006)
7. Gil , L., Castellón, I., Coll-Florit, M.: "Hacia una definición de la similitud verbal para la extracción de eventos". En: Actas del Congreso AESLA 2014 (en prensa).
8. Padró, L., Stanilovski, E.: FreeLing 3.0: Towards Wider Multilinguality. En: Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA, 2473-2479 Istanbul, Turkey (2012)
9. Vázquez, G., Fernández, A., Beà, E.: SenSemCat: Corpus de la lengua catalana anotado con información morfológica, sintáctica y semántica. En: Casanova, E., Calvo, C. (eds.) Actas del 26 Congreso Internacional de Lingüística y Filología Románicas, vol. VIII, pp. 159-170. De Gruyter (2013)
10. Vázquez, G., Fernández, A.: Ampliación del Banco de Datos de Verbos del español SenSem. En: Cantos, P., Sánchez, A. (ed.), A Survey on Corpus-based Research. Panorama de investigaciones basadas en corpus, pp. 957-969. U. Murcia (2009)