Interfaz de explotación del corpus SenSem

Ana Fernández Montraveta*
Glòria Vázquez García**
David Teruel Ledesma*

*Universitat Autònoma de Barcelona
**Universitat de Lleida

RESUMEN

En este artículo presentamos una interfaz de búsqueda de información desarrollada para la explotación de un corpus anotado a nivel sintáctico y semántico que se ha elaborado en el marco del proyecto SenSem[1]. Nuestro objetivo es doble: por un lado, presentar las potencialidades de la herramienta, centrándonos en las diferentes posibilidades que permite la interfaz en relación a la expresión de la información lingüística contenida en el corpus; y, por otro, dar cuenta del tipo de generalizaciones que se pueden extraer a través de la adquisición de información sobre subcategorización.

ABSTRACT

In this paper we present a search interface which has been developed for the exploitation of a corpus annotated syntactically and semantically. Said corpus has been developed within the framework of the SenSem project. Our objective is twofold: first, we would like to present the potentialities of the tool, focusing on the different search possibilities that it allows and that reflect the linguistic information that has been considered and annotated in the corpus. Secondly, we want to explore the kind of generalizations that can be inferred by means of the acquisition of subcategorization patterns.

***Palabras Clave**: anotación y explotación de corpus, esquemas de subcategorización, construcciones*

***Keywords**: corpus annotation and exploitation, subcategorization frames, constructions*

1. INTRODUCCIÓN

El proyecto SenSem se enmarca dentro de la línea de investigación de anotación de corpus y de creación de recursos que facilitan el procesamiento de los lenguajes naturales. Una vez finalizados los procesos de anotación del corpus[2] y de corrección de errores, se procederá a la creación de un recurso léxico en el que se prevé reflejar el comportamiento sintáctico-semántico de cada sentido verbal a partir de los datos obtenidos del corpus incluyendo ejemplos. El corpus y el léxico constituirán lo que se denomina un banco de datos del español. Otros proyectos parecidos en el campo de la anotación sintáctico-semántica de corpus en la actualidad son *Adesse* (García de Miguel 2004) y *FrameNet* del español (Subirats et al 2003).

En *SenSem* los verbos se describen a nivel de sentido, a diferencia de otros recursos de estas características (Muñiz et al. 2003). Las oraciones del corpus se anotan a nivel de oración y se describen en términos de construcciones, pares de forma y significado (Goldberg 1995; Croft 2001; Fillmore 2003). Por lo que se refiere a la forma, la expresamos a partir de las categorías y funciones de los constituyentes, distinguiendo además entre argumentos y adjuntos. En cuanto al significado, tenemos en cuenta, por un lado, la relación entre los participantes del

evento y el evento propiamente dicho (roles semánticos) y, por otro, la semántica de la construcción.

En este artículo presentamos la interfaz de búsqueda que se ha desarrollado para consultar el corpus anotado y cuáles son las diversas combinaciones de búsquedas que permite. Una de las consideraciones metodológicas que se ha utilizado ha sido la de almacenar los diferentes elementos descritos de forma desglosada de cara a una mejor explotación de los datos. Así, por ejemplo, es posible obtener la interpretación de la semántica global de la oración y, a la vez, realizar búsquedas de aspectos más específicos, como la combinación de roles semánticos y funciones sintácticas.

2. INTERFAZ DE BÚSQUEDAS Y VISUALIZACIÓN DE LOS RESULTADOS[3]

La interfaz diseñada incluye diversas posibilidades de búsqueda. El contenido viene condicionado, evidentemente, por las decisiones tomadas en las fases de constitución y anotación del corpus y de establecimiento de los criterios de anotación (Vázquez et al 2005). Se ha trabajado con los 250 verbos más frecuentes en español, cuyo listado se puede consultar en la primera columna de la interfaz (v. fig. 1)[4]. Por lo que se refiere a las fuentes textuales, las oraciones han sido extraídas de un corpus periodístico.

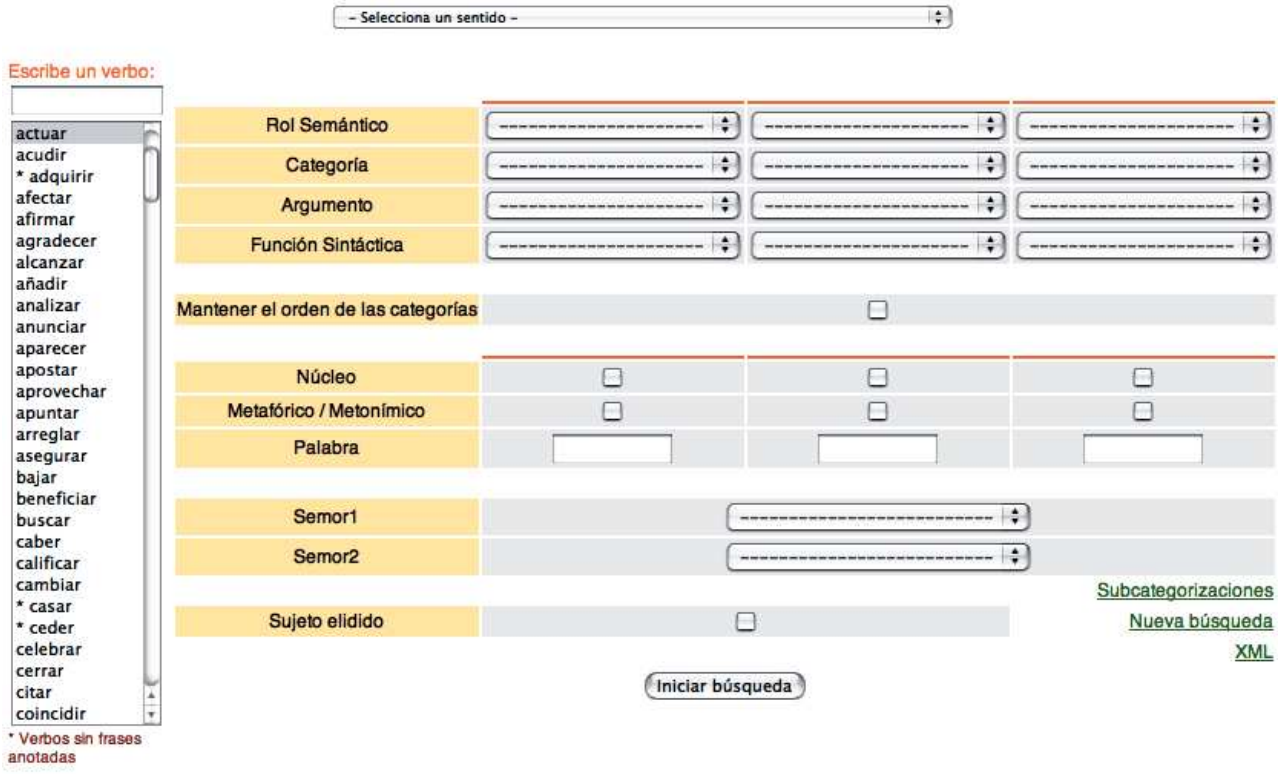

Figura 1: Menú de la interfaz de búsqueda

En la segunda columna de la figura (color naranja) pueden verse reflejados los diferentes criterios de búsqueda que permite el programa (rol semántico, categoría, etc.). Estos parámetros pueden combinarse entre sí. La única restricción que existe al definir una búsqueda se refiere al número de constituyentes, ya que sólo se pueden buscar tres

simultáneamente. Por último, el orden puede ser un elemento más a tener en cuenta en el establecimiento de los criterios de búsqueda (v. apartado 4 para una descripción más exhaustiva de las búsquedas combinadas).

La búsqueda puede realizarse para un determinado sentido verbal o para el conjunto de los sentidos de un verbo. Hasta el momento no es posible la búsqueda indiscriminada en todos los verbos de la base de datos pero esperamos disponer de esta opción próximamente.

En cuanto a la visualización de los resultados (v. fig. 2), el elemento objeto de búsqueda se presenta incluido en el contexto de una oración y resaltado en naranja, de manera que sea fácilmente identificable.

Imprimir - XML	Resultados de la búsqueda: **42 frases**
bajar-5	***Reducir la intensidad, el volumen o la cantidad de algo.***
Anotación	9281.- Y explicó que de 1992 a 1996 el precio **bajó** , aunque reconoció que en 1998 hubo un aumento que " no cubrió ni el índice de precios al consumo " .
Anotación	9275.- ¿ **Bajarán** realmente los impuestos ?
Anotación	9276.- " Después del gol , nosotros **bajamos** mucho , y el Atlético subió mucho .
Anotación	9269.- " El aforo no **debería bajar** de las 8.000 localidades " , añade .
Anotación	9266.- De entre los candidatos a las elecciones legislativas , el ministro José Montilla (PSC) , con 5,9 puntos , sigue en primer puesto Joan Puigcercós (ERC) , con 5,5 , y Josep Antoni Duran Lleida (CiU) , con 5,3 , **bajan** dos décimas .
Anotación	9257.- Los libreros estiman que las ventas **pueden bajar** entre el 15 y el 20 %
Anotación	9255.- En Europa , a finales de septiembre , las reservas de gasóleo para calefacción **habían bajado** un 3,4 % con respecto a las del año anterior .
Anotación	9254.- Pero las ventas del diario **siguieron bajando** hasta pasar de 407.000 a 380.300 ejemplares en los últimos dos años .
Anotación	9251.- La cosecha de grano **bajó** de 512 millones de toneladas en 1998 a 431 millones en el 2003 .
Anotación	9248.- Recordó que el objetivo de dedicar un 0,7 % del producto interior bruto (PIB) a ayuda al desarrollo " tiene 34 años de antigüedad " y que en el 2000 **había bajado** hasta el 0,2 % .
Anotación	9193.- La actividad del entorno radical **ha bajado** ante la pérdida de recursos .
Anotación	9191.- La cifra de beneficiarios sólo **bajó** en Andalucía y Navarra .
Anotación	9192.- " Después del gol - remarcó - , nosotros **hemos bajado** mucho .
Anotación	9190.- Las existencias de gasóleo para calefacción en EEUU **han bajado** un 6 % en un año .
Anotación	9187.- En la segunda parte , **bajó** el ritmo del partido y el Levante dejó la iniciativa al rival .
Anotación	9184.- " **Ha bajado** sustancialmente el consumo de fruta , verdura o pescado blanco , que son alimentos saludables , y ha aumentado el de carnes grasas , platos preparados y zumos envasados - - afirma Salvador - - .

Figura 2: Visualización de los resultados de la búsqueda: bajar en una construcción anticausativa

Asimismo, es posible visualizar la anotación completa de las oraciones independientemente de los parámetros utilizados para la búsqueda consultando el botón "Anotación", situado a la izquierda de la oración. Al solicitar esta información, se abre una nueva pantalla (v. fig. 3) donde se expresa gráficamente la anotación.

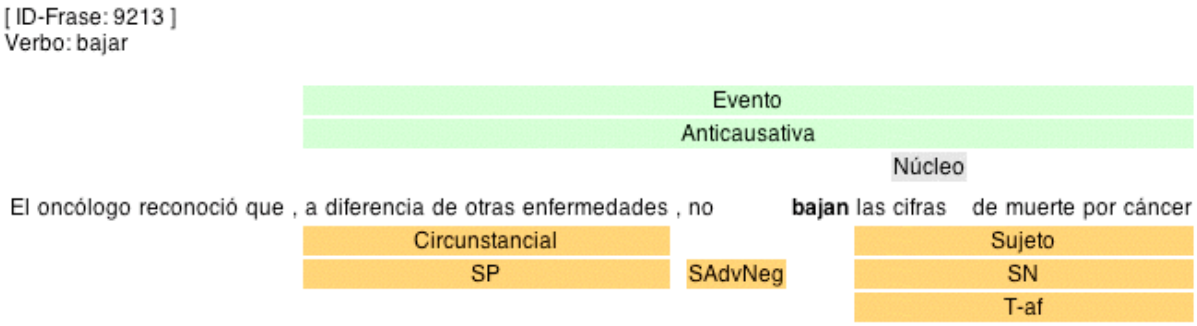

Figura 3: Anotación completa de una oración

Cada color indica un parámetro distinto, de forma que la anotación se interpreta de forma rápida. En la parte inferior, en color naranja, se puede consultar la anotación de cada complemento, sea o no argumental. En la parte superior, en color gris, se indican los núcleos, y su uso metafórico o metonímico y, en color verde, se marca la información relativa al significado aspectual de la oración y el alcance de la oración anotada.

Para finalizar, destacamos que la interfaz permite guardar los resultados de la búsqueda en ficheros con formato texto o XML.

3. BÚSQUEDA DE INFORMACIÓN ASOCIADA A PALABRAS AISLADAS

Las búsquedas que permite la interfaz a nivel de palabra son de dos tipos:

a) oraciones que contienen una determinada palabra
b) palabras que son el núcleo de sintagmas argumentales.

En cuanto al primer tipo, esta búsqueda es útil cuando se desea conocer si una determinada palabra, sin restricción respecto a su categoría, aparece en el corpus. Pueden buscarse hasta tres palabras dentro de una misma oración.

El segundo tipo de búsqueda permite recoger el conjunto de substantivos que ocupan las posiciones argumentales de SN y SP, independientemente de su función. Una de las aplicaciones de este tipo de búsqueda es la representación de las restricciones de selección en las entradas verbales del léxico. Se prevé establecer el hiperónimo compartido por estos sustantivos a través del cotejo de una jerarquía conceptual como *WordNet* (Fellbaum 1998) para identificar los tipos semánticos de los constituyentes.

Con el objetivo de no sesgar los resultados obtenidos al aplicar esta metodología, durante el proceso de anotación se marcan aquellos núcleos que son usados metafórica o metonímicamente para que no sean tenidos en cuenta a la hora de establecer el correspondiente hiperónimo. Veamos un ejemplo de uso metonímico del sujeto de una oración del verbo *actuar*:

> *Los ayuntamientos y los responsables universitarios* ***deberán actuar*** *aprovechando las sinergias que se establecen entre dos poderes locales...*

4. BÚSQUEDA DE INFORMACIÓN ASOCIADA A CONSTITUYENTES

A la hora de solicitar una búsqueda relativa al tipo de constituyentes, podemos utilizar los siguientes parámetros:

a) relevancia argumental
b) categoría sintagmática
c) función sintáctica
d) rol semántico

Como se ha mencionado anteriormente, se pueden solicitar búsquedas de hasta tres constituyentes y decidir si se mantiene o no el orden establecido. Evitar un orden estricto nos permite obtener mayor generalización en los resultados. Por ejemplo, si buscamos la secuencia de roles agente y manera en el verbo *actuar* sin mantener el orden, el sistema nos devolverá, entre otras, las siguientes frases:

> *... la defensa del Depor*$_{ag}$ ***actuó*** *a un buen nivel*$_{ma}$ *y se sacudió...*
> *Camuflados entre estos jóvenes* $_{ma}$*...,* ***actúan*** *grupúsculos agresivos* $_{ag}$*...*

Se ha utilizado un sistema visual para localizar las búsquedas de más de un constituyente. Así, si sólo se rellena la información de la primera columna, el color naranja de la línea horizontal que aparece encima (v. fig. 1) es el que se usa para marcar el constituyente en el contexto de la oración (v. fig. 2). En cambio, si se rellena más de una columna, el programa designa un color distinto para cada una y usa los mismos colores en la visualización de los resultados.

Además, los cuatro parámetros mencionados (relevancia argumental, categoría, función y rol) pueden ser combinados entre sí en la misma búsqueda, lo cual puede ser muy útil para estudiar, por ejemplo, las diferentes funciones sintácticas de los roles semánticos. Así, podría ser interesante observar el comportamiento de determinados roles que se han asociado con funciones sintácticas específicas, como es el caso del agente y el sujeto.

En tanto que los adjuntos no son relevantes semánticamente para el verbo, no han sido asociados a ningún rol semántico, pero sí se puede consultar su función y categoría. Como es sabido, la distinción entre argumento y adjunto no está exenta de problemas y es un tema que no está resuelto en la teoría lingüística. Asimismo, hemos comprobado que es uno de los aspectos de mayor discrepancia entre los anotadores (Alonso et al. 2005). Esperamos, por tanto, que un estudio pormenorizado de este aspecto de la anotación nos pueda orientar y corregir en la decisión de considerar el papel argumental de determinados sintagmas.

En cuanto a la información relativa a la función sintáctica, este parámetro se complementa con la posibilidad de encontrar también las oraciones con sujeto elidido. Se ha decidido incluir esta posibilidad en la interfaz ya que creemos que puede ser útil para un estudio pormenorizado de dicho fenómeno.

5. BÚSQUEDA DE INFORMACIÓN ASOCIADA AL CONJUNTO DE LA ORACIÓN

En este apartado vamos a definir la información que se puede solicitar en relación al conjunto de la oración. Existen básicamente dos posibilidades:

a) búsqueda de patrones de subcategorización,
b) búsqueda de información semántica.

En cuanto al primer tipo, el sistema permite obtener información sobre los esquemas de subcategorización de un verbo encontrados en el corpus consultando el botón "Subcategorizaciones", situado en el extremo inferior derecho de la pantalla principal (v. fig. 1). Como se puede observar en la figura 4, el sistema nos muestra una nueva pantalla donde se presentan los resultados para el verbo escogido y donde se pueden elegir diferentes visualizaciones de los patrones, ya sea formados sólo por las categorías (como se observa en dicha figura), los roles semánticos o las funciones o bien por la combinación de dos o tres de estos parámetros. Además, en el patrón se indican también los sujetos elididos, resaltados en rojo, así como la frecuencia, a través de un número situado al final de cada patrón. Se puede escoger el orden de presentación de los resultados: según el tipo de los constituyentes o el número de ocurrencias.

En la actualidad, se muestran todas las realizaciones de los argumentos, por lo que el número de patrones asociados a un verbo es muy extenso. Estamos trabajando en la compactación de las categorías sintácticas con el fin de presentar los datos de forma más generalizada.

En cuanto a la semántica de la oración, se han descrito dos aspectos: la semántica de la construcción y la interpretación aspectual de la oración.

Por lo que respecta a esta última se han descrito dos niveles de anotación: el del sintagma verbal y el de la oración. En el primer caso, a través del campo "semor1" (semántica oracional 1) hemos definido la estructura eventual del sintagma verbal. Las posibles etiquetas son: evento, proceso y estado. Desde el punto de vista léxico, algunos verbos, como *analizar,* son designados como ambiguos, entre proceso y evento, y es en el momento de la anotación cuando se decide la etiqueta pertinente en cada caso:

> *... hemos* analizado *los precios y el mercado y hemos tomado la decisión apropiada.*
> *Mientras los responsables de los Mossos d'Esquadra* analizaban *los riesgos de entrar por la fuerza en el módulo , ...*

En la primera oración se considera que el evento está limitado por la existencia de un "análisis" final que propicia que se "tome la decisión apropiada". En la segunda oración el valor de proceso viene dado por la ausencia de telicidad marcada por el uso de "mientras".

Por otro lado, en el campo "semor 2" se puede asignar una interpretación aspectual de tipo habitual cuando sea necesario. Para ello se tendrán en cuenta más elementos, como determinados adjuntos y el tiempo verbal. Por ejemplo, el verbo *analizar*, etiquetado como proceso o evento a nivel de SV (semor1), pasa a tener una interpretación estativa en la siguiente oración (semor2):

con ese panorama genético , Andrés Ortega dirige la edición españo-la de Foreign Policy , la revista que analiza *el mundo .*

Figura 4: Visualización de los patrones de subcategorización

El campo "semor2", además de la información aspectual oracional, incluye otras etiquetas que hacen referencia al tipo de construcción que expresa la oración, por ejemplo, al elemento que constituye el foco comunicativo (construcción anticausativa, pasiva, impersonal), la reflexividad y la reciprocidad u otros fenómenos como la expresión de un dativo de interés. Así, a diferencia de lo que ocurre con la semántica oracional 1, en el campo de la semántica oracional 2, una oración puede ser anotada con más de una etiqueta. Por ejemplo, una frase como *Esa solicitud puede ser rechazada si en el país donde ha sido detenido el delincuente se le inicia un proceso por los mismos hechos por los que fue reclamado* se anota como "antiagentiva" y "dativo de interés".

CONCLUSIÓN

En este artículo hemos presentado una herramienta para la explotación de un corpus anotado del español. La elaboración de este corpus es una contribución en un campo incipiente de la anotación de corpus, el de los niveles sintáctico y semántico. La interfaz de búsqueda que se ha desarrollado permite la consulta eficiente y rápida de los fenómenos reflejados en la anotación.

La herramienta de explotación creada es muy versátil, ya que permite combinar búsquedas de diferentes niveles y para diferentes elementos de la misma oración al mismo tiempo. Además,

la forma de visualizar los resultados pretende facilitar el acceso rápido de la información, aún cuando se ha requerido más de un objeto de búsqueda. También se pueden consultar otros aspectos de la anotación que no se han solicitado para obtener una visión más amplia de la anotación.

Por último, los verbos del corpus han sido anotados a nivel léxico-semántico, lo cual implica una inversión de tiempo importante, pero es un valor añadido del corpus creado y de la interfaz, ya que el usuario podrá obtener la información para cada sentido verbal, que es fundamental a la hora de poder usar de forma válida la información obtenida.

---

1. Este trabajo ha sido realizado gracias a la financiación otorgada por el Ministerio de Ciencia y Tecnología (MCyT, BFF2003-06456).

2. En la actualidad se han anotado 19.400 oraciones, de un total de 25.000.

3. La interfaz actual es de uso restringido para los miembros del proyecto, mientras se acaban de anotar y corregir por completo las oraciones del corpus. Disponemos de una interfaz de demostración de la anotación en : http://grial.uab.es/demo/.

4. En la actualidad los verbos marcados con un asterisco no se encuentran todavía disponibles.

REFERENCIAS BIBLIOGRÁFICAS

Alonso, L., J. A. Capilla, I. Castellón, A. Fernández y G. Vázquez. 2004. "The SenSem Project: syntactico-semantic annotation of sentences in Spanish". *Proceedings of the International Conference Recent Advances in Natural language Processing*, Bulgaria, 39-46.

Croft, W. 2001. *Radical Construction Grammar*. Oxford: Oxford University Press.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: The MIT Press.

Fillmore, Ch. 2003. *Form and meaning in Grammar*. CSLI Lecture Notes. Center for the Study of Language and Information.

García de Miguel, J. y S. Comesaña. 2004. "Verbs of Cognition in Spanish: Constructional Schemas and Reference Points". *Linguagem, Cultura e Cogniçao: Estudos de Linguística Cognitiva*. Eds. A. Silva, A. Torres y M. Gonçalves. Coimbra: Almedina, 399-420.

Goldberg, A. 1995. *Construction Grammar*. Chicago: Chicago University Press.

Muñiz, E., M. Rebolledo, G. Rojo, O. Santalla y S. Sotelo. 2003. "Description and exploitation of a BDS: a Syntactic Database about verb Government in Spanish". *Proceedings of Recent Advances in Natural Language Processing*, Bulgaria, 297-303.

Subirats-Rüggeberg, C. y M. R. L. Petruck. 2003. "Surprise: Spanish FrameNet!". *Proceedings of the International Congress of Linguists*, Praga. http://framenet .icsi.berkeley.edu/~framenet/papers/SFNsurprise. pdf.

Vázquez, G., A. Fernández y L. Alonso. 2005. "Guidelines for the syntactico-semantic annotation of a corpus in Spanish". *Proceedings of the International Conference Recent Advances in Natural language Processing*, Bulgaria, 603-607.