INTERFAZ DE CONSULTA DEL CORPUS DEL ESPAÑOL SENSEM

Ana Fernández Montraveta
U. Autònoma de Barcelona

Gloria Vázquez
U. de Lleida

1. EL BANCO DE DATOS SENSEM

En este trabajo presentamos una interfaz de búsqueda que permite la explotación de los datos recogidos en el corpus SenSem (Vázquez et al. 2008). SenSem es un banco de datos del español en cuya creación llevamos trabajando los últimos 8 años.[1] Dicho banco de datos se compone, por un lado, de un corpus constituido por 850.000 de palabras y un léxico verbal creado a partir de los datos recogidos en el corpus y que presenta más de 1.100 entradas –sentidos verbales.

El objetivo de este recurso es la descripción del comportamiento sintáctico y semántico de los verbos y de las oraciones del español usando una metodología de base empírica basada en el estudio del comportamiento de la lengua en textos escritos (corpus) pertenecientes al uso real. Para ello se ha procedido a crear el corpus recopilando textos y a su anotación. Posteriormente, se volcarán los datos en las entradas verbales de los predicados de las oraciones anotadas.[2]

Actualmente, en SenSem nos encontramos en la fase de ampliación del proyecto, en la que estamos elaborando diversas mejoras. En este artículo presentamos parte de los resultados de dichas mejoras a partir de la nueva interfaz de búsqueda del corpus.[3] En el apartado 2 se presenta la estructura básica de la herramienta y en los siguientes apartados (hasta el 7 incluido) se van describiendo una a una cada sección del menú principal. En el apartado 8 se describe cómo se visualizan los resultados obtenidos al realizar una búsqueda. Finalmente, se presentan las conclusiones, juntamente con las áreas de aplicación más importantes y las líneas de trabajo futuro.

Otros proyectos relacionados con el mencionado para la lengua española son Framenet español (Subirats-Rüggeberg, Carlos y Miriam R. L. Petruck (2003)) y ADESSE (García-Miguel, José M et al. (2005), Albertuz Carneiro, Francisco (2007). En dichos proyectos se prevé, sobre todo, dar cuenta de la información sintáctico-semántica a nivel de constituyentes, indicando sus funciones sintácticas y semánticas, principalmente, además del sentido verbal. En el caso de ADESSE se añade también

[1] El primer proyecto fue "SenSem: Banco de datos sintáctico y semántico del español" (Ministerio de Ciencia y Tecnología BFF2003-06456) y el actual, llamado "Ampliación de la BD léxica y el corpus sintáctico-semántico de semántica oracional del español SenSem" se está llevando a cabo gracias a la financiación del Ministerio de Educación y Ciencia - HUM2007-65267.
[2] La base lexicográfica resultante de este volcado estará disponible al finalizar el proyecto.
[3] Para la descripción de la interfaz anterior, v. Fernández *et al*. 2007.

otro tipo de información relativa al nivel oracional, que es el tipo de construcción y la modalidad.

La diferencia principal respecto a los proyectos mencionados es que en SenSem, además de la información sintáctico-semántica de los participantes y la semántica de la construcción, se añade la codificación de aspectos innovadores en dicho campo, como es la aspectualidad. Además se ha optado por una nomenclatura que, en la medida de lo posible, no está asociada a ninguna teoría en concreto, con el fin de crear una herramienta que pueda ser usada por un número de usuarios lo más amplio posible.

2. LA INTERFAZ DE BÚSQUEDA[4]

Las mejoras que incorpora el nuevo buscador tienen que ver con la visualización de aspectos de la anotación que hasta el momento no podían ser consultados, como la modalidad, la polaridad y la aspectualidad. A través de esta herramienta también se pueden recuperar nuevas oraciones anotadas, todas ellas pertenecientes a un nuevo registro, el literario, que se han incorporado en la nueva fase del proyecto, ya que en la primera fase sólo se había trabajado con el registro periodístico. Actualmente, del subcorpus literario se puede consultar ya el 50% y a finales del 2010 se espera tener disponible el 100%.

Además de las mejoras cuantitativas mencionadas, también se han llevado a cabo mejoras de tipo cualitativo. Por un lado, desde el punto de vista organizativo, se han reestructurado los menús de la primera versión de la interfaz relacionados con la semántica de la construcción. Dicha reestructuración presenta importantes ventajas que comentaremos más adelante (v. ap. 5).

Por otro lado, a nivel informático se ha mejorado el diseño gráfico de la interfaz y se ha renovado el motor de búsqueda. En el primer caso, la interfaz es más amigable y la visualización de los resultados y de la anotación es más clara; respecto al segundo aspecto, el sistema es más efectivo a la hora de realizar las búsquedas.

En la figura 1 se presenta el aspecto de la interfaz y su menú de selección básico. En el espacio de la derecha, se mostrarían las oraciones recuperadas según el criterio o los criterios establecidos por el usuario, ya que las opciones elegidas se pueden acumular para filtrar la búsqueda sin límite alguno. En la columna de la izquierda de la pantalla aparece la estructura básica del menú, que contiene 5 apartados (idioma y corpus, verbo, oración, participantes y palabras) que iremos presentando a lo largo del artículo.

[4] La herramienta de búsqueda descrita en este trabajo puede consultarse en la siguiente dirección web: http://grial.uab.es/tools/buscador.

Cada uno de estos submenús permite realizar una selección por parte del usuario de aquello que se quiere consultar, teniendo en cuenta que todas las opciones se pueden ir añadiendo como criterios de búsquedas. En dicha figura se visualizan los submenús correspondientes a los campos “Idioma y corpus”, que permiten acotar los textos de la búsqueda, y “Verbo”, que se usa para seleccionar verbos concretos o sentidos de los mismos (v. ap. 3). A través del apartado “Oración” se pueden recuperar frases que ejemplifican diversos fenómenos lingüísticos que tienen que ver con la semántica global de la oración (v. ap. 4 y 5). La sección “Participantes” permite realizar búsquedas específicas en relación a los constituyentes de las oraciones (v. ap. 6). Por último, también se pueden realizar búsquedas por palabras (v. ap. 7).

Figura 1. Interfaz de búsqueda del corpus SenSem[5]

3. IDIOMA, REGISTRO Y VERBO

La primera opción para filtrar oraciones en el buscador consiste en seleccionar el idioma del texto y/o el fragmento de corpus que se quiere consultar. En la actualidad se puede consultar sólo el español pero ya se está trasladando la anotación al catalán y esperamos tener resultados públicos para finales del presente año.

Mediante el identificador *Fuente* se ofrecen los dos registros recogidos, el periodístico y el literario, desglosando, en cada caso, los diarios y las obras literarias que han sido utilizados como fuente para la obtención de las frases.[6]

[5] La interfaz está disponible también para los idiomas inglés y catalán.
[6] El detalle de la fuente de donde se ha obtenido cada oración está incluida en la base de datos, así como otros datos también específicos sobre autores y fechas.

Para el español el corpus está constituido por 30.000 oraciones, lo cual supone aproximadamente unas 850.000 palabras.[7] El 83,3% de dicho corpus pertenece al registro periodístico. Concretamente, este subcorpus está constituido por 25.000 oraciones (100 oraciones para cada uno de los 250 lemas verbales incluidos en el estudio), que están compuestas por unas 700.000 palabras. El 16,7% restante lo constituyen fragmentos de textos literarios de tipo narrativo de autores españoles del siglo XX y XXI. Se trata de un subconjunto compuesto por 150.000 palabras y 5.000 oraciones (20 oraciones para cada uno de los 250 verbos). Con el primer subcorpus se dio cobertura al 58% de los sentidos descritos en la base léxica verbal. Se espera aumentar esta cifra con la inclusión de los textos literarios. También se espera aumentar el número de fenómenos lingüísticos recogidos para cada sentido (mayor número de construcciones y esquemas sintáctico-semánticos). Todo ello se evaluará al final del proyecto.

Los escollos que se han encontrado en la ampliación del registro textual han sido de dos tipos. Por un lado, ha sido algo dificultoso obtener textos literarios de las características descritas en formato electrónico, ya que, aunque existen muchos portales literarios, la mayoría de obras recogidas son traducciones y clásicos. Por otro lado, algunos lemas que pueden tener una elevada frecuencia en un corpus general de la lengua no son habituales en el registro literario (*efectuar*). En estos casos ha sido difícil conseguir 20 ocurrencias de dichos verbos en la primera fase de extracción de obras literarias. Se calculó que, para conseguir el número de oraciones que faltaba, hubiera sido necesario el añadido de textos con un número de palabras muy elevado. Para superar este problema se inició una segunda fase en la que se recurrió al uso del CREA (Corpus de Referencia del Español Actual, de la Real Academia de la Lengua Española).[8] Concretamente, se ha utilizado este recurso para extraer 1.267 oraciones, que constituyen un 25,34% del subcorpus literario de SenSem), que están siendo anotadas como el resto de frases del proyecto.

Por lo que respecta a la organización del corpus en la base de datos, todas las frases se asignan, en primer lugar, a un verbo (entrada verbal) y, posteriormente, a un sentido del mismo. Para la asignación del verbo se procedió a lematizar los textos y extraer el número de frases requerido para cada verbo. Para la asignación del sentido se ha procedido manualmente.

[7] Totalmente anotadas están las palabras que forman parte de un argumento; parcialmente anotadas las que son parte de un adjunto. El contexto (palabras en la oración que no dependen sintácticamente del verbo) no se han anotado.

[8] La interfaz de consulta de este corpus permite hacer búsquedas en subpartes del mismo, por lo que se pudieron mantener los criterios de composición del subcorpus literario SenSem. Como desventaja cabe señalar que en el CREA sólo se pueden consultar formas y no lemas, por lo que se optó por buscar formas previsiblemente frecuentes, como la 3ª persona singular o plural del pasado perfecto, imperfecto o futuro.

En la interfaz se ha pretendido reflejar esta misma organización, ya que creemos que presenta un número importante de ventajas al usuario a la hora de diseñar las búsquedas. En concreto, en la sección “Verbo”, el usuario puede elegir entre 3 opciones de búsqueda en este caso: todos los verbos, un solo verbo del conjunto de todos los verbos, un solo sentido del conjunto de sentidos de un verbo. Respecto al primer caso, esta opción es muy útil para observar el alcance de un fenómeno lingüístico de forma general en la lengua. En cuanto al segundo caso, nos permite ver la distribución de un fenómeno entre los sentidos de un mismo verbo. Por último, el usuario también puede focalizarse en el análisis de un fenómeno para sentidos específicos. Hay que tener en cuenta, sin embargo, que para algún sentido no hay ocurrencias asociadas, ya sea porque sean usos menos frecuentes en la lengua o porque en la selección de las oraciones, que es aleatoria, no ha coincidido su aparición. En este sentido, observamos una limitación del corpus presentado por lo que se refiere a su explotación en cuanto al estudio de frecuencias de uso.

4. ASPECTUALIDAD, MODALIDAD Y POLARIDAD

La siguiente sección del marco de la izquierda (“Oración”) nos permite llevar a cabo búsquedas relacionadas con la anotación realizada a nivel oracional: la aspectualidad, la modalidad, la polaridad y la construcción. En este apartado nos vamos a centrar en las tres primeras.

Por lo que respecta a la aspectualidad, esta información se encuentra reflejada en los identificadores *Aspecto* y *Aspectualidad*. La metodología utilizada en la anotación para dar cuenta del significado relacionado con el aspecto entronca con la utilizada por Xiao y McEnery 2004. Se trata de un tipo de anotación que muestra cómo se va construyendo de forma composicional el significado aspectual de las oraciones, teniendo en cuenta desde el tipo eventivo del predicado hasta el tipo de argumentos de que se acompaña, las desinencias morfológicas y los diversos adjuntos que se utilizan (Smith 1997).

En el identificador *Aspecto* el usuario puede elegir entre tres opciones sobre el tipo eventivo del predicado. Las posibles interpretaciones son: *estado*, *evento* y *proceso*. La etiqueta *evento* se utiliza cuando se detecta la presencia de un límite cuantitativo en la acción, mientras que la etiqueta *proceso* es utilizada cuando se da una ausencia de este límite. Esta información puede coincidir o no con el aspecto léxico declarado en el sentido verbal. Así, un verbo como escribir 1, que no está

determinado léxicamente por lo que se refiere a la ausencia o presencia de límite, se actualiza en cada oración según el sintagma que lo modifique.[9]

Con el identificador *Aspectualidad* se define aquella información de tipo aspectual que aportan otros elementos de la oración más allá de la complementación dentro del SV, como los auxiliares, las desinencias verbales o determinados adjuntos. Los valores que se han tenido en cuenta para la descripción de la aspectualidad son tres: en primer lugar, si la oración en su conjunto presenta una lectura *perfectiva* o *imperfectiva*; en segundo lugar, si se da una lectura *habitual*, relativa a acciones que se presentan como reiteradas en el tiempo;[10] y, por último, si la frase es estativa, diferenciando entre el aspecto *estativo permanente* y *temporal*.

Finalmente, bajo el término *Modalidad* se presenta información sobre el carácter asertivo o no asertivo de la proposición. Asociada a la modalidad, el usuario también dispone de información relativa a la etiquetación de la oración con respecto a la *Polaridad*, que puede ser positiva o negativa.

5. SEMÁNTICA DE LA CONSTRUCCIÓN

La información que se presenta relacionada con la semántica de la construcción en la sección "Oración" ha sido reestructurada respecto a la versión anterior, en la cual se había optado por presentar un listado de construcciones (anticausativa, pasiva, impersonal, etc.). En este segundo proyecto se ha intentado presentar dicha información desde una perspectiva lo más general posible desde el punto de vista de la terminología y la teoría lingüística, así como de la lengua objeto de estudio, en tanto que la idea es que el modelo sea lo más exportable posible. Es por este motivo que se ha decidido partir de la estructura informacional de las oraciones para reflejar este tipo de fenómenos. Se ha diferenciado, pues, entre las construcciones en que el sujeto lógico es el tema o tópico de la oración (a través del identificador *Topicalización del sujeto lógico*), y las construcciones en que el elemento topicalizado es distinto al sujeto lógico (a través del identificador *Topicalización de otros participantes*). La estructura presentada nos ha permitido hacer una distinción entre frases activas donde el sujeto lógico coincide con el sintáctico y otras frases, entre las que se incluyen las pasivas[11] y las anticausativas, donde el sujeto lógico no coincide con el sintáctico.

Tanto para los casos de topicalización del sujeto lógico como de topicalización de otros participantes, se presenta un menú desplegable donde el usuario puede

[9] Una oración como "Lo escribí a mano y lo regalé a mis clientes" ha sido anotada como evento, mientras que "¿ Ha escrito sobre el 11- M?" ha sido anotada como proceso.

[10] La siguiente oración del verbo *acercarse* es un ejemplo de frase anotada con aspectualidad habitual: "Cuando *se acercan* las elecciones la gente se pone más nerviosa".

[11] Por el momento, se han incluido en este apartado tanto las pasivas pronominales como sintácticas, aunque algunos estudios (Pinuer 2005) revelan que algunas pasivas de este último tipo pueden ser casos de rematización y no de tematización o topicalización.

solicitar en cada caso la recuperación de todas las oraciones o bien diferenciar según dos criterios. En el primer tipo de topicalización, el usuario puede escoger, en primer lugar, el rol semántico del sujeto lógico topicalizado y, en segundo lugar, puede solicitar sólo los casos de reflexividad o reciprocidad. En cuanto al segundo tipo de topicalización, se puede elegir también, por un lado, el rol semántico del participante destopicalizado y, por otro, el mecanismo formal usado en la oración para dicha topicalización (pronominal, sintáctico o impersonal).[12]

Esta forma de estructurar la interfaz creemos que presenta muchas ventajas ya que va a permitir a los usuarios del recurso definir mejor sus búsquedas, aunque a primera vista no sea tan práctico como el uso de etiquetas del tipo *pasiva*. Por ejemplo, un investigador interesado en el fenómeno de la pasiva de verbos agentivos (rol semántico agente) va a poder filtrar las oraciones que deben ser objeto de su estudio de forma más adecuada y evitarse entre sus resultados oraciones como "No era cosa de ir entonces a cambiarlo, *se perdería tiempo y ocasión en ello*", donde la construcción pronominal en cursiva no se corresponde con una acción agentiva. O, visto desde otra perspectiva, las etiquetas usadas en la interfaz pueden hacer consciente al usuario de que el término *construcción pasiva*, tradicionalmente asociado a significados agentivos, también puede usarse en sentido más amplio para designar oraciones de verbos que no son agentivos, como el ejemplificado anteriormente. Algo parecido podría decirse sobre el uso de los términos *construcción media* o *anticausativa*, para los cuales tampoco hay un consenso entre los propios lingüistas. Consideramos que, aunque también puede haber discrepancias sobre cómo definir los distintos tipos de roles semánticos, su uso presenta muchas más ventajas que los términos anteriores, ya que su significado es bastante intuitivo y, combinándolos con los conceptos de focalización y sujeto lógico, creemos que el usuario puede realizar consultas bastante acordes con sus intereses.

Por último, en la parte inferior de la sección "Oración" también se le ofrece al usuario la posibilidad de recuperar oraciones que se hayan anotado según reflejen un caso de construcción de dativo o de elisión de sujeto.

6. PARTICIPANTES

En la sección "Participantes" se pueden realizar búsquedas sobre la sintaxis y/o la forma para uno o diversos constituyentes a la vez. Esta búsqueda puede realizarse a través de cualquiera de las cuatro propiedades que se han usado para anotar dichos

[12] Esta diferenciación de mecanismos es útil sobre todo para los distintos tipos de pasiva, pero también nos hemos encontrado con diversos tipos de mecanismos asociados a la anticausatividad. Respecto a las construcciones impersonales nos referimos aquí a las pronominales. En un futuro, se ha previsto incluir otros mecanismos, como la dislocación.

elementos en el corpus: rol semántico, categoría sintáctica, función sintáctica o tipo de constituyente (argumento / adjunto).

Cada una de estas cuatro categorías presenta un menú desplegable y en las tres primeras las búsquedas se pueden realizar a distintos niveles de generalización. Tomemos como ejemplo el caso del identificador *Rol semántico*. Cabe decir que dentro del proyecto se ha convenido crear algunos roles más allá de los prototípicos. El motivo es que los listados provenientes de las propuestas teóricas no nos han sido suficientes para la anotación masiva de oraciones. Así, por ejemplo, además de las etiquetas agente y experimentador, disponemos de la etiqueta agente-experimentador. La primera (agente) se usa para los casos de voluntariedad y control (*ordenar* 'poner en orden') y la segunda (experimentador) para los participantes de procesos mentales que experimentan sin quererlo (*lamentar*)[13]. La tercera (agente-experimentador), en cambio, se ha creado para etiquetar aquellos participantes de procesos mentales de tipo voluntario, como *estudiar*.

Figura 2. Submenú del identificador *Rol semántico* (sección "Participantes")

Como puede verse en la figura 2, el sistema permite solicitar oraciones con el rol semántico agente, sea del tipo que sea (agente – todos), o bien sólo el agente más prototípico (*ordenar*), o bien los otros subtipos por separado. Además del agente-experimentador, se ha usado también el agente plural (*compartir*), el agente-destino (*adquirir*), el agente-origen (*decir*), el agente-tema objeto desplazado (*acercarse*). Por último, existe el doble rol agente / causa, pensado para aquellos casos en que no sea posible desambiguar entre ambas opciones por el contexto.

7. PALABRAS

Alternativamente, el usuario puede optar por la búsqueda por palabra, es decir, por forma, de cualquier elemento textual del corpus. Como complemento a esta

[13] No todos los participantes que experimentan un sentimiento son etiquetados como experimentadores, sino sólo aquellos que no ven afectado su estado. Por tanto, el objeto de una oración como "El ruido asustó al niño" será considerado tema afectado.

búsqueda básica, para la herramienta descrita se han incluido en esta sección los identificadores *Núcleo* y *Anotado*, que tienen una utilidad pensada para la obtención del tipo semántico de los argumentos. En cuanto a este último, permite buscar solamente aquellas palabras que forman parte de la anotación de una frase con el fin de descartar las búsquedas de aquellas palabras que aparecen en los contextos de las oraciones anotadas. Por otro lado, el identificador *Núcleo* sirve como filtro para requerir que la palabra buscada esté anotada como núcleo o no del sintagma al que pertenece. Asimismo, el identificador *Núcleo* también ofrece la opción de búsqueda a las palabras anotadas con sentido metafórico.

8. RESULTADOS DE LAS BÚSQUEDAS

El usuario puede consultar el número total de oraciones que responden a la búsqueda, y no sólo un extracto limitado a una cifra. Como puede observarse en la figura 2 los resultados se despliegan en el marco derecho de la pantalla de la interfaz. Al principio se indica el número de frases recuperadas y se visualizan un máximo de 100 por pantalla (página). Dichos resultados, se presentan ordenados por orden alfabético según el verbo al que estén asociados (en verde). Al lado de la forma verbal aparece entre paréntesis cuántos sentidos de dicho lema aparecen en los resultados de la búsqueda. Seguidamente se presentan las frases recuperadas distribuidas por cada uno de los sentidos a los que están asociados. El número de sentido se indica en negrita, seguido de la definición. Por ejemplo, para la búsqueda ejemplificada en la figura 3, el primer resultado se refiere al verbo *arreglar*, de cuyos 10 sentidos sólo ha cumplido los criterios de la búsqueda el sentido 2. En el extremo derecho de la línea donde se visualiza el sentido, aparece el número de frases que cumplen los criterios de búsqueda del total de frases de que se dispone en la base de datos con dicho sentido. En el caso que nos ocupa, sólo se ha recuperado una oración de las 58 asociadas al sentido 2 de *arreglar* (1/58 frases).

Cada oración se presenta con un identificador a la izquierda, que es el localizador de la misma en la base de datos. El verbo de la oración que es núcleo del segmento anotado aparece en negrita. Al final de la oración se presentan dos enlaces: “anotación” y “más info”. Este último sirve para recuperar información más detallada sobre la fuente, así como el contexto de la oración en el párrafo. El primer enlace sirve para ver gráficamente la información asociada a la frase. En la figura 4 aparece la anotación de la oración de *arreglar 2*. La oración continúa hacia la derecha, y el usuario puede desplazarse con la barra horizontal de color azul para leer la frase completa.

También se puede solicitar la visualización de todas las anotaciones a la vez con el botón que aparecen en el extremo superior derecho.

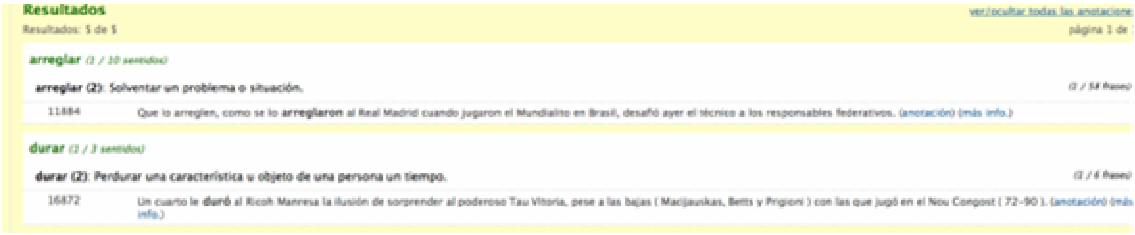

Figura 3. Resultados de una búsqueda con la información sin expandir.

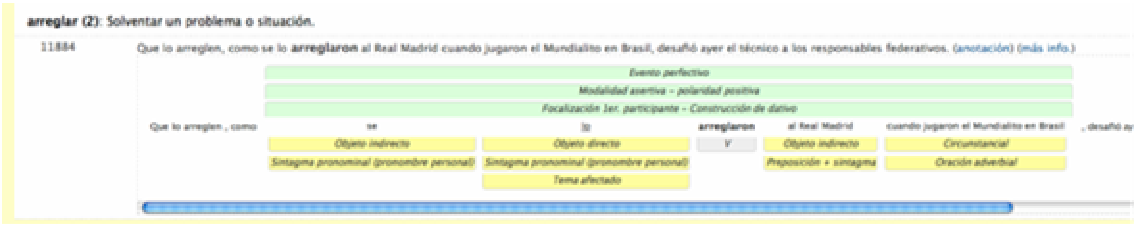

Figura 4. Vista de la anotación

9. CONCLUSIONES Y LÍNEAS FUTURAS

En este artículo se ha presentado la nueva interfaz de búsqueda del corpus SenSem. Dicha herramienta presenta una serie de mejoras tanto cuantitativas como cualitativas respecto a la versión anterior. Así, se ha incrementado el número y tipo (registro) de frases anotadas y el tipo de fenómenos lingüísticos. También se ha reestructurado la información de algunos menús de búsqueda.

Entre las posibilidades de búsqueda que ofrece el sistema, quisiéramos destacar cuatro en concreto: a) se pueden realizar consultas sobre todo el corpus y visualizar sin límites los resultados, b) se pueden sumar todos los criterios de búsqueda; c) se puede consultar información sobre la semántica oracional, como la topicalización de los participantes, la aspectualidad y la modalidad de las oraciones; y d) se ha intentado huir de enfoques lingüísticos particulares con el fin de que sea de provecho para el mayor número de usuarios posible.

Cabe mencionar que la herramienta presentada no muestra su aspecto final, ya que nuestra intención es ir introduciendo cambios en la medida que vayamos avanzando en aspectos de la anotación que aún no están resueltos. Como futuras mejoras prevemos subclasificar las oraciones no asertivas, las llamadas construcciones de dativo y las llamadas recíprocas.

Por otro lado, pretendemos también incrementar el número de fenómenos incluidos dentro de la topicalización de los participantes distintos al sujeto lógico, ya

que deberían tenerse en cuenta los casos de construcciones dislocadas sin énfasis. Además, con el fin de ser coherentes con el modelo adoptado y de ofrecer una descripción más completa de los fenómenos lingüísticos, deberían tenerse en cuenta también los posibles casos de rematización. Ambos tipos de fenómenos no han sido codificados directamente en el corpus pero pueden ser identificados a través de la presencia de determinados rasgos, por lo que se considera que ambas tareas son susceptibles de ser realizadas sin un coste elevado.

Las utilidades de este corpus son diversas. Creemos que puede ser interesante su uso para realizar estudios lingüísticos muy variados e innovadores sobre diversos temas específicos del campo de la sintaxis y la semántica. Además, permite extraer datos interesantes de cara a la construcción de léxicos verbales con información sobre los usos oracionales de los verbos, como las construcciones y las restricciones de selección. Este tipo de información puede ser de utilidad para la creación tanto de obras lexicográficas diseñadas para aprendices de español como L2 como de módulos léxicos de distintos sistemas relacionados con el procesamiento del lenguaje natural (por ejemplos, sistemas de traducción automática). Por último, la existencia de corpus anotados permite la creación de gramáticas fundamentadas en modelos estadísticos que se usan para el análisis textual automático.

REFERENCIAS

Albertuz Carneiro, F. (2007). "Sintaxis, semántica y clases de verbos: Clasificación verbal en el proyecto ADESSE". En Cano López, P. (coord): *Actas del VI Congreso de Lingüística General, Santiago de Compostela,* Vol. 2, Tomo 2, *Las lenguas y su estructura* (IIb), 2015-2030.

Fernández, A., G. Vázquez y D. Teruel (2007). "Interfaz de explotación del corpus SenSem". En Maizal, R. *et al*. (eds.), *Aprendizaje de lengua, uso del lenguaje y modelación cognitiva. Perspectivas aplicadas entre disciplinas*. Madrid: UNED, 1501-1508.

García-Miguel, J. M., L. Costas y S. Martínez (2005). "Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE". En Wotjak, G. y J, Cuartero Otal (eds.) *Entre semántica léxica, teoría del léxico y sintaxis*. Frankfurt am Main: Peter Lang, 373-384.

Pinuer, C. (2005). “Relieve sintáctico en el español escrito de Chile: Las construcciones ecuacionales y ecuandicionales. *Revista Signos, 38*(*57*), 75-88.

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

Subirats-Rüggeberg, C. y M. R. L. Petruck (2003). “Surprise: Spanish FrameNet!” En E. Hajicova, A. Kotesovcova y J. Mirovsky (eds.) *Proceedings of CIL 17. CD-ROM*. Prague: Matfyzpress.

Vázquez, G. y A. Fernández (en prensa). "Ampliación del Banco de Datos de Verbos del español SenSem". En *Actas del III Congreso Internacional de Lexicografía,* 2008.

Xiao, R. y T. McEnery (2004). *Aspect in Mandarin Chinese: A corpus-based study*. Amsterdam: John Benjamins.