AMPLIACIÓN DEL BANCO DE DATOS DE VERBOS DEL ESPAÑOL SENSEM

Gloria Vázquez, Ana Fernández Montraveta[1]
Universidad de Lérida
Escuela Universitaria de Informática de Sabadell.

Resumen
En el marco del proyecto SenSem se ha constituido un corpus del español de aproximadamente un millón de palabras anotado a nivel morfológico, sintáctico y semántico y un léxico verbal que incluye información extraída de dicho corpus. Uno de los aspectos más destacables de dicho recurso es que incluye aspectos lingüísticos innovadores en el ámbito de la semántica, concretamente, la oracional, como la construcción, la aspectualidad y la modalidad. En este artículo nos vamos a centrar en los dos últimos, ya que son en los que se está trabajado actualmente. El reflejo de la información relativa a la aspectualidad y la modalidad es muy útil tanto para llevar a cabo estudios empíricos como también para mejorar los sistemas automáticos del procesamiento del lenguaje que requieren un tratamiento profundo de las lenguas, como los sistemas de extracción de información, las aplicaciones de pregunta y respuesta y los sistemas de traducción automática.
Palabras clave: corpus anotado, aspectualidad, modalidad, léxico

I. INTRODUCCIÓN

SenSem es un banco de datos de verbos del español compuesto por un *léxico verbal* formado por más de 1.000 sentidos verbales (que se corresponden con las diferentes acepciones de 250 verbos) y un *corpus anotado* asociado a dicho léxico. La primera versión[1] de este banco de datos se finalizó en el 2006 y actualmente se está trabajando en su ampliación[2].

El objetivo del proyecto es definir el comportamiento sintáctico-semántico de las entradas verbales a partir de la información extraída del corpus, de aquí que en esta segunda fase del proyecto se esté trabajando en la ampliación del corpus. El uso de esta metodología permite que la descripción léxica sea más amplia y más exhaustiva, y, en cierta medida, más objetiva que la que se lleva a cabo solamente a partir de la introspección.

Otro valor añadido del recurso creado es que incluye aspectos lingüísticos innovadores en el ámbito de la semántica, concretamente, la semántica oracional, por lo que se refiere a tres aspectos: la construcción, la aspectualidad y la modalidad.[3] En este artículo nos vamos a centrar

[1] **Mailing address**: Gloria Vázquez. Universitat de Lleida. Pl. Víctor Siurana, 1. 25003 Lleida. gvazquez@dal.udl.cat Ana Fernández Montraveta. Universitat Autònoma de Barcelona
Escola Universitària d'Informàtica de Sabadell. Emprius, 2. 08202 Sabadell. ana.fernandez@uab.es

en los dos últimos, ya que aquellos en los que se está trabajado actualmente en la ampliación del proyecto.

El reflejo de la información relativa a la aspectualidad y la modalidad que subyace en las oraciones es muy útil tanto para llevar a cabo estudios empíricos sobre estas cuestiones como también para mejorar los sistemas automáticos del procesamiento del lenguaje (PLN) que requieren un procesamiento profundo de las lenguas, como los sistemas de extracción de información y las aplicaciones de pregunta y respuesta,. Así, por lo que se refiere a la modalidad, "the inferences derivable from factual events are obviously different from those judged as possible or non-existent" (Saurí et al. 2006). En cuanto a la información aspectual, en la medida en que se disponga de corpus en otras lenguas con el mismo tipo de anotación, se van a poder realizar estudios contrastivos muy valiosos para mejorar los sistemas de traducción automática (Murata et al. 2006).

En el terreno de la anotación de corpus, es habitual que en los proyectos en que se está trabajando en el campo de la aspectualidad y la temporalidad se lleve a cabo también la anotación de aspectos relativos a la modalidad, ya que presentan conexiones entre sí. Por ejemplo, los tiempos asociados con el modo indicativo no expresan claramente un significado de modalidad pero aquellos relacionados con el modo subjuntivo o imperativo sí.

Por ejemplo, existe un corpus para el inglés denominado TimeBank, formado por 300 documentos, mayormente noticias periodísticas (Pustejovsky et al. 2003), en el cual se utiliza el lenguaje TimeML (Saurí et al. 2006) para anotar la información aspectual, temporal y la relativa a la modalidad. Así, en dicho proyecto se ha llevado a cabo la anotación de información aspectual extraída del tiempo conjugado o de determinados adverbios, así como otro tipo de información sobre aspecto y modalidad que subyace la relación que se establece entre eventos expresados con estructuras sintácticas predefinidas. Por ejemplo, en el caso de *begin*+V, se etiqueta la acción como incipiente desde el punto de vista aspectual, y, en el caso de *regret* +O, se marca O como modalidad factiva. También se tienen en cuenta otras estructuras sintácticas más genéricas, como las oraciones relativas o subordinadas temporales, cuya modalidad es también anotada como factiva.

Por otro lado, se ha constituido también un corpus japonés-inglés formado por 39.660 frases bilingües que ha sido anotado también en relación a la expresión del tiempo, el aspecto y la modalidad (Murata et al. 2005 y 2006). Para ello se ha partido de la información que aportan 12 auxiliares en inglés (*be able to, used to, will*, entre otros), los tiempos presente y pasado, los tiempos perfectivos e imperfectivos y los progresivos y no progresivos, así como el modo imperativo. Por lo que se refiere al japonés, básicamente la información referente al tiempo, el aspecto y la modalidad, también se expresa en el verbo (al final de éste) y en algunos adverbios de la oración.

Consideramos que desde el proyecto SenSem se va a contribuir de forma especial en el campo de la anotación semántica de corpus. En primer lugar, a nuestro conocimiento, no existe

en español ningún corpus anotado con información referente a la aspectualidad y pocos en relación a la modalidad[4].

En segundo lugar, en el caso de la anotación aspectual, en este proyecto se aplica un sistema de anotación innovador, en la línea de la propuesta de Xiao & McEnery 2004. Dicho sistema es modular y 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). Con el fin de dejar constancia de las transformaciones que se van sucediendo aspectualmente hasta configurar el aspecto de la oración, se mantienen en el corpus las anotaciones de cada nivel (verbo, SV y oración).

Nuestra aportación desde el proyecto SenSem en relación a la modalidad es, por el momento, muy básica, ya que en la actualidad la anotación que estamos realizando implica estrictamente la diferenciación entre las oraciones asertivas y no asertivas, con polaridad negativa o positiva. No obstante, teniendo en cuenta el panorama de la lingüística de corpus en este ámbito, creemos que el resultado obtenido no será menospreciable.

En este artículo vamos a presentar en detalle en qué consiste la ampliación del corpus, tanto por lo que se refiere al tamaño de éste (ap. 1) como al nuevo tipo de anotación (ap. 2). Asimismo, también se indica la futura remodelación del léxico verbal a partir de la nueva información anotada (ap. 3). Por último, se presentan las conclusiones.

II. LA AMPLIACIÓN DEL TAMAÑO DEL CORPUS

El corpus SenSem está formado por unos 13 millones de palabras. Dicho corpus se constituyó a partir de textos periodísticos durante la primera fase del proyecto. Los motivos por los cuales se escogieron textos de este registro fueron varios:

1) Los textos periodísticos incluyen temática muy variada (economía, deportes, cultura, política, etc.), lo cual previsiblemente permitiría ejemplificar usos léxicos distintos.
2) El lenguaje periodístico *per se* es de formalidad neutra, aunque también puede incorporar textos de formalidad más elevada, e incluso es esperable algún uso perteneciente a un registro informal, lo cual también aporta variabilidad al corpus resultante por lo que se refiere a los niveles de lenguaje;
3) La prensa incluye también textos de ficción, lo cual implica, por un lado, que se amplía también la variedad de usos léxicos, sobre todo de tipo metafórico, y por otro lado, que previsiblemente se ejemplifica lenguaje coloquial;
4) Si uno de los usos del léxico SenSem es el PLN, es más útil obtener datos provenientes de textos de registro neutro, como el periodístico, porque los textos que son procesados automáticamente pertenecen mayormente a dicho registro y no al coloquial.
5) La obtención de textos periodísticos en formato electrónico es una empresa fácil.[5] Así, aunque hemos sido siempre conscientes de que la validez de los datos obtenidos a partir del

estudio de corpus aumenta en función del equilibrio del corpus de partida, también había que sopesar el tiempo invertido en la obtención de textos, ya que nuestro objetivo principal era la anotación sintáctico-semántica de las oraciones.

El corpus fue utilizado en primer lugar para extraer los 250 verbos más frecuentes del español con el fin de llevar a cabo la anotación de las oraciones que contenían estos verbos. Una vez detectados, se elaboró un diccionario con las diferentes acepciones de cada verbo llegando a un total de 1.140 entradas (v. ap. 3). Con el uso de dicho corpus se ejemplificaron el 58% de los sentidos del léxico elegido, es decir, se pudieron completar con datos del corpus 661 entradas.

Aunque no interpretamos que sea un mal resultado, nuestro objetivo en esta segunda fase del proyecto es aumentar la cobertura del léxico, por lo que nos planteamos conseguir ejemplos de sentidos que quedaban por cubrir y, a la vez, aumentar el número de oraciones para algunos sentidos cuya representación en el corpus era menor de 10 oraciones. Para ello hemos optado por aumentar el corpus con textos pertenecientes íntegramente al registro literario, donde previsiblemente se localizarían acepciones distintas, ya que, aunque en los textos periodísticos incluyen, como ya se ha avanzado, fragmentos pertenecientes a la no ficción, el porcentaje de éstos no es muy elevado y el género literario en el que se adscriben es poco variado. Por otro lado, como nuestro fin no es únicamente el campo del PLN sino también el de la descripción lingüística, consideramos que la presencia de lenguaje informal debía también ser tenida en cuenta.

Hasta el momento, en la segunda fase del proyecto, dedicada a esta ampliación, se han recopilado los textos literarios a través de Internet. Cabe decir que la tarea de búsqueda de dichos textos no ha sido del todo fácil, ya que aunque existen diversos sitios web desde donde se puede acceder a fragmentos de obras literarias, no tantos incluyen textos de las características que buscábamos: extractos de obras de autores de origen español, del siglo XX o XXI y pertenecientes al género de novela o ensayo. En la tabla 1 se puede consultar la nómina de autores y obras, con la fecha de publicación y el número de palabras utilizado en cada caso.

Tabla 1. Obras que componen el subcorpus literario de SenSem

Año de publicación	Autor	Título	Nº palabras
1902	Vicente Blasco Ibáñez	Cañas y barro	74.990
1914	Miguel de Unamuno	Niebla	54.000
1940	Ortega y Gasset	Creer y pensar	5000
1977	Alonso Zamora	Sin levantar cabeza	3315
1995	Enrique Cerdán	Los ahorcados del cuarto menguante	13.456
1996	Arturo Pérez-Reverte	Capitán Alatriste	5000
2000	Rafael López Rivera	El don	39.000

Como puede observarse, en esta segunda fase se han añadido aproximadamente 200.000 palabras más. Aunque la proporción de ambos registros puede parecer excesivamente

descompensada, ya que la presencia de lenguaje literario queda relegada al 1,5%, a la práctica consideramos que el resultado es aceptable. Así pues, ya en la primera fase del proyecto, los textos que fueron procesados y anotados no fueron todos los que forman el corpus de 13 millones de palabras, sino que para llevar a cabo la anotación se creó un subcorpus a partir del anterior formado por un total de 25.000 frases (unas 700.000 palabras). Ahora, a este número de oraciones se añadirán 5.000 (unas 150.000 palabras) pertenecientes al registro literario, por lo que éste estará representado en un 16,7% y el periodístico en un 83,3%. Cabe decir, además, que la extracción de las oraciones se ha realizado de forma aleatoria, pero se ha fijado un número de 120 oraciones para cada verbo, 100 obtenidas del conjunto de textos de origen periodístico y 20 del grupo de obras literarias.

III. LA AMPLIACIÓN DEL TIPO DE ANOTACIÓN

En la primera fase del proyecto (Castellón et al. 2006) se pretendió cubrir los aspectos básicos de la anotación de la sintaxis y la semántica de las oraciones, como las categorías sintagmáticas, las funciones sintácticas, las funciones semánticas (roles), la distinción entre argumentos y adjuntos y la semántica de la construcción. Además, se anotó también la semántica léxica del verbo, es decir, se desambiguó el sentido de los verbos núcleos de las oraciones anotadas indicando la acepción de éstos utilizando como referencia lexicográfica la creada en el propio proyecto, elaborada a partir de diferentes fuentes ya existentes (v. ap. 3).

En la segunda etapa del proyecto, los objetivos marcados relativos a la ampliación del tipo de anotación son los siguientes:

1) incorporar la anotación morfológica a nivel de palabra en todo el corpus seleccionado[6]
2) incorporar la anotación aspectual a nivel sintagmático y oracional en todo el corpus seleccionado[7]
3) incorporar la anotación de la modalidad en todo el corpus seleccionado
4) anotar la segunda parte del corpus seleccionado (v. ap. 1) con la información sintáctico-semántica utilizada durante la primera fase del proyecto

El primer objetivo puede darse ya casi por concluido. La anotación se ha realizado de forma automática con la herramienta FreeLing (Atserias et al. 2006) y falta llevar a cabo la revisión de errores. Además de algunos problemas relacionados con la segmentación y puntuación de las oraciones, que ya están en fase de solución, el número de errores previstos por problemas de precisión de la herramienta se prevé que sea bajo (según sus creadores la precisión de esta herramienta es de un 97%).

Por lo que se refiere al segundo objetivo, es uno de los hitos más importantes del proyecto por lo que supone de innovador, y consiste en la anotación de las oraciones con información aspectual. Hasta el momento se han anotado 10.000 frases y se prevé que a finales del 2008 ya se habrá anotado la mitad del corpus y a finales del 2009 su totalidad.

Para llevar a cabo este tipo de anotación se tiene en cuenta, además del tipo eventual léxico (v. ap. 3), la información de tipo aspectual que pueden aportar los diferentes elementos que comparecen en una determinada oración, ya sean auxiliares, desinencias verbales o determinados complementos o adjuntos.

Así pues, se parte de la base de que la información aspectual que expresa una oración proviene de diferentes fuentes y que se obtiene de forma composicional, teniendo en cuenta tanto el llamado aspecto situacional (*situational aspect*) como el aspecto relacionado con el punto de vista (*pointview aspect*), que, como es sabido, son independientes pero deben considerarse conjuntamente a la hora de describir el aspecto de una oración (Smith 1997, Xiao & McEnery 2004).

En algunos casos, la presencia de un elemento puede cambiar el valor aspectual relativo a la presencia o ausencia de límites en la acción, valor que aporta inicialmente el verbo léxicamente (*Aktionsart*). En otros casos, se añaden valores aspectuales, como el de la (im)perfectividad o la habitualidad. Así pues, la información aspectual asociada a una oración podrá contener más de una etiqueta.

Por ejemplo, el uso de determinados tiempos verbales, como los compuestos o el pretérito perfecto, añaden un valor perfectivo (1), mientras que otros, como el presente o pretérito imperfecto, aportan un significado imperfectivo (2). Este tipo de anotación se ha podido automatizar en algunas ocasiones (cuando el valor relativo a la (im)perfectividad que aporta un determinado tiempo verbal siempre es el mismo)[8], ya que el corpus está anotado con información morfosintáctica que incluye la codificación de los tiempos verbales.

(1) No se practicó ninguna detención, aunque *a cuatro personas se les abrió diligencias por amenazas a la autoridad*.

(2) Después de una etapa en la que el Gobierno del PP ha vivido de espaldas a los gobiernos autonómicos y en confrontación con ellos, *se abre una nueva etapa* que es la de la España plural y la de la cooperación franca y leal entre los distintos niveles de la Administración.

Por otro lado, algunos tiempos típicamente imperfectivos pueden conllevar una lectura habitual, a menudo combinados, aunque no necesariamente, con elementos que expresan frecuencia, ya sean adverbios como *siempre, a menudo*, etc., o auxiliares como *soler*, entre otros. La lectura habitual (3) se aplica a oraciones en las que participan predicados que léxicamente expresan procesos o eventos e implica la ausencia de unas coordenados espacio-temporales concretas puesto que dichas acciones se presentan como reiteradas en el tiempo.

(3) (…) Eduardo Bautista, considera que el consumidor español "se está espabilando mucho" porque *se abre a todos los productos* y sabe cada vez más lo que quiere.

Asimismo, la cuantificación de determinados argumentos puede incidir en la interpretación de si el estado de cosas descrito en la oración es limitado (*bounded*) o no (*unbounded*). Así, el uso de un complemento no cuantificado implica la ausencia de límite en la acción, por lo que se considerará que la oración expresa un proceso (4), independientemente de si el verbo léxicamente es considerado un evento (v. ap. 3).

(4) Corrió, *abrió espacios*, bajó a recibir, peleó sin importarle la entidad del rival y marcó sus primeros dos goles, el tercero y el sexto del Barça.

Respecto al tercer objetivo, el tipo de información referente a la modalidad que se pretende anotar es, inicialmente, muy básica, pero no por ello menospreciable, teniendo en cuenta la poca presencia de corpus anotados en español con información de este tipo. Así, en la actualidad se están marcando las oraciones de polaridad negativa y también las de modalidad no asertiva.

En algunas ocasiones ha sido posible realizar una preasignación automática de estos valores. Por lo que se refiere a la polaridad, ha sido sencillo detectar la aparición de palabras que expresan negatividad. En el caso de la no asertividad, nos hemos basado en la presencia de determinados símbolos (interrogación, admiración), determinados tiempos verbales (subjuntivo, condicional) o determinadas perífrasis (*poder* + infinitivo, *tener que* + infinitivo, etc.).

Respecto a las oraciones anotadas como no asertivas, se está trabajando en su subclasificación y se prevé que al final del proyecto se dispondrá de una anotación más pormenorizada del tipo de modalidad expresada en ellas. Así, por ejemplo, como en el caso de las perífrasis aspectuales, actualmente, ya se dispone de una propuesta de clasificación de los distintos tipos de significado que expresan las perífrasis modales (Topor et al. 2006).

Por último, se prevé que el cuarto objetivo se llevará a cabo el último año de ejecución del proyecto, es decir, en 2009, ya que inicialmente hemos priorizado los aspectos relativos a la anotación más innovadores. Con la consecución de esta tarea, el corpus SenSem quedará homogeneizado en cuanto al tipo de anotación incluida.

IV. EL LÉXICO VERBAL SENSEM Y SU AMPLIACIÓN

Como ya se ha avanzado, el léxico verbal SenSem (Fernández et al. 2006) está formado por 1.140 entradas verbales, entendiendo que cada entrada es un sentido distinto. Estos sentidos se corresponden con las diferentes acepciones de los 250 verbos más frecuentes del español excluyendo los usos arcaicos o restringidos. El léxico que actualmente se puede consultar en Internet (http://grial.uab.es/lexico_es_SenSem), se corresponde al recurso creado en la primera fase del proyecto. Se prevé que a finales del 2009 estará disponible la segunda versión, que contendrá las novedades que se describen en este apartado.

Vamos a dividir la descripción del léxico en dos subapartados, diferenciando aquella información que es común a todas las entradas (3.1) de la que sólo contienen un subgrupo de éstas (3.2).

IV.1 Información asociada a todos los sentidos de SenSem

Todas las entradas verbales, independientemente de si se dispone o no de ejemplos de dichos sentidos en el corpus, han sido descritas con una mínima descripción sintáctico-semántica: definición lexicográfica, sinónimos, clase eventual léxica (*Aktionsart*) y roles semánticos.

Las *definiciones* son originales, aunque, obviamente, se ha partido de diversas obras lexicográficas. Las fuentes lexicográficas que se han utilizado han sido principalmente el *Diccionario de la Real Academia de la Lengua Española* y el *Diccionario Salamanca de la Lengua Española*. En la construcción de este léxico, no se tuvieron en cuenta los usos arcaicos o muy restringidos.

Los *sinónimos* se incorporaron de forma asistemática durante la primera fase del proyecto y utilizando la intuición del lingüista. En esta segunda fase se ha incorporado para cada sentido verbal su correspondiente con la red semántica WordNet (Fellbaum 1998) en las versiones 1.5, 1.6 y 2.1. Próximamente se incorporará el enlace con los *synsets* correspondientes a la última versión de WordNet (3.0). Recordamos que en WordNet la unidad del léxico es el sentido, por lo que cada entrada incorpora diversos ítems sinónimos (*synset*). Así pues, la inclusión del identificador de los *synsets* correspondientes en nuestro léxico supone un incremento substancial de los sinónimos de los sentidos de SenSem. Por otro lado, otra ventaja es que el léxico SenSem amplía sus utilidades al incorporar esta información, ya que WordNet es un referente mundial en el campo de la lexicografía computacional y el de la anotación y desambiguación semántica dentro del PLN.

En cuanto a la *clase eventual*, consideramos que esta información es clave en la caracterización semántica de un predicado y no necesariamente se desprende de la definición lexicográfica, por lo que debe codificarse a parte. En este sentido, la incorporación de la clase eventual en la entrada léxica del recurso creado es uno de los valores añadidos que lo caracterizan y lo distinguen en relación a otros léxicos. Para la descripción de esta información se han utilizado 4 tipos eventuales: estado (185 sentidos, *tener*), proceso (353 sentidos, *correr*), proceso/evento (56 sentidos, *analizar*) y evento (546 sentidos, *romper*). La novedad de esta clasificación es que hemos incorporado el tipo *proceso/evento* porque consideramos que existen predicados léxicamente no es posible clasificarlos ni como procesos ni como eventos.

Respecto a los *roles semánticos*, se incluye en cada entrada léxica el listado de las etiquetas que definen la relación semántica entre los argumentos y el verbo. El listado utilizado como punto de partida es la adaptación del conjunto de etiquetas consensuado por tres equipos investigadores en anteriores proyectos (Fernández et al. 2002).

IV.2 Información asociada a los sentidos con representación el corpus

De las 1.140 entradas mencionadas, actualmente un 58% (concretamente, 662 sentidos verbales) incluyen además otra información extraída del corpus anotado en la primera etapa del proyecto: las *estructuras de subcategorización,* los *significados oracionales de cada construcción* y la *frecuencia del sentido* en el corpus. A diferencia de la información de nivel básico que contienen todas las entradas, esta información extraída del corpus sólo la incorporan algunas entradas porque sólo se ha podido incluir para aquellos sentidos documentados en dicho

corpus. Actualmente, como ya se ha avanzado (v. ap. 1), se está trabajando en la ampliación de éste con el fin de aumentar la cobertura actual de descripción de sentidos.

Esta información, por tanto, a diferencia de la descrita anteriormente, no proviene ni de la introspección lingüística ni de ninguna otra fuente lexicográfica, sino que depende directamente de los datos obtenidos del corpus. No hay que olvidar, sin embargo, que los datos codificados dependen de la interpretación que ha hecho el anotador, sobre todo en lo referente a aquellos problemas todavía pendientes en la teoría lingüística, como la interpretación de algunas funciones sintácticas (objeto preposicional, complemento circunstancial), y especialmente en lo referente a la semántica. Es fácil que en relación a este tipo de anotación pueda haber divergencias entre diferentes lingüistas. Así pues, se ha realizado un proceso de revisión de las frases una vez anotadas con el fin de, en primer lugar, localizar posibles errores humanos de la anotación y, en segundo lugar, consensuar aquellos aspectos más problemáticos desde el punto de vista lingüístico (Vázquez et al. 2006).

Respecto a las *estructuras de subcategorización* en cada entrada verbal, se incluye el listado de las diferentes configuraciones argumentales que se han reflejado en el corpus periodístico ya anotado ejemplificadas con todas las oraciones extraídas de éste que responden a dichos patrones. Estas configuraciones incluyen el tipo de sintagmas, la función sintáctica y semántica de cada uno y la frecuencia de cada estructura argumental en relación al total de oraciones identificadas para el sentido verbal que se está describiendo.

Esta información se puede consultar en dos niveles de especificidad: en el primer nivel, el que presenta las estructuras de subcategorización de forma más generalizada, se omite la función semántica, no se diferencian las estructuras coincidentes si el orden de aparición de los constituyentes no es coincidente y se agrupan algunos sintagmas en la misma categoría (por ejemplo, el SPRON –sintagma formado por un pronombre- se encubre bajo la categoría de SN y, en el caso del sujeto, también se incluye bajo esta etiqueta el caso de los sujetos elípticos). En el segundo nivel, se subespecifican los esquemas del primer nivel aportando información relativa a los roles semánticos, respetando el orden de aparición de los sintagmas e indicando el tipo específico de categoría y si hay sujeto elíptico.

Respecto a la *semántica oracional de la construcción*, en cada entrada se da cuenta de los diferentes significados construccionales en que aparece cada verbo. Se han anotado, por tanto, las oraciones anticausativas, pasivas (diferenciando si provienen de un agente o no, si son sintácticas o pronominales y si hay concordancia con el sujeto o no), impersonales y causativas indirectas. Se deduce que las oraciones que no tienen ninguna de estas etiquetas son oraciones que presentan la estructura argumental en su forma menos marcada, es decir, con sujeto agente, causa o experimentador. Dentro de esta misma categoría que hemos denominado semántica construccional, aunque a otro nivel, se han anotado también las oraciones recíprocas, reflexivas y de dativo[9].

En cuanto a *la frecuencia del sentido*, cabe recordar que para cada uno de los 250 verbos analizados se extrajeron en la primera fase del proyecto 100 frases, por lo que la frecuencia del

sentido se ha de interpretar en base a esta cifra, ya que todavía no se ha llevado a cabo la anotación del subcorpus literario.

La ampliación del corpus con oraciones provenientes del registro literario nos permitirá, previsiblemente, dos mejoras en el léxico que citamos a continuación: por un lado, añadir nuevas estructuras de subcategorización y nuevos significados construccionales al listado asociado a cada verbo y, por otro, redistribuir el peso de cada estructura en términos de frecuencia. En definitiva, con la consecución de estas dos tareas, se conseguirá mejorar la descripción del comportamiento sintáctico-semántico de los verbos objeto de estudio.

Respecto a la información aspectual anotada en el corpus, al finalizar el proyecto se prevé ampliar la base de datos léxica actual, de forma que se podrá observar como un predicado que léxicamente se adscribe a una determinada clase aspectual puede variar su aspectualidad según el contexto.

En cuanto a la información relativa a la modalidad codificada en el corpus, en tanto que ésta no está relacionada con las características léxicas de los predicados, no se incorporará en la base de datos lexicográfica.

V. CONCLUSIONES

En este artículo presentamos la ampliación que se está llevando a cabo del banco de datos verbales del español SenSem, que contiene un recurso léxico y otro textual interconectados e interdependientes. En la primera fase, el corpus se anotó con un tipo de información sintáctico-semántica, que, una vez procesada, se utilizó para describir el comportamiento de los verbos. En la segunda etapa, se está ampliando el número de palabras del corpus y el tipo de anotación. Dicha ampliación se está llevando a cabo en diversas etapas.

En primer lugar, se ha ampliado el corpus periodístico inicial con la inclusión de textos provenientes del registro literario del español actual y se está llevando a cabo la anotación de la semántica aspectual a nivel de oración, así como de la modalidad. También se ha mejorado el léxico actual al incorporar los identificadores de la red semántica WordNet para cada entrada. En segundo lugar, se etiquetará el corpus literario con la misma información que contiene el corpus periodístico. En tercer lugar, se introducirá de nuevo en el léxico la información aportada a partir de la nueva anotación efectuada.

Esta última tarea implica, por un lado, la redefinición de las entradas ya creadas durante la primera etapa del proyecto (nuevos ejemplos asociados, posibles nuevas construcciones semánticas y configuraciones sintácticas y la redistribución de la frecuencia de ambas). Por otro, se describirán entradas verbales que habían quedado sin ejemplificar en esa primera etapa debido a la ausencia de representatividad de dichos sentidos en el corpus periodístico. Por último, se incluirá nueva información sobre la semántica aspectual en todas las entradas verbales. Por lo que se refiere a la modalidad, consideramos que dicha información no mantiene una relación clara con el tipo de predicado, por lo que no se prevé incorporarla al léxico.

Respecto a la información aspectual, hasta el momento, sólo se había codificado el tipo eventual del sentido verbal. Una vez finalizado el proyecto, se dispondrá de una descripción pormenorizada de los posibles significados aspectuales que adquiere cada verbo en las oraciones y la frecuencia de esos usos en cada caso. Dichos significados aspectuales aportan información aspectual que no proviene de la pieza léxica verbal y que pueden implicar, incluso, la modificación del tipo eventivo de ésta. En el caso de las lenguas románicas las desinencias verbales aportan información referente al tiempo y el aspecto. Así pues, los corpus anotados morfológicamente, que hoy en día son bastante numerosos y existen para un número amplio de lenguas, son un paso previo y necesario para la anotación aspectual. Pero, tanto en estas lenguas como en idiomas pertenecientes a otra tipología, existen otras marcas aspectuales más allá de las contenidas a nivel morfológico y que pueden modificar su contenido. Sin embargo, a nuestro conocimiento, la anotación de la aspectualidad oracional de forma global es un campo todavía por explorar. En este proyecto pretendemos aportar resultados en este campo utilizando una metodología de base empírica.

Al final de proyecto, a finales del 2009, el banco de datos SenSem, que incluye el léxico y el corpus anotado, contendrá información muy novedosa en el campo de la semántica en la descripción del español.

REFERENCIAS BIBLIOGRÁFICAS

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. & Padró, M. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. http://www.lsi.upc.edu/~nlp/papers/atserias06.pdf. Acceso: 13.03.09.

Castellón, I., Fernández, A., Vázquez, G., Alonso, L. & Capilla, J. A. (2006). The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 355-359.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Londres: The MIT Press, Cambridge.

Fernández, A., Saint-Dizier, P., Vázquez, G., Benamara, F. & Kamel, M. (2002). The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexicons. En *Proceedings of the Language Engineering Conference*, 89-98.

Fernández, A., Vázquez, G. & Castellón, I. (2006). SenSem: a Databank for Spanish Verbs. En *Proceedings of the X Ibero-American Workshop on Artificial Intelligence, IBERAMIA*. Ribeirão Preto, Brasil.

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

Murata, M., Utiyama, M., Uchimoto, K., Ma, Q. & Isahara, H. (2005). Correction of Errors in a Modality Corpus Used for Machine Translation Using Machine-learning. *ACM Transactions on Asian Language Information Processing*, 4:1, 18-37.

Murata, M., Ma, Q., Uchimoto, K., Kanamaru, T. & Isahara, H. (2006). Japanese-to-English translations of tense, aspect, and modality using machine-learning methods and comparison with machine-translation systems on market. En *Language Resources and Evaluation*, 40, 233-242.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L. & Lazo, M. (2003). The TIMEBANK Corpus. En *Proceedings of Corpus Linguistics*, p. 647-656.

Topor, M., Fernández, A. & Vázquez, G. (2006). Perífrasis verbales del español y rumano: correspondencias y vacíos léxicos. *Studies in Contrastive Linguistics.* Santiago de Compostela, 1061-1067.

Vázquez, G., Alonso, L., Capilla, J. A., Castellón, I. & Fernández, A. (2006). SenSem: sentidos verbales, semántica oracional y anotación de corpus. *Procesamiento del Lenguaje Natural*, 37, 113-120.

Saurí, R., Verhagen, M. & Pustejovsky, J. (2006). Annotating and Recognizing Event Modality in Text. En *FLAIRS Conference,* p. 33-339.

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

Subirats-Rüggeberg, C. & Petruck, M. R. L. (2003). Surprise: Spanish FrameNet!. En *Proceedings of the International Congress of Linguists* (Workshop on Frame Semantics), Praga. http://www.icsi.berkeley.edu/pubs/ai/subirats-petruck.pdf. Acceso: 13.03.09.

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

NOTAS

[1] Ministerio de Ciencia y Tecnología (BFF2003-06456).
[2] Ministerio de Educación y Ciencia (HUM2007-65267).
[3] En la primera versión del banco de datos, se anotó, además de los roles sintácticos y los roles semánticos de los argumentos oracionales, la semántica oracional relacionada con la construcción y que se refiere al significado que aporta la expresión de determinados argumentos, su función semántica y su posible focalización en la estructura funcional de la oración. Dicha información se incluye también en algunos de los proyectos, como Adesse (García de Miguel & Comesaña, 2004).

[4] Por lo que se refiere a la anotación de la modalidad en corpus, sí que existen corpus anotados de la lengua oral, como el *Corpus Oral de Referencia del Español Contemporáneo o PRESEEA*, pero no de la lengua escrita. Por lo que se refiere a la lengua escrita, tenemos constancia que el Corpus ARTHUS está anotado también a este nivel.

[5] Los textos que forman el corpus SenSem pertenecen a *El Periódico de Catalunya*, fundamentalmente, entidad a la cual agradecemos su colaboración.

[6] Cuando decimos "todo el corpus seleccionado" estamos refiriéndonos al subcorpus de 850.000 palabras, que incluye tanto las oraciones provenientes del registro periodístico como a las extraídas de las obras literarias.

[7] Como se ha avanzado, en la primera fase del corpus, sólo se incluyó información aspectual a nivel oracional referente a la característica de la habitualidad.

[8] Dicha automatización no se ha llevado a cabo, por ejemplo, con el tiempo futuro, porque no está claro que valor relativo a la (im)perfectividad confiere.

[9] Las oraciones etiquetadas como construcciones de dativo incluyen diversos tipos de dativos que no son argumentales (dativos de interés, posesivos, etc.).