RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 50 (1), I Sem. 2012, pp. 13-38.

CL ISSN 0033 - 698X

CONSTITUCIÓN DE UN CORPUS DE SEMÁNTICA VERBAL DEL ESPAÑOL. METODOLOGÍA DE ANOTACIÓN DE NÚCLEOS ARGUMENTALES[*]

BUILDING A VERBAL SEMANTICS CORPUS OF SPANISH. METHODOLOGY FOR LABELLING PHRASE HEADS

IRENE CASTELLÓN
Universitat de Barcelona. Barcelona, España
icastellon@ub.edu

SALVADOR CLIMENT
Universitat Oberta de Catalunya. Barcelona, España
scliment@uoc.edu

MARTA COLL-FLORIT
Universitat Oberta de Catalunya. Barcelona, España
mcollfl@uoc.edu

MARINA LLOBERES
Universitat Oberta de Catalunya. Barcelona, España
mlloberes@uoc.edu

GERMAN RIGAU
Euskal Herriko Unibersitatea. País Vasco, España
german.rigau@ehu.es

RESUMEN

El presente artículo detalla la metodología y el desarrollo del proyecto de desambiguación semántica de los núcleos argumentales de SenSem, un corpus equilibrado constituido por 100 oraciones para cada uno de los 250 verbos más frecuentes del español. El resultado, unido a desarrollos anteriores del proyecto, es un corpus ricamente etiquetado con información sintáctica y semántica, conectado a una base de datos que recoge la información pertinente para cada sentido verbal, por lo que el recurso resultante es adecuado para estudios empíricos centrados en el verbo. Como resultado del proceso se presenta asimismo un

[*] Esta investigación se ha llevado a cabo gracias a los proyectos: FFI2008-02579-E/FILO y TIN2009-14715-C04-03 del Ministerio de Ciencia e Innovación de España.

análisis crítico de WordNet 1.6 del español como recurso de anotación lexico-semántica de corpus y una guía de criterios de anotación, ambos de utilidad para tareas similares de etiquetado con WordNet.

Palabras clave: Lingüística de corpus, anotación semántica, WordNet, corpus SenSem.

ABSTRACT

The SenSem Corpus and Database (Alonso, Capilla, Castellón, Fernández y Vázquez, 2007) consists of a verb-oriented balanced corpus of Spanish linked to syntactic and semantic database of predicates and sentences. The corpus consists of 100 sentences for each of the 250 more frequent verbs of Spanish. It is labelled with rich semantic and syntactic information which is structured in the database according to verb senses, thus providing an invaluable resource for verb-focused linguistic empirical research. In this paper we present the process and methodology adopted for labelling nominal argument-structure heads with WordNet sense-id's. As a by-product, both a critical assessment of Spanish WordNet 1.6 as a resource for semantic labelling and a labelling criteria guide are discussed and provided so that they might be useful in future similar research.

Keywords: Corpus linguistics, semantic annotation, WordNet, SenSem corpus.

Recibido: 29.06.2011. Aceptado: 23.12.2011.

1. INTRODUCCIÓN

El uso de corpus lingüísticos informatizados para acceder al conocimiento del uso real del lenguaje se inicia en el último cuarto del siglo pasado, de manera simultánea a la popularización de los ordenadores y su progresiva capacidad para procesar textos. Podemos distinguir entre el procesamiento de dicha información por parte de lingüistas –lo que dio origen a la lingüística de corpus en tanto que método para inducir o contrastar empíricamente análisis lingüísticos– y el procesamiento mediante ordenadores. El auge del procesamiento informático de corpus, a caballo del cambio de siglo, impulsó y facilitó la anotación lingüística de los mismos. Obviamente, un corpus enriquecido con información lingüística resulta mucho más útil para la investigación que el puro texto sin etiquetar y cuanto más rica sea dicha anotación, mayor será el conocimiento que se podrá extraer o explotar. El uso de corpus etiquetados es doble (cf. Navarro, 2007: 15 y ss): por una parte, el aprendizaje automático (desarrollo y optimización de algoritmos de aprendizaje a partir de grandes cantidades de ejemplos anotados por especialistas) y, por otra, su uso como referente de la evaluación de los sistemas, en tanto que muestra de análisis correcto anotado por humanos (*gold standard*).

Así, el rápido desarrollo en la última década del procesamiento del lenguaje natural (PLN), y en especial de los métodos de aprendizaje automático, ha incrementado enormemente el valor de los corpus anotados manualmente. Ello ha causado el desarrollo de más y mayores corpus y, también, el aumento de la complejidad de la información con que se anotan. Lógicamente, dicho aumento de la complejidad en la anotación hace más costosa la tarea, pero también más útil el recurso. De la anotación morfológica y la lematización se ha pasado a la anotación con información sintáctica y, más recientemente, a la semántica, siendo esta última la menos desarrollada hasta el momento.

El tratamiento semántico en el PLN es imprescindible para la creación de cualquier aplicación inteligente, como sistemas de pregunta-respuesta, resumen automático, sistemas de diálogo o traducción automática. La extracción de conocimiento semántico de textos anotados es asimismo crucial para el desarrollo de formas más sofisticadas de codificación de la información: ontologías, clasificación o representación de la información, web semántica, etc.

En el terreno de las líneas básicas de investigación en PLN, una de las tareas que avanza con mayor dificultad es la resolución automática de la ambigüedad semántica del léxico (*Word Sense Disambiguation*, WSD) (Agirre y Edmonds, 2007). Por otra parte, el desarrollo de métodos fiables de adquisición automática de preferencias selectivas de los verbos será un recurso enormemente valioso tanto para la WSD como para sistemas de extracción o comprensión automática de la información.

Es importante hacer notar que, paradójicamente, pese a existir un consenso bastante amplio entre los lingüistas, en el sentido de que el verbo es el núcleo y el principal centro estructural de la información lingüística de las oraciones, prácticamente no se han desarrollado anotaciones de corpus centradas en el predicado verbal y en la estructura sintáctica y semántica (temática, funcional) que impone a las construcciones. Para cubrir dicho vacío se está desarrollando la anotación sintáctico-semántica del corpus SenSem, una parte de la cual se presenta en este artículo.

En concreto, se presenta el proceso de anotación semántica de los núcleos sustantivos de los argumentos verbales de un corpus del español, cuyo objetivo básico es la adquisición de preferencias semánticas asociadas a los sentidos de los predicados verbales. Este trabajo se enmarca y es la continuación del proyecto de anotación semántica del corpus SenSem[1] (Alonso, Capilla, Castellón, Fernández y Vázquez 2007; Vázquez y Fernández, 2008), mediante el cual se constituyó un banco de datos compuesto por una base de datos verbal y un corpus de 25.000

[1] Proyecto financiado por el Ministerio de Ciencia y Tecnología de España en las siguientes acciones: BFF2003-06456 (2004/2006); HUM2007-65267 (2007). Disponible en: http://grial.uab.es/fproj.php?id=1&idioma=es.

oraciones correspondientes a los 250 verbos más frecuentes de la lengua española, anotados con información sintáctica y semántica del verbo (sentido) y de la construcción de la que es núcleo (tipos de constituyentes y de construcción, funciones, papeles temáticos, información aspectual). La fase actual[2] de este proyecto de largo alcance se integra a su vez en el proyecto KNOW2[3] centrado en el desarrollo de recursos y sistemas inteligentes de procesamiento del lenguaje.

En consonancia con las estrategias del proyecto KNOW2[4] se ha utilizado para el etiquetado de sustantivos la base de conocimiento WordNet del español en su versión 1.6 (ESPWN1.6), integrada en el *Multilingual Central Repository*[5] (MCR) (Atserias, Villarejo, Rigau, Agirre, Carroll, Magnini y Vossen, 2004), una macrored semántica y ontológica multilingüe. WordNet es, en este momento, el recurso estándar para la anotación léxico-semántica de corpus en el área del procesamiento del lenguaje natural. ESPWN1.6 fue construido básicamente por traducción (Atserias, Climent, Farreres, Rigau y Rodríguez, 2000) y posteriormente validado manualmente, respetando la estructura relacional del WordNet original inglés, desarrollado en la Universidad de Princeton (Fellbaum, 1998).

El objetivo fundamental del trabajo que aquí se presenta es la obtención de un corpus ricamente anotado, orientado a la semántica verbal y su libre distribución entre la comunidad investigadora. Sin embargo, el desarrollo del proceso conlleva tres tareas inseparables, cuya descripción consideramos también de utilidad para la comunidad:

1. El análisis detallado de la adecuación de WordNet (en concreto ESPWN1.6) para la anotación semántica de corpus.
2. El establecimiento de propuestas de solución de los problemas derivados de los casos de inadecuación de WordNet.
3. Como resultado de lo anterior y de otras condicionantes del proceso, la obtención de un conjunto de criterios de anotación, incluyendo instrucciones para anotadores, procedimiento de anotación de casos especiales, soluciones a problemas habituales y, especialmente, criterios para la desambiguación manual de significados.

En el siguiente apartado se presenta el estado de la cuestión en anotación semántica de corpus y, a continuación, la metodología utilizada en la presente investigación (§3), la evaluación de ESPWN1.6 como recurso léxico-semántico en relación a los problemas de etiquetado (§4), el conjunto de soluciones adoptadas

[2] Ministerio de Ciencia e Innovación español. FFI2008-02579-E/FILO (2008/2011).

[3] Ministerio de Ciencia e Innovación español. TIN2009-14715-C04 (Plan Nacional de I+D+i 2008-2011).

[4] Disponible en: http://ixa.si.ehu.es/know2.

[5] Disponible en: http://adimen.si.ehu.es/web/MCR.

para la resolución de dichos problemas más la correspondiente guía de criterios (§5), los resultados obtenidos en el proyecto (§6) y finalmente las conclusiones y la propuesta de trabajo futuro (§7).

2. ANOTACIÓN SEMÁNTICA DE CORPUS

Como ya se ha avanzado, SenSem es un banco de datos del español compuesto por un corpus y una base de datos verbal. El corpus, en su primera versión, consta de 100 oraciones para cada uno de los 250 verbos más frecuentes de la lengua española (Davies, 2002). Las oraciones y el núcleo verbal están etiquetados a nivel sintáctico y semántico.

En la etapa del proyecto que aquí se presenta se ha realizado la anotación semántica de los núcleos nominales de los argumentos de SenSem. Los sentidos léxicos utilizados para dicha anotación son los que constituyen ESPWN1.6, una red lexico-semántica que consta de sentidos agrupados en conjuntos de sinónimos o *synsets* (‘*synonym sets*’) y conectados por relaciones semánticas (por ejemplo, hiponimia y meronimia). Se ha usado además como base de conocimiento de apoyo MCR, el cual integra WordNet con múltiples ontologías de propósito general como la EuroWordNet Top Ontology (TO) (Álvez, Atserias, Carrera, Climent, Laparra, Oliver y Rigau, 2008) y SUMO (Niles and Pease, 2003).

La lengua inglesa fue la primera en desarrollar corpus anotados semánticamente, en proyectos como SemCor (Miller, Chodorow, Landes, Leacock y Thomas, 1994) o FrameNet (Fillmore, Johnson y Petruck, 2003). Sin embargo probablemente sea PropBank (Palmer, Gildea y Kingsbury, 2003) el proyecto más cercano a SenSem, en tanto que afronta directamente la anotación de corpus a nivel de predicados verbales. PropBank es el resultado de añadir anotación sobre predicados y argumentos al corpus anotado sintácticamente Penn Treebank II (Marcus, Santorini y Marcinkiewicz, 1993), procedente del corpus Wall Street Journal. El objetivo principal de PropBank es servir como banco de datos empírico para una teoría semántica neutra.

En cuanto al recurso utilizado para la anotación en este proyecto, WordNet, los desarrollos más relacionados con SenSem son Ontonotes (Hovy, Marcus, Palmer, Ramshaw y Weischedel, 2006), SemCor (Miller *et al.*, 1994) y MultiSemCor (Bentivogli y Pianta, 2005).

Ontonotes tiene por objeto describir un gran corpus compuesto por distintos géneros (noticias, conversaciones telefónicas, weblogs, usenets, televisión, etc.) en tres idiomas (inglés, chino y árabe), con información estructural (sintaxis y estructura predicado-argumento) y semántica (sentidos de palabras vinculados a WordNet). Se basa en recursos doblemente verificados: al validar la anotación sintáctica de Penn Treebank y la estructura predicado-argumento de PropBank.

Su representación semántica incluye la anotación de los sentidos de los sustantivos y verbos.

SemCor es un corpus textual de 700.000 palabras etiquetado morfosintáctica y semánticamente. Los textos que lo constituyen proceden del Brown Corpus (Francis y Kucera, 1964) y han sido etiquetados sintácticamente utilizando el desambiguador (*tagger*) desarrollado por Brill (1995). El etiquetado semántico se realizó manualmente para todos los nombres, verbos, adjetivos y adverbios (en un total de 200.000), los cuales se han asociado a su correspondiente sentido en WordNet.

MultiSemCor (Bentivogli y Pianta, 2005) es un corpus paralelo inglés-italiano basado en SemCor. MultiSemCor cuenta, además de la información propia de SemCor, con enlaces interlingüísticos con los sentidos correspondientes al WordNet italiano. La aplicación al español de la solución tomada en MultiSemCor para el italiano, es decir, la traducción del corpus Semcor al español, proporcionaría un buen recurso general para el WSD, pero no tanto para la distinción de sentidos verbales y sus argumentos y la adquisición de restricciones selectivas, debido a que el número de oraciones asociadas a cada verbo no es equilibrado.

En cuanto a recursos de este tipo para el español, los principales son Adesse (García-Miguel y Albertuz, 2005) y AnCora (Taulé, Martí y Recasens, 2008). Ambos recursos, al igual que SenSem constan de un corpus y una base de datos verbal interconectados.

Adesse es la variante semántica del corpus Arthus (Rojo, 2001) y constituye una versión ampliada de la Base de Datos Sintácticos del Español Actual[6] (BDS) desarrollada por la Universidad de Santiago de Compostela que contiene información sintáctico-semántica sobre las cláusulas verbales. Para cada ejemplo se han anotado, en el nivel sintáctico-semántico, la función, la categoría y una anotación en rasgos semánticos básica (*animado* y *no animado*). A nivel semántico se han anotado el rol semántico, la acepción verbal (clasificación propia) y la clase semántica. A partir de esta información se ha construido la base de datos consignando las diferentes construcciones en las que participa un predicado verbal (alternancias y frecuencias). Adesse realiza una identificación de los sentidos o acepciones del verbo, aunque se indica que se realiza a un nivel muy general que se irá especificando progresivamente, y no se ha conectado con los sentidos de WordNet.

AnCora consta de dos corpus, uno del catalán y otro del español, con diferentes niveles de anotación sintáctico-semántica: categoría morfológica, constituyentes, funciones sintácticas, estructura argumental y papeles temáticos, clase semántica verbal, sentidos de WordNet nominales y entidades referenciales. Como resultado del proceso de anotación se construyen léxicos verbales para las dos lenguas que incluyen información sobre la clase semántica del verbo y la subcategorización

[6] Disponible en: http://www.bds.usc.es/.

sintáctica, la estructura argumental y los roles temáticos.

Las diferencias fundamentales entre SenSem y los corpus AnCora y Adesse se describen a continuación. En primer lugar, estos últimos no se han construido de manera equilibrada por lo que respecta a las ocurrencias verbales, por lo que los ejemplos no siempre son representativos de los usos de los predicados –por ejemplo, existen predicados verbales con un único ejemplo– ya que su objetivo es más la anotación de múltiples niveles lingüísticos que la adquisición de información léxica verbal. SenSem, en cambio, está diseñado y construido para la adquisición y anotación de datos lingüísticos asociados a la unidad verbal, por lo que los ejemplos asociados a los verbos están equilibrados (100 ejemplos por cada verbo). Otro aspecto que los diferencia es que AnCora y Adesse están basados en la postulación de clases verbales; en consecuencia, los predicados se clasifican en grupos y sus miembros adoptan necesariamente un tipo de construcción determinado, aspecto que condiciona y en cierta manera limita la descripción verbal ya que o bien no se da explicación a ciertas construcciones o bien se establece un número elevado de clases para poder dar cuenta de la diversidad verbal. Este principio metodológico puede sesgar la adquisición de información por su dificultad de tratar las excepciones –por ejemplo, comportamiento gramatical no acorde con el esperado de acuerdo con la clase asignada al verbo. En cambio, en SenSem las clases semánticas se toman en cuenta a posteriori, una vez realizada la anotación, por lo que se pueden realizar clasificaciones distintas y cruzadas en función de la propiedad sintáctico-semántica que se focalice.

3. METODOLOGÍA DE ANOTACIÓN

La metodología de anotación semántica de los núcleos argumentales de SenSem se basa en la establecida para la creación del corpus Eusemcor (Agirre, Aldezabal, Etxeberria, Iruskieta, Izagirre, Mendizabal y Pociello, 2006), en la cual, de forma simultánea e inseparable, se realizó la anotación de un corpus del euskera y la corrección y ampliación del WordNet de dicha lengua, con lo que se aseguraba la coherencia final entre ambos recursos.

La tarea que aquí se presenta ha sido realizada por seis lingüistas, tres graduados y tres posgraduados, desarrollando los primeros únicamente la función de anotadores y los tres últimos las funciones de anotadores y árbitros simultáneamente. La tarea ha tenido un coste de 8 personas/mes, ha durado 6 meses y se ha desarrollado dividida en las siguientes etapas:

1. Anotación morfosintáctica automática del corpus mediante FreeLing (Padró, Collado, Reese, Lloberes y Castellón, 2010) y consiguiente identificación de los sustantivos que debían ser anotados semánticamente.

2. Adaptación de la interfaz de anotación de Eusemcor para las necesidades específicas de SenSem.
3. Prueba de anotación para calcular el acuerdo entre los diferentes anotadores (Carrera, Castellón, Climent y Coll-Forit, 2008) y elaboración de criterios de desambiguación.
4. Anotación efectiva del corpus.
5. En paralelo a 4, sesiones de coordinación para la implementación y extensión de los criterios iniciales, pruebas de acuerdo entre anotadores y redefinición y fijación de los criterios.

Con carácter previo se definió el dominio de la anotación para adecuar la tarea a su objetivo final básico: la representación de las preferencias selectivas de los predicados verbales incluidos en la base de datos léxica de SenSem. En consecuencia, se anotaron únicamente los núcleos semánticos nominales de los argumentos ya anotados previamente. No se anotaron complementos ni modificadores de dichos sustantivos excepto en el caso de los sintagmas partitivos (por ejemplo, "un trozo de pan"), para los que se ha anotado tanto el núcleo sintáctico ("trozo") como el semántico ("pan"). Tampoco se anotaron los pronombres, ya que ello hubiera implicado la necesidad de resolver la correferencia, una tarea que aún no se ha abordado en el proyecto.

El etiquetado morfosintáctico automático es sin duda una fase imprescindible para el tratamiento del corpus en tanto que implica la detección masiva de elementos a anotar manualmente y su posterior carga en la interfaz de anotación, pero no resultó exento de problemas debido a diversos desajustes formales. El principal es que FreeLing no delimita adecuadamente las unidades léxicas multipalabra: tiende a marcarlas como palabras distintas y en consecuencia se presentan como tales al anotador. Como estrategia para superar esta limitación se adoptaron dos decisiones:

(i) Anotar el núcleo sintáctico con la información semántica propia de la forma léxica compleja; por ejemplo, en idiomatismos y colocaciones muy convencionalizadas como "mesa electoral" se anota "mesa" con el rasgo propio de *organización*.
(ii) Crear un operador de anotación (<MLTW>) para poder asignar al referido núcleo sintáctico la información de que forma parte de una unidad más compleja.

Para optimizar el esfuerzo y minimizar errores e inconsistencias, el etiquetado manual se realizó mediante una interfaz implementada como servicio web[7]; para ello se adoptó la interfaz desarrollada por el grupo IXA de la Universidad del

[7] http://ixa2.si.ehu.es/spsemcor.

País Vasco para la anotación de corpus del euskera antes mencionada, para acoger las características lingüísticas y formales de nuestro proyecto. Dicha interfaz está orientada a la búsqueda y anotación completa de todas las ocurrencias de un determinado lema en el corpus. Esta metodología asegura una mayor coherencia en la anotación por dos motivos:

(i) El lingüista afronta en un único bloque de trabajo la anotación completa de un único lema en todo el corpus, con lo que obtiene una visión holística de su semántica y en especial de la división de significados del lema en ESPWN1.6;
(ii) Un mismo anotador etiqueta todas las ocurrencias de un mismo lema en el corpus.

La interfaz es una herramienta flexible que permite la búsqueda por lema, por categoría morfosintáctica y por identificador de oración. Una vez seleccionado un lema, presenta al anotador todos sus posibles significados (*synsets*) en ESPWN1.6 y el contexto disponible en el corpus. Además, permite al lingüista modificar la categoría morfosintáctica asignada automáticamente a la ocurrencia.

En las primeras pruebas de anotación se establecieron instrucciones genéricas de procedimiento orientadas a optimizar la comprensión por parte del anotador de, dado un lema y los *synsets* que lo realizan, qué tipo de entidades o situaciones cubre cada uno de ellos. Para alcanzar dicha comprensión, el anotador recibe instrucciones de explorar distintos elementos informativos en ESPWN1.6 relacionados con el *synset* en cuestión, en los siguientes términos:

(i) Deben examinarse las relaciones semánticas del *synset*. Como mínimo el primer hiperónimo y el primer nivel de hipónimos. Si es posible, se examina toda la línea de hiperonimia.
(ii) Deben examinarse siempre los rasgos semánticos de las ontologías integradas en MCR, especialmente TO y SUMO por ser las más informativas ontológicamente y las elaboradas con mayor detalle.
(iii) Lo anterior implica que el anotador, en contra de su intuición lexicográfica habitual, no debe fiarse únicamente de las glosas, ya que éstas no son en WordNet definiciones precisas y por tanto suelen ser poco informativas. Sin embargo, también es conveniente examinarlas, pero en ese caso hay que tener en cuenta que suelen ser más acertadas las de WN1.6 inglés; las del WN1.6 en español pueden ser traducciones de las inglesas.

La fijación de los criterios concretos de desambiguación y anotación semántica se inició en la fase de prueba de anotación y continuó en las fases 4 y 5 hasta estabilizarse. Esta fase del proyecto permitió:

(i) En el aspecto procedimental, la realización del etiquetado de forma progresivamente más rápida y consistente; y
(ii) La obtención de un análisis de la utilidad y la problemática del uso de WordNet como recurso para la anotación semántica de sustantivos del español extrapolable a otros proyectos y lenguas.

En la fase de prueba se llevó a cabo un experimento de anotación (Carrera *et al.*, 2008) mediante el cual cuatro lingüistas anotaron en paralelo los sustantivos de 50 oraciones del corpus sin que se les suministrara más criterio que las instrucciones generales arriba indicadas. El acuerdo entre anotadores que se obtuvo fue sólo del 40%. En una segunda fase se pusieron en común los cuatro procesos de anotación, lo que condujo a la redacción consensuada de un primer conjunto de criterios de desambiguación. La revisión del trabajo basada en dicho consenso condujo a una reanotación, en la que se alcanzó un 84.32% de acuerdo entre anotadores. A partir de este momento se dividió la tarea de anotación, ya sin intersección entre anotadores. Sin embargo se mantuvieron reuniones periódicas para incrementar y refinar los criterios de desambiguación.

A continuación, en §4 se expone la evaluación del recurso ESPWN1.6 y en §5 los criterios de anotación dimanantes del análisis y el proceso.

4. EVALUACIÓN CRÍTICA DE WORDNET 1.6 COMO RECURSO DE ANOTACIÓN DE CORPUS

Clasificaremos la problemática general en cinco tipos de problemas derivados del uso de ESPWN1.6 como recurso de anotación: técnicos, estructurales, de ambigüedad o indefinición del significado, de incompletitud, y de carácter interlingüístico.

4.1. Problemas técnicos

En primer lugar consideramos problemas derivados de decisiones propias del diseño inicial de WordNet en cualquiera de sus versiones.

- **Descripción lexicográfica.** WordNet no incorpora definiciones lexicográficas precisas de los conceptos, sino artefactos más laxos: glosas y ejemplos. En muchos casos éstos entran en conflicto o incluso parecen contradecir el significado que dimana de las relaciones ontológicas del concepto en cuestión.
- **Morfología.** El pretratamiento de WordNet se realizó con FreeLing para poder lematizar las formas. Uno de los problemas que se encontró fue el hecho que la lematización de FreeLing y de WordNet no coincide. Estas diferencias afectan

básicamente a distintos criterios de lematización en formas con diferentes géneros (por ejemplo, 'hermana' en FreeLing tiene como lema 'hermano' cuando 'hermana' tiene un lema propio en WordNet) o a diminutivos (como por ejemplo: 'pueblecito' o 'jovencito'). En estos últimos, FreeLing utiliza directamente el diminutivo como lema (por considerar el diminutivo como resultado de un proceso derivativo y no flexivo), mientras que WordNet no incluye diminutivos que mantienen el significado del término original como variantes de los conceptos. Así, estas divergencias han provocado algunos problemas de anotación.

4.2. Problemas estructurales de WN1.6

Se trata en este caso de problemas derivados de la dificultad de una construcción adecuada de WordNet. Estos inconvenientes tienen su raíz en que la aplicación al léxico real de las estructuras ontológicas es un problema en muchos casos complejo.

- **Autohiponimia.** Es habitual que dos sentidos de un mismo lema aparezcan en situación taxonómica de hiponimia-hiponimia directa. En estos casos WordNet expresa niveles de generalidad distintos de lo que podría considerarse un mismo concepto. Por ejemplo, el lema 'trabajo' incluye siete sentidos en WordNet, entre los cuales trabajo_4 ("trabajo productivo") y trabajo_1 ("actividad dirigida a realizar algo") se encuentran en una relación de autohiponimia, ya que trabajo_4 tiene asignado como hiperónimo directo trabajo_1. En algunos casos la expresión del diferente nivel de generalidad es aún menos consistente, ya que aparecen en WordNet ambos sentidos, el más general y el más específico, sin relación estructural entre ellos. Éste es el caso del lema 'ritmo'. En WordNet, hay tres sentidos de este lema que forman una relación de autohiponimia ya que ritmo_3, en el sentido de velocidad de movimiento, y ritmo_6, en el sentido de velocidad a la que se repite algún evento, son, en realidad, hipónimos de ritmo_1, "magnitud relativa a una unidad de tiempo".
- **Falsa hiponimia.** En otra tipología de casos, la relación taxonómica es inexacta o falsa en términos ontológicos. Guarino (1998) las describe agrupadas en las siguientes categorías:

    - **Confusión de sentidos** en situaciones de herencia múltiple. En este caso sentidos diferentes se han colapsado en un mismo *synset*; por ejemplo, [Estatua_de_la_Libertad] como hipónimo simultáneo de [monumento] y [figura].
    - **Reducción de significado.** En este caso la hiponimia codifica el significado sólo de manera parcial. Por ejemplo, [organización] es hipónimo de [grupo] cuando es claramente algo más que un grupo de gente.

- **Sobregeneralización.** Se observa en situaciones de excesiva heterogeneidad de los cohipónimos, por lo que en realidad WordNet está atribuyendo indirectamente al hiperónimo un significado más general del que le es propio. Suelen ser soluciones coyunturales en la construcción de WordNet que obvían una estructuración taxonómica que debería ser más compleja.
- **Confusión entre tipo y rol.** Se trata de un error ontológico habitual en WordNet, por el cual la relación taxonómica, que por definición se debe dar entre tipos de entidad, se codifica entre un tipo y un rol. Por ejemplo, persona_1 como hipónimo de agente_causal_1. Una persona no es ontológicamente, es decir siempre, un agente causal, sino que éste es un rol que puede adoptar en un cierto conjunto de situaciones.
- **Confusión de tipo de relación.** La más habitual es la confusión entre taxonomía y meronimia; por ejemplo, 'hueso' (hueso_1 en WordNet) está codificado como hipónimo de 'tejido' (tejido_2 en WordNet), cuando en realidad un hueso no es "un tipo" de tejido sino una entidad "hecha de" tejido; en consecuencia la relación entre ambos *synsets* debería ser de meronimia.

4.3. Problemas de ambigüedad e indefinición en la distinción de sentidos

La excesiva granularidad, o la excesiva proliferación de sentidos, es el problema de WordNet más mencionado en la bibliografía, especialmente en la relativa a desambiguación automática de significados. La consecuencia directa de este problema es que las distinciones semánticas establecidas por WordNet son difíciles de identificar, incluso por anotadores humanos. Este problema arranca sin duda de uno teórico que aún no ha obtenido una solución clara en lingüística: ¿qué es un sentido de palabra? O incluso más, ¿puede el significado de las palabras ser dividido en sentidos?

Desde la lingüística cognitiva se ha argumentado (Taylor, 1995) que las palabras deben ser consideradas categorías radiales organizadas en torno a prototipos, lo que implica la imposibilidad, o como mínimo la gran dificultad, de distribuirlas en clases cerradas –es decir, sentidos excluyentes entre sí–. Por su parte autores como Kilgarriff (1997) argumentan que los sentidos de palabra tan sólo pueden ser inferidos, de manera emergente, a partir de las ocurrencias de las palabras en corpus.

En cualquier caso, si se adopta como es nuestro caso una metodología de anotación de corpus basada en el uso de WordNet como inventario de significados, se asume, aunque sea desde un punto de vista procedimental, la división de las palabras y los conceptos en sentidos diferenciados. Sin embargo, surge el problema de la falta de sistematicidad de WordNet a este respecto. Para un mismo lema pueden coexistir en WordNet varios sentidos muy similares y difíciles de diferenciar con vacíos de significado: la inexistencia de uno de sus sentidos obvios. Por otra parte,

distinciones regulares y conocidas, como los diversos tipos de polisemia regular (Apresjan, 1973) no se aplican (o dejan de aplicar) de manera sistemática.

Clasificaremos este problema general de WordNet, la **excesiva granularidad de significados**, en los siguientes tipos básicos:

- **Polisemia regular.** Como es sabido, algunos tipos de entidad o situación son sistemáticamente polisémicos, como en el caso de los eventos y su estado resultante. Por ejemplo, si se habla de "la educación de los jóvenes", es difícil discernir sobre si se está hablando de cómo se está educando o del resultado, esto es, el estado final de la persona en tanto que "educada" –ambas posibilidades son *synsets* en ESPWN1.6.
- **Diferenciación de sentidos a causa del punto de vista.** Es habitual en WordNet encontrar diversos *synsets* que en realidad están codificando el mismo tipo de entidad; la razón de la proliferación es que dicha entidad se está observando desde puntos de vista distintos. Por ejemplo, 'familia' tiene en ESPWN1.6 tres sentidos correspondientes, respectivamente, a "un grupo de gente que vive junta", "un grupo social" y "un grupo de gente relacionado por matrimonio o consanguineidad". La dificultad de atribuir uno de los tres sentidos a una ocurrencia de la palabra en el corpus es evidente, ya que en realidad los tres se están refiriendo al mismo tipo de entidad. Otro caso es el de las palabras que tienen dos sentidos, uno que denota un evento y otro que denota la percepción del mismo, como en el caso de 'problema', en el que, entre otros, hay *synsets* que denotan un hecho que causa dificultades, y otros que denotan el sentimiento o la percepción de dificultades. Entonces, ante frases como "X tiene problemas", es difícil discernir si estamos ante un hecho o una conceptualización del observador.
- **Sentidos modulados por el contexto.** Sólo un contexto muy rico puede permitir desambiguar entre tres posibles significados de 'interno' en ESPWN1.6: el niño que vive en el colegio durante el curso, la persona confinada en una institución (como una cárcel o un hospital), o el médico en prácticas residente en un hospital. Este tipo de polisemia sutil (ya que en el fondo se trata de tres variaciones de un mismo significado: alguien que vive en una institución) dificulta la anotación de oraciones, ya que suele faltar un contexto que se hallaría a nivel de párrafo o de discurso.
- **Intersección de sentidos.** Se dan casos en que dos sentidos de una palabra no cubren aspectos totalmente disjuntos del área de denotación, sino que intersectan. Por ejemplo, pista_1 y pista_6 remiten a áreas delimitadas para la práctica del deporte; en el primer caso se trata de una porción de terreno preparada para el deporte y en el segundo a unas instalaciones o servicios con el mismo fin. Pero se da el caso de que (a) existen terrenos de juego estructurados en tanto que instalaciones deportivas, a la vez que (b) existen terrenos de juego

que no lo están y también (c) instalaciones deportivas que no incluyen terrenos de juego o "pistas". Por lo tanto, mientras que las pistas de tipo (b) y (c) son claramente clasificables respectivamente como pista_1 y pista_6, las de tipo (a) pueden serlo en ambas, ya que ponen de manifiesto que los sentidos de pista_1 y pista_6 intersectan.

Sin duda estos tipos de ambigüedad son hechos del lenguaje, no problemas de WordNet. El problema de WordNet es que, en vez de tratar dichos problemas de alguna manera estructuralmente elegante y simple, opta por añadir nuevos sentidos de palabras ante las dudas, y, además, ello no se realiza de manera sistemática en todas las palabras sino a todas luces a partir de las distintas casuísticas ante las que se encontraron los distintos desarrolladores de la base de conocimiento.

Un problema distinto aunque relacionado surge no del diseño de WordNet sino de qué información proporciona:

- **Información insuficiente o confusa.** Una distinción semántica puede ser clara, pero ESPWN1.6 no proporciona información clara de cuál es la distinción ni a partir de la glosa ni a partir de las relaciones estructurales –que son, de hecho, las dos maneras que tiene WordNet de ofrecer información semántica. Por ejemplo, en el caso anterior del lema 'pista', aunque ya hemos visto que la distinción no es clara, a ello se añade el hecho de que las respectivas glosas son prácticamente idénticas, casi paráfrasis, y sus dos primeros niveles de hiperonimia remiten a conceptos distintos, pero todos ellos relativos a áreas o porciones de terreno conceptualizados de manera ligeramente diferente.

4.4. Problemas de incompletitud de ESPWN1.6

Es obvio que WordNet, pese a ser la mayor base de conocimiento léxico-semántico, no recoge todo el vocabulario existente de una lengua. Por supuesto ello no es posible porque el léxico está sujeto a la creatividad y otras formas de constante incremento. Sin embargo, hay que dejar constancia de que ello es un problema importante para la anotación de corpus y como tal debe ser afrontado. Las principales dimensiones de incompletitud de WordNet son las siguientes:

- **Falta de *synset*.** Existen sentidos de un lema que no aparecen en WordNet, el recurso no recoge el concepto. Un ejemplo lo encontramos en 'sanidad', en el sentido "Conjunto de servicios gubernativos ordenados para preservar la salud del común de los habitantes de la nación, de una provincia o de un municipio" (DRAE[8]), sentido que se asocia a ocurrencias como "sanidad española", "sani-

[8] Disponible en: http://drae.rae.es/.

dad catalana", etc. En WordNet no se expresa este concepto (cuya traducción al inglés seria algo similar a "health_system" o "social security"), por lo que tampoco se ha creado el concepto en español, ya que el WordNet español se construyó por proyección del inglés.

- **Falta de *variant.*** Asimismo también pueden faltar sinónimos de conceptos existentes (el *synset* es incompleto). Por ejemplo, en la estructura de WordNet en inglés hay una *variant* para *living will*, que en español se corresponde a "testamento vital". De todas formas, en la estructura de WordNet en español, "testamento vital" no está recogido en la versión utilizada.
- **Incompletitud indeterminada.** Algunos casos están a medio camino entre la falta de *synset* o la falta de *variant*. Es el caso de 'baile' (como en "baile de cifras"). Se puede considerar que un concepto cercano a 'baile' aparece en WordNet ('variación'), aunque este *synset* no es el concepto exacto para este lema, ya que 'baile' implica reiteración en la variación. Las posibles soluciones darían lugar a nuevos problemas, ya que si ESPWN1.6 incorporara para este concepto un nuevo *synset* se produciría intersección de sentidos (cf. apartado anterior). Si, en cambio, incorporara 'baile' como *variant* en el *synset* de "variación" se produciría una reducción de sentido de las descritas por Guarino (1998) –cf. § 4.2–.
- **Sentidos metafóricos.** WordNet no recoge de manera sistemática los significados creados por extensión metafórica. No se trata de un defecto reprochable, puesto que virtualmente cualquier palabra puede ver su significado extendido mediante metáfora conceptual. El problema de WordNet es su inconsistencia, ya que recoge las extensiones metafóricas en algunos casos, pero en otros no, de manera aparentemente aleatoria; por ejemplo en ESPWN1.6 'puente' incorpora, además del sentido arquitectónico original, los sentidos de prótesis dental y de figura gimnástica, pero en cambio no el de período festivo. Es más, en algunos casos incluye el sentido metafórico pero no el original, como en el de 'pie', en el que ESPWN1.6 incorpora el sentido correspondiente a la base de una montaña pero no el de la extremidad inferior del ser humano. Ambos ejemplos son achacables, sin duda, a incompletitudes en la construcción del WordNet español a partir del inglés, en el primer caso porque el significado de 'puente' como "festivo" no está lexicalizado en inglés y en el segundo a una simple incompletitud.
- **Falta de unidades multipalabra.** De la misma manera, ESPWN1.6 incorpora o no unidades multipalabra sin seguir ningún criterio aparente; por ejemplo, WordNet recoge como multipalabra 'agente secreto', que tiene una lectura no composicional y, a su vez, 'agente de policía', que tiene una lectura forzosamente composicional.
- **Falta de nombres propios.** ESPWN1.6 incorpora un cierto número de nombres propios, pero naturalmente no la totalidad de los existentes. Los nombres

propios incorporados pertenecen culturalmente a los Estados Unidos, también a causa de la construcción de ESPWN1.6 a partir del WordNet inglés (ENGWN1.6).

4.5. Problemas de carácter interlingüístico: versiones idiomáticas de WN1.6

Diversos problemas de los descritos hasta el momento tienen su base en mayor o menor medida en el hecho de que WN1.6 del español se ha construido partiendo de WN1.6 del inglés: manteniendo su estructura y realizando en muchos casos traducciones, por lo que el nivel de adaptación al español es limitado. En este apartado, sin embargo, recogeremos los problemas directamente debidos a este hecho.

- **Falta de equivalente en español.** En el anterior apartado recogíamos la falta de expresión de un concepto en WordNet. En éste el caso es que el concepto existe en ENGWN1.6 (y por tanto se podría anotar el corpus con el mismo), pero no se ha recogido su equivalente al construir el WordNet del español. Por ejemplo, 'juez' es monosémica en ESPWN1.6 ya que sólo aparece su sentido habitual propio del sistema jurídico. En cambio no aparece en el sentido de "evaluador", que sí se halla presente en el WordNet inglés ([judge_2, evaluator_1]).
- **Desdoblamiento indebido de *synset* a causa de diferencias morfológicas.** Dado que el inglés no tiene flexión de género se dan casos de palabras que en inglés tienen dos lemas mientras que en español sólo tienen uno. Ello causa, al crear ESPWN1.6 por proyección de ENGWN1.6, el desdoblamiento del lema español en *synsets* diferentes en función del género. Un caso claro es el de 'tío' que en inglés se codifica en dos *synsets* distintos: [aunt_1] y [uncle_1]. Ello ha provocado la existencia por proyección de dos *synsets* en español, [tia_1] y [tio_1] cuando por la morfología del español únicamente debería existir uno, tal como ocurre con todas las palabras que tienen flexión de género. Un caso similar se da en los *synsets* [hermano] y [hermana], que se han creado por proyección de los *synsets* del inglés [brother] y [sister].
- **Sesgo cultural.** Se dan casos en que la codificación de conceptos en el WordNet inglés está muy marcada culturalmente, por lo que les falta la deseada objetividad interlingüística. Este factor se proyecta de manera inadecuada en el WordNet en español causando desajustes diversos. Por ejemplo, el lema 'president' (presidente) no recoge el caso de los presidentes de gobierno en estados de tipo monárquico o similar, en los que el presidente no es la primera autoridad del Estado. Ello es debido al seguimiento directo del caso norteamericano, en que el presidente lo es del Estado. Por ello, ENGWN1.6, y ESPWN1.6 por proyección, únicamente recoge la posibilidad de 'presidente' como (a) primera

autoridad de una república o (b) primera autoridad de un Estado. Ello impide la correcta anotación de frases como "el Presidente Zapatero".

5. SOLUCIONES Y GUÍA DE CRITERIOS DE ANOTACIÓN

En este apartado detallaremos las soluciones adoptadas para los problemas expuestos en §4, los cuales se han recogido en una guía para los anotadores. Como se verá, la mayoría de ellas son soluciones de tipo pragmático condicionadas por la decisión de anotar sin modificar ESPWN1.6, ya que afrontar una tarea conjunta de anotación y modificación de WordNet no se consideró realista en los términos del proyecto que aquí se describe. Sin embargo, la descripción de problemas de §4 puede significar una buena base de inicio para afrontar modificaciones de WordNet en proyectos futuros.

5.1. Solución de problemas técnicos

Presentamos las diferentes soluciones tomadas para los casos citados anteriormente.

- Para solventar el problema de las **definiciones lexicográficas laxas** se definieron los siguientes procedimientos de anotación:

    - Es necesario consultar siempre las relaciones del *synset*, como mínimo las de hiponomia e hiperonimia, para hacerse una idea del concepto que se está tratando.
    - Es necesario consultar los rasgos semánticos, especialmente los de la EWN TO y los de SUMO que nos sitúan en la clase ontológica del concepto.
    - Las glosas son normalmente menos informativas que las relaciones y que los rasgos semánticos; se tienen que consultar pero con prudencia. Se debe prestar más atención a la glosa del inglés que a la del español ya que estas últimas suelen ser traducciones no demasiado acertadas de las glosas del inglés.

- **Morfología.** En estos casos la interfaz no nos permitía superar la divergencia de lema entre FreeLing y WordNet, así se ha optado por elaborar una lista de este tipo de incidencias anotando el *synset* que corresponde al diminutivo. Es decir el anotador realiza la tarea; pero no en la interfaz sino en un listado independiente y se anota esta situación en los comentarios.
- **Operadores.** Asimismo, para poder matizar ciertas anotaciones fue necesario establecer una serie de operadores que permitieran más flexibilidad a la hora de seleccionar el *sysnet*. Los operadores utilizados están listados en la Tabla I.

Tabla I. Operadores utilizados en la anotación.

Operador	**Explicación**
MTF	Uso metafórico de un *synset* existente siempre que no exista ya la interpretación metafórica en WordNet.
MLTW	Multiword. Se trata de una multipalabra en la que se ha asociado el núcleo a un *synset*.
FLT	No existe o no se encuentra el sentido en WordNet.
HIP	Falta contexto y la anotación es una hipótesis.
DUD	Duda, falta contexto y no hay hipótesis.
PRT	Partitivo. Por ejemplo, un millón de pesetas, un vaso de agua, un kilo de tomates.
MET	Metonimia. Anotamos un *synset* pero indicamos que la ocurrencia es una metonimia de este concepto.

5.2. Solución de problemas estructurales

- **Autohiponimia.** Siempre que se encuentren dos *synsets* que representen un mismo concepto (o muy cercano), uno más genérico y uno más preciso, si el contexto no ayuda a la interpretación, se escoge el *synset* más genérico (por ejemplo: 'asociación_1' vs. su hipónimo 'asociación_3'). No tomar esta medida implicaba tiempo de decisión para el anotador y poca sistematicidad entre anotadores. De esta forma las decisiones que se toman en cuanto a la anotación de un término genérico vs. uno específico que es una concreción del anterior siempre son más generalistas que particulares, lo que para el objetivo final de la anotación resulta más adecuado. Observemos como ejemplo la distribución de tres sentidos del lema 'asociación':

    - grupo social_1 < organización_1 < **asociación**_1 < **asociación**_5
    - grupo social_1 <colectividad_1 < **asociación**_2

El sentido 1 se glosa genéricamente como "Grupo organizado de personas" y el 5 como "Grupo de personas con intereses similares", con sinónimos como "club" o "fraternidad". La diferencia con "asociación 2" es más clara ya que éste refiere a grupos no específicamente organizados, por ejemplo, "sector" en el sentido ocupacional o económico.

- **Falsa hiponimia.** Este tipo de errores han sido tratados individualmente caso por caso igual que en los problemas de ambigüedad e indefinición de sentidos que explicamos a continuación.

5.3. Problemas de ambigüedad e indefinición en la distinción de sentidos

- En los casos de dificultad en la distinción de sentidos y ante la imposibilidad de elaborar criterios generales, los árbitros han elaborado una guía para cada una de las palabras que los anotadores detectaban como difíciles por tener sentidos demasiado finos, estas guías se elaboraban con el acuerdo de dos o tres árbitros. En la Figura 1 podemos observar la guía elaborada para el lema 'consejo'.

Sentidos del lema 'consejo':
consejo_1: es un comité.
consejo_2: es una recomendación.
consejo_3: se trata de directrices.
consejo_4: es también un comité.
consejo_5: se trata de un "chivatazo". Mala traducción de ESPWN.
consejo_6: es un comité no estable, improvisado.

Instrucciones de anotación:
- Entre 2, 3 y 5 utilizar únicamente 2, a menos que se encuentren construcciones donde el sentido de 2 y 5 esté claramente diferenciado (poco probable).
- 1 y 4 son prácticamente iguales y ambos hipónimos de "unidad administrativa". Se anota siempre consejo_1.
- consejo_6 sólo se utilizará cuando sea muy evidente que el contexto habla de comités de constitución no estable. Si no se utiliza 1.

Figura 1. Guía para la desambiguación del lema 'consejo'.

- Otros criterios afectan no a palabras concretas sino a clases específicas de sustantivos, como es el siguiente:

  - Ante casos de polisemia entre entidad objetiva y el correspondiente uso social (por ejemplo, entre año físico y año social), a menos que la evidencia en contra sea muy notoria, se elige el uso social.

- Para casos de proliferación de sentidos de difícil decisión se ha procedido a la agrupación de sentidos de EWNESP1.6. Por ejemplo, tomemos las oraciones siguientes, susceptibles de ser anotadas con experiencia_2 o experiencia_3:

  (i) Más que en su obra literaria, es en estos dibujos donde Hugo vuelca su desbordante imaginación y manifiesta sus propias experiencias y vivencias personales como el exilio, la muerte de sus hijos o su intensa actividad política.
  (ii) Esta es la segunda vez que los transportes públicos realizan esta experiencia.

De acuerdo con WordNet, *experiencia_2* refiere al "contenido de la observación o la participación directas en un evento", mientras que *experiencia_ 3* refiere al "evento tal y como se percibe". Dada la dificultad casi metafísica de distinguir entre ambos sentidos, se ha optado por agruparlos. El total de agrupaciones realizadas es de 58, afectando a un total de 129 *synsets*.

5.4. Problemas de incompletitud

- **Sentidos metafóricos y metonímicos.** Las ocurrencias en que el sentido metafórico o metonímico no está especificado en WordNet se anotan usando el *synset* correspondiente a una interpretación literal, y se marcan con el operador <MTF> para indicar su interpretación metafórica; por ejemplo, 'barón' de un partido u organización y 'barón' en sentido nobiliario; el segundo sentido está recogido en WordNet pero no el primero.
- **Nombres propios, fechas y cantidades de dinero.** Se anotan mediante las categorías establecidas en las MUC[9] (Tabla II). Aunque WordNet contiene un número importante de nombres propios, muchos de estos son locales –pertenecen a la cultura o el ámbito estadounidense. Este es el caso de "Agencia Tributaria", un concepto propio del español y, por consiguiente, no recogido en la estructura del WordNet del inglés (tampoco en la del español):

  (i) Acabo de recibir de la Agencia Tributaria el tríptico con la nueva asignación tributaria, o sea, lo que marcaremos en nuestra declaración de la renta de 1999 para desviar a la Iglesia católica o a otros fines de interés social: el 0,5239% de lo que pagamos.

Así ante la carencia de conceptos relacionados con los nombres propios optamos por asignar un *synset* genérico de clase (humano, organización, etc.). La propuesta MUC se ha convertido en estándar para tareas de identificación y categorización de nombres propios en tratamiento de la información. Por esta razón y por considerarlas adecuadas para el establecimiento de preferencias selectivas se han adoptado en el presente proyecto. Sin embargo, como excepción a esta regla, los nombres propios que contengan el nombre común en su expresión se anotan con el *synset* de dicho nombre común. Por ejemplo, para 'Restaurante Bemby' se asigna el *synset* correspondiente a [restaurante]. De esta manera, se obtiene información más precisa que la aportada por la categoría MUC.

[9] Message Understanding Conference, competición para el desarrollo y evaluación de sistemas de extracción de información. Disponible en: http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html.

Tabla II. Categorías del MUC para nombres propios o entidades referenciales.

ID Synset	Categoría
09639711	*Money*
09894531	*Figure*
00015594	*Age/Time*
00020056	*Other (quantity)*
09760609	*Unit of measurement*
06381267	*Place*
00017297	*Event*
00010123	*Natural object*
00002086	*Life_form*
00004123	*Person*
05997592	*Organization*
00011937	*Artifact*
03569523	*Vehicle*
04493671	*Concept*

- **Falta de sentidos.** Si el sentido de un lema que aparece en el corpus no aparece como *synset* en ESPWN1.6, se documenta y se describe el caso para posteriores revisiones del recurso, por ejemplo, 'batería' en el sentido de músico. El mismo proceso se aplica para las variantes (sinónimos) que tampoco existen en WordNet español. Un caso especial de falta de sentido es el de falta de sentido metafórico o metonímico (cf. supra).
- **Multipalabras.** Como se ha descrito en §3, no es posible la anotación de léxico multipalabra en este corpus. En consecuencia, el anotador debe proceder de la siguiente manera: si la interpretación de la multipalabra es compositiva (por ejemplo, colegio electoral) el anotador consulta ESPWN1.6 y anota analíticamente las partes de la multipalabra. Si no es compositiva y está presente en ESPWN1.6 (por ejemplo, célula madre), se etiqueta como <MTW> (*multiword*). Si ni es compositiva ni está recogida en ESPWN1.6 (por ejemplo, puesta en marcha), se etiqueta con los operadores <MTW> y <FLT> (*falta sentido*).

5.5. Problemas de carácter interlingüístico

En determinadas ocasiones, debido a que el ENGWN1.6 es más completo que ESPWN1.6, ciertos conceptos están recogidos en el primero pero no en el segundo. Otro problema de carácter interlingüístico es la divergencia entre las estructuras conceptuales de las dos lenguas. Un caso frecuente es la diferencia en la distinción de género gramatical y el concepto genérico que engloba ambos, por ejemplo, [sobrino] y [sobrina] no están conectados directamente por un mismo

hiperónimo en ENGWN1.6 ya que se distingue por géneros al corresponder a léxico distinto en inglés. Dada la imposibilidad tanto de anotar una ocurrencia cuyo lema no aparezca en WordNet, como de modificar la estructura actual de ESPWN1.6, se opta por incluir el caso en una lista para la construcción de la siguiente versión –proyecto en marcha (Fernández-Montraveta, Vázquez y Fellbaum, 2008).

6. RESULTADOS

Como resultado del trabajo presentado, el corpus SenSem se ha anotado semánticamente en un total de 23.307 formas correspondientes a 3.693 lemas. Ello representa la anotación del 82,6% del total del corpus. En esta fase, con el fin de alcanzar un máximo de rentabilidad a cambio de esfuerzo, no se han anotado los lemas que ocurren en el corpus menos de 5 veces. Por ello, queda pendiente de anotar un 17.4% del corpus. 91 lemas no se han anotado por no estar representados en ESPWN1.6; la mayoría son nombres propios culturalmente locales, por lo cual no se hallan presentes en la estructura de ENGWN1.6. Se conserva la lista para ser incluidos en la próxima versión de ESPWN. Mostramos a continuación un ejemplo de oración etiquetada en este proyecto.

- *Actuación de David Rees-Williams, que interpreta obras de Bach, Franck, Sweelinck, Alain y Yon.*

En esta oración, *obras* se ha anotado con el *synset* [05277178n] formado por el conjunto sinonímico {composición_2, pieza_musical_1, pieza_1, obra_3}, cuyo hipónimo es [02518101n creación_5] y su glosa "Producción musical: *las cuatro estaciones es una obra de Beethoven; obra musical*". Los sentidos alternativos en ESPWN son:

- {obra_1} > {producción_1, producto_2}: "Aquello que ha sido producido o conseguido a través del esfuerzo o la actividad de una persona o cosa*: la erosión es la obra del viento y el agua en el tiempo; la obra de un hombre imaginativo; obra benéfica; se lo debía a la obra pionera de John Dewey*".
- {obra_2} > {producción_5, producto_final_1}: "Producción total de un escritor o artista: *la obra de Salvador Espriu*".
- {obra_4} > {espectáculo_4 show_1}: "Representación pública de una obra teatral, cinematográfica, etc.: *querían ver alguno de los espectáculos de Broadway*".
- {obra_5} > {construcción_1}: "Lugar donde se construye algo, especialmente un edificio: *los obreros llegan puntuales a la obra*".

El resto de información que ofrece SenSem (anotación fruto de proyectos anteriores) sobre la construcción nucleada por 'interpretar' es la siguiente: *Evento imperfectivo; Polaridad positiva: Focalización 1er participante; "que": [Sujeto, Sintagma pronominal (pronombre relativo); Agente]; "obras de Bach..."[Objeto directo, Sintagma nominal (Nombre común); Tema].*

Este haz de información, desarrollado para cien oraciones de cada uno de los 250 verbos más frecuentes del español, ofrece una gran potencialidad para el desarrollo de proyectos de lingüística teórica y aplicada, estudios empíricos sintáctico-semánticos, desambiguación de significados o desarrollo de aplicaciones docentes.

7. CONCLUSIONES Y TRABAJO FUTURO

En este artículo se ha presentado la metodología y el desarrollo de un proyecto de desambiguación semántica de los núcleos argumentales de SenSem, un corpus equilibrado constituido por 100 oraciones para cada uno de los 250 verbos más frecuentes del español. El resultado, unido a desarrollos anteriores de SenSem, es un corpus etiquetado sintáctica y semánticamente: sentido verbal, sentido de los núcleos argumentales, tipos de constituyentes, funciones de los argumentos, papeles temáticos, tipo de construcción e información aspectual. El corpus está conectado a una base de datos que recoge la información pertinente para cada sentido verbal, por lo que el recurso resultante es muy adecuado para estudios empíricos centrados en el verbo. En concreto, uno de los objetivos futuros de esta investigación es la adquisición y representación de preferencias selectivas para su inclusión en una gramática de dependencias (Lloberes, Castellón y Padró, 2010).

El recurso utilizado para la anotación de los núcleos nominales ha sido la base de conocimiento WordNet del español en su versión 1.6. El proceso de etiquetado ha sido realizado por seis lingüistas y se ha dividido en las siguientes etapas: 1) etiquetado morfosintáctico automático del corpus; 2) implementación de una interfaz de anotación que permite búsquedas por lema, categoría morfosintáctica e identificador de oración; 3) pruebas de anotación y definición de criterios; 4) anotación efectiva del corpus y extensión de criterios, entre los cuales destaca el establecimiento de operadores para poder identificar usos metafóricos, metonimias, multipalabras y partitivos.

Como resultado colateral del proceso se ha realizado un análisis crítico de WordNet 1.6 del español, aplicable en muchos aspectos a los WordNets en general, como recurso de anotación lexico-semántica de corpus. Los puntos débiles detectados en el recurso han dado lugar a la elaboración de una guía de criterios de anotación, que también puede resultar útil para los investigadores que emprendan tareas similares de etiquetado con WordNet. Asimismo, la casuística detectada está siendo aplicada por el grupo en la construcción de la versión 3.0 del WordNet

español (Fernández *et al.*, 2008): inclusión de usos metafóricos o metonímicos del léxico, agrupación de sentidos, incorporación de sentidos o palabras ausentes, mejora o sustitución de glosas, etc.

El recurso presentado se puede obtener libremente bajo licencia GPL en http://grial.uab.es/sensem/download/main.

REFERENCIAS

Agirre, E., Aldezabal, I., Etxeberria, J., Iruskieta, M., Izagirre, E., Mendizabal, K. y Pociello, E. 2006. "Improving the Basque WordNet by corpus annotation", en *Proceedings of the Third International WordNet Conference,* pp. 287-290.

Agirre, A. y Edmonds, Ph. (Eds.). 2007. *Word sense disambiguation algorithms and applications*. *Text, Speech and Language Technology*, Vol. 33. Berlin, Heidelberg, New York: Springer-Verlag.

Alonso, L., Capilla, J.A., Castellón, I., Fernández, A. y Vázquez, G. 2007. "The SenSem Project: Syntactico-Semantic annotation of sentences in Spanish". En Nikolov, K., Bontcheva, K., Angelova, G. y Mitkov, R. (Eds.). *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*. Amsterdam, Philadelphia: Benjamins Publishing Co., pp. 89-98.

Álvez J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A. y Rigau, G. 2008. "Complete and Consistent Annotation of WordNet Using the Top Concept Ontology", en *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, pp. 1529-1534.

Apresjan, J. 1973. "Regular Polysemy", en *Linguistics* 142, pp. 5-32.

Atserias, J., Climent, S., Farreres, J., Rigau, G. y Rodríguez, H. 2000. "Combining multiple methods for the automatic construction of Multilingual WordNets". En Nikolov, K., Bontcheva, K., Angelova, G. y Mitkov, R. (Eds.). *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*. Amsterdam, Philadelphia: Benjamins Publishing Co, pp. 143-149.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B. y Vossen, P. 2004. "The MEANING Multilingual Central Repository", en *Proceedings of the Second International WordNet Conference-GWC*, pp. 23-30.

Bentivogli, L. y Pianta, E. 2005. "Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus", en *Natural Language Engineering, Special Issue on Parallel Texts* 11 (3), pp. 247-261.

Brill, E. 1995. "Unsupervised learning of disambiguation rules for part of speech tagging". En Yarowsky, D. y Church, K. (Eds.), *Proceedings of the Third Association for Computational Linguistics Workshop on Very Large Corpora*. Cambridge: MA, pp. 1-13.

Carrera, J., Castellón, I., Climent, S. y Coll-Forit, M. 2008. "Towards Spanish verbs' selectional preferences automatic acquisition. Semantic annotation of SenSem corpus", en *Proceedings of the 6th international conference on Language Resources and Evaluation (LREC 2008*), pp. 2397-2402.

Davies, M. 2002. "Un corpus anotado de 100.000.000 de palabras del español histórico y moderno", en *Procesamiento del Lenguaje Natural* 29, pp. 21-27.

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

Fernández-Montraveta, A., Vázquez, G. y Fellbaum, C. 2008. "The Spanish version of WordNet 3.0". En Storrer, A., Geyken, A., Siebert, A., Würzner, K.M. (Eds.). *Text resources and lexical knowledge*. Berlin: Mouton de Gruyter, pp. 175-182.

Fillmore, C. J., Johnson, C. R. y Petruck, M. R. L. 2003. "Background to Framenet", en *International Journal of Lexicography* 16-3, pp. 235-250.

Francis, W. N. y Kucera, H. 1964. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. USA: Brown University.

García-Miguel, J. y Albertuz, F. J. 2005. "Verbs, semantic classes and semantic roles in the ADESSE project". En Erk, K., Melinger, A. y Schulte im Walde, S. (Eds.). *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pp. 50-55.

Guarino, N. 1998. "Some Ontological Principles for Designing Upper Level Lexical Resources". En *Proceedings of the First International Conference on Language Resources and Evaluation.* Granada: ELRA, pp. 527-534.

Hovy E., Marcus, M., Palmer, M., Ramshaw, L. y Weischedel, R. 2006. "OntoNotes: The 90% Solution". En *Proceedings of HLT/NAACL,* pp. 57-60.

Kilgarriff, A. 1997. "I don't believe in word senses", en *Computers and the Humanities,* 31 (2), pp. 91-113.

Lloberes, M., Castellón, I. y Padró, L. 2010. "Spanish freeling dependency grammar". En *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Malta: ELRA, pp. 693-699.

Marcus, M. P., Santorini, B. y Marcinkiewicz, M.A. 1993. "Building a large annotated corpus of English: the Penn Treebank", en *Computational Linguistics* 19 (2), pp. 313-330.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C. y Thomas, R. G. 1994. "Using a semantic concordance for sense identification". En *Proceedings of the ARPA Human Language Technology Workshop*. Plainsboro, New Jersey, pp. 240-243.

Navarro F. de B. 2007. Metodología, construcción y explotación de corpus anotados semántica y anafóricamente. Tesis Doctoral. Universidad de Alicante.

Niles I. y Pease A. 2003. "Linking Lexicons and Ontologies: Mapping WordNet to the suggested upper model ontology". En *Proceedings of the 2003 Internatio-*

nal Conference on Information and Knowledge Engineering. IKE'03, June 2003, Las Vegas, Nevada, Vol. 2, CSREA Press, pp. 23-26.

Padró, L., Collado, M., Reese, S., Lloberes, M. y Castellón, I. 2010. "FreeLing 2.1: Five years of open-source language processing tools". En *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Malta: ELRA, pp. 931-936.

Palmer, M., Gildea, D. y Kingsbury, P. 2003. "The Proposition Bank: An annotated corpus of semantic roles", en *Computational Linguistics* 31 (1), pp. 71-106.

Rojo, G. 2001. "La explotación de la Base de Datos Sintácticos del español actual". En De Kock, J. (Ed.) *Lingüística con corpus. Gramática española Enseñanza e investigación.* Salamanca: Ediciones de la Universidad de Salamanca.

Taulé, M., Martí, M.A. y Recasens, M. 2008. "AnCora: Multilevel Annotated Corpora for Catalan and Spanish". En *Proceedings of the 6th conference on International Language Resources and Evaluation (LREC 2008)*, Marraquesh: ELRA, pp. 96-101.

Taylor, J.R. 1995. *Linguistic Categorization.* Oxford: Oxford University Press.

Vázquez, G. y Fernández, A. 2008. "Annotation de corpus: Sur la délimitation des arguments et des adjoints", en *SKY Journal of Linguistics* 2, pp. 244-269.