3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. *

M. Palomar (1), M. Civit (2), A. Díaz (3), L. Moreno (4), E. Bisbal (4), M. Aranzabe (3), A. Ageno (5), M.A. Martí (2), B. Navarro (1)
(1) Departamento de Lenguajes y Sistemas informáticos
(Universidad de Alicante) {mpalomar,borja}@dlsi.ua.es
(2) CLiC Centre de Llenguatge i Computació (Universitat de Barcelona)
civit@clic.fil.ub.es, amarti@ub.edu
(3) IXA Taldea (Euskal Herriko Unibertsitatea)
{jipdisaa,maxux}@si.ehu.es
(4) Departamento de Sistemas Informáticos y Computación
(Universidad Politécnica de Valencia)
{ebisbal,lmoreno}@dsic.upv.es
(5) Departament de Llenguatges i Sistemes Informàtics
(Universitat Politècnica de Catalunya) ageno@lsi.upc.es

Resumen: En este artículo presentamos los resultados del proyecto 3LB, consistente en el desarrollo de tres corpus (para el catalán, el castellano y el euskera) anotados sintáctica y semánticamente. Se exponen los criterios que se han seguido para las diferentes anotaciones, las diferentes herramientas desarrolladas para los distintos etiquetados, así como los resultados de evaluación de la anotación.
Palabras clave: Corpus, anotación sintáctica, anotación semántica, castellano, catalán, euskera

Abstract: In this paper, we present the results of the 3LB project, which consist on the development of three corpora (one for Catalan, one for Spanish and one for Basque) with syntactic and semantic annotation. We show the criteria followed for each annotation, the different tools developed for each tagging and the results of annotation evaluation.
Keywords: Corpus, syntactic annotation, semantic annotation, Spanish, Catalan, Basque

1. Introducción

En este artículo presentamos los resultados del proyecto 3LB, cuyo objetivo ha sido el desarrollo de tres corpus anotados a nivel lingüístico, con información sintáctica y semántica: **Cat3LB**, **Cast3LB** y **Eus3LB**. Los tres corpus son parcialmente comparables, ya que un 25 % de los mismos procede de noticias de agencia de las mismas fechas. Los resultados de este proyecto, los corpus anotados, son de libre disposición para investigación.

La anotación sintáctica se ha realizado a dos niveles, constituyentes para los tres corpus, y funciones en el caso de catalán y castellano, y dependencias para el euskera. Presentamos una evaluación cuantitativa y cualitativa de su anotación.

* Este trabajo ha sido parcialmente financiado por los proyectos PROFIT (FIT-15 0500-2002-244) y X-TRACT-II (BFF2002-04226-C03-03)

El desarrollo y adaptación de herramientas para las anotaciones sintáctica y semántica ha facilitado el proceso de anotación así como la calidad de los resultados finales.

El artículo se estructura como sigue: la sección 2 presenta el etiquetado sintáctico, que incluye los criterios desarrollados para la anotación; la evaluación de los grados de acuerdo y las herramientas que se han utilizado. En la sección 3 se tratan las directrices adoptadas para el etiquetado semántico, la herramienta desarrollada y los resultados obtenidos. Finalmente, en la sección 4 presentamos las conclusiones y las líneas futuras.

2. El etiquetado sintáctico

El esquema de anotación que se ha utilizado para anotar los corpus es dependiente de las características de la lengua. Mientras el catalán y el castellano son lenguas que presentan una clara estructura de con-

stituyentes, que pueden ocupar una posición más o menos libre en la oración, el euskera presenta un orden libre de palabras en la oración. Por ello, los esquemas de anotación que se han seguido en la sintaxis son distintos y, mientras en el primer caso se anotan constituyentes y funciones, en el segundo se anotan dependencias.

2.1. Líneas de etiquetado sintáctico para constituyentes y funciones

La anotación sintáctica del castellano fue la que primero se inició y la metodología, en lo relativo a la anotación de constituyentes, fue la siguiente:

- chunking automático previo con TACAT (Atserias y Rodríguez, 1998) y GramEsp (Civit, 2003a).
- anotación, por parte de cinco lingüistas, de 100 oraciones del texto para establecer los principios básicos de anotación;
- anotación de 220 oraciones para verificar la adecuacion del esquema anteriormente definido. Modificación y ampliación de las guías de anotación;
- revisión de la anotación de todas las frases anteriores para la detección de errores y refinamiento de los criterios;
- anotación de 700 frases, también por parte de los cinco anotadores, con el fin de verificar si se producía ya un acuerdo significativo entre los diferentes anotadores, como así ocurrió.

Tras este proceso en paralelo, y con la guía ya completa (Civit, 2003c), se ha anotado el resto de las oraciones del texto por parte de un equipo de dos anotadores y por separado.

Por otra parte, con las primeras mil oraciones del corpus anotadas a nivel de constituyentes, se inició la anotación de las funciones sintácticas. El proceso fue similar, puesto que se anotó primero un conjunto de 100 oraciones, con la ayuda de una versión preliminar de la guía. Se compararon los resultados entre los anotadores, dos en este caso, se actualizó y revisó la guía (Civit, 2003b) y se procedió a anotar el resto del corpus a este nivel.

La anotación sintáctica del catalán se inició con posterioridad a la del castellano, por lo que las anotadoras ya tenían experiencia previa. El proceso fue el mismo que en el caso del castellano, pero mucho más rápido. Tras la comparación de la anotación de 200 oraciones se realizó una primera comparación de los resultados; se refinaron y explicitaron los criterios (Valverde, Civit, y Bufí, 2004) y luego se prosiguió la anotación ya por separado. Avanzada ya la anotación de constituyentes, se inició la anotación de funciones. Las mismas 200 oraciones fueron anotadas por dos personas, con una versión inicial de la guía de funciones que se amplió considerablemente con los ejemplos y consideraciones resultado del análisis de las discrepancias (Civit, Bufí, y Valverde, 2004). A continuación se procedió a la anotación de las frases restantes.

2.2. Líneas de etiquetado sintáctico para dependencias

La anotación sintáctica del corpus Eus3LB se realizó siguiendo el modelo de la gramática de dependencias. A continuación detallamos la metodología definida:

- tres lingüistas anotaron 20 oraciones con la finalidad de definir el sistema de etiquetado. Como resultado de este proceso se establecieron los criterios de anotación que se recogieron en el *manual de anotación* (Aduriz et al., 2003).
- se seleccionaron 150 nuevas oraciones que fueron anotadas en paralelo por otros dos lingüistas haciendo uso del manual anteriormente descrito. Estas oraciones se caracterizan por tratarse de estructuras representativas del euskera. Esta tarea finalizó con una descripción completa del sistema de etiquetado (Aranzabe et al., 2003). Una vez completada la descripción, se procedió a su verificación repitiendo la anotación de las 150 oraciones. Se comprobó que el grado de acuerdo entre los anotadores era suficiente por lo que se pasó a la siguiente fase.
- anotación de las 50.000 palabras del corpus realizada por tres lingüistas. Inicialmente se llevó a cabo un etiquetado manual en el que surgieron nuevos problemas que fueron gradualmente solucionados. Con la finalidad de agilizar esta tarea y evitar errores en la transcripción de las dependencias y/o campos de las mismas, se diseñó y puso

en marcha la herramienta computacional 3LB *AbarHitz* (de Ilarraza, Garmendia, y Oronoz, 2004).

2.3. Datos

En el caso de Cast3LB se han anotado 100.000 palabras, que corresponden a unas 4.000 oraciones tanto a nivel de constituyentes como de funciones. Las oraciones contienen un promedio de 25 palabras. El tiempo de anotación para constituyentes era de 10 frases cada hora, mientras que para las funciones se anotaban 25 oraciones en el mismo tiempo.

Para Cat3LB se han anotado 106.000 palabras a nivel de constituyentes, correspondientes a unas 2.700 oraciones (la media de palabras por oración es de 39), de las cuales la mitad también se han anotado a nivel de funciones sintácticas. Los tiempos de anotación eran los mismos que para el castellano.

En el caso de Eus3LB, se han anotado a nivel sintáctico (dependencias) 56.000 palabras que suponen 3.708 oraciones (aproximadamente 15 palabras por oración) y el ritmo de anotación era de 6 frases a la hora.

2.4. Evaluación del grado de acuerdo

No existiendo medidas específicas para la comparación cuantitativa del acuerdo entre anotadores, se ha decidido usar las que se pueden considerar las primeras medidas objetivas (y más estandarizadas actualmente) para la evaluacion de gramáticas y/o métodos de análisis. Se trata de las métricas definidas en los workshops Parseval (Black et al., 1991), que comparan la similitud de los resultados obtenidos con los árboles de análisis de referencia (los previamente considerados *correctos* o *gold standard*). Estas medidas se basan en la comparación de los constituyentes de ambos árboles de análisis, en dos niveles: parentizado (considerando sólo los límites de los constituyentes) y etiquetado (considerando tanto los límites como su etiqueta). En concreto se ha utilizado la *precisión* (proporción de constituyentes planteados como hipótesis que son correctos), la *cobertura* (proporción de constituyentes correctos que son planteados como hipótesis) y la *cobertura de paréntesis consistentes* (proporción de constituyentes del árbol evaluado cuyos límites no se cruzan con los límites de los del árbol de referencia, **P.c.**).

Sin embargo, en nuestro caso no estamos evaluando la anotación proporcionada por un cierto método de análisis, sino comparando las anotaciones realizadas por dos linguistas, no existe un *gold standard*. Por ello hemos decidido comparar los análisis en los dos sentidos y considerar ambas medidas a la hora de calcular las medias: los conceptos de precisión y cobertura dejan pues de tener sentido, y se unifican en una sola medida de comparación, que denominaremos indistintamente precisión, etiquetada (**P.e.**) o parentizada (**P.p.**).

La evaluación para el castellano se efectuó en cinco fases a lo largo de las cuales se fueron resolviendo los problemas de desacuerdo, desde una primera fase en que se establecieron los principios básicos de la anotación, hasta una quinta fase correspondiente a los resultados considerados definitivos. La figura 1 muestra la evolución de las medidas a lo largo de estas cinco fases. Se observa que la precisión etiquetada llega a mejorar cerca de un 27 % desde la fase inicial a la final, la precisión parentizada en más de un 20 %, y la consistencia en el parentizado en casi un 15 %. Los criterios surgidos de la evaluación cualitativa de las desavenencias a lo largo de estas fases se han aplicado también al catalán, para el cual sólo se ha efectuado la evaluación de acuerdo final.

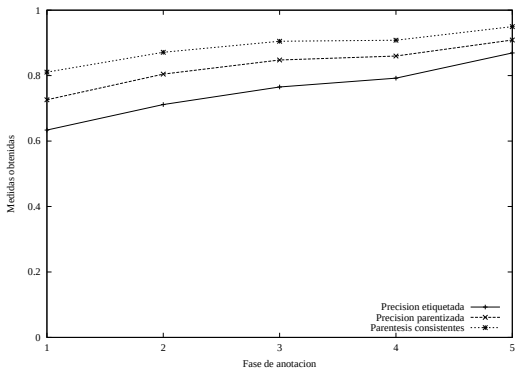

Figura 1: Evaluación de la anotación

Una de las principales discrepancias entre anotadores que apareció en las primeras fases de análisis fue la consideración como locuciones o no de estructuras complejas del tipo *desde que*, *dar lugar a*, etc., lo cual modificaba la longitud de las frases. Como esto afecta profundamente a nuestras medidas, éstas se han evaluado también sólo para aquellas frases cuyas longitudes son iguales. El cuadro 1 muestra pues todos los resulta-

dos finales obtenidos para el castellano y el catalán. El hecho de que los resultados para ambas lenguas sean asombrosamente similares, superando en el segundo caso holgadamente el 90 % de acuerdo, confirma que éste sea probablemente el límite en la consistencia de la anotación humana en el caso de estas lenguas.

	P.e.	P.p.	P.c
Todas las frases			
Castellano	0.86927	0.90889	0.94958
Catalán	0.87647	0.90953	0.94321
Frases de igual longitud			
Castellano	0.91529	0.94036	0.96985
Catalán	0.91981	0.93964	0.96512

Cuadro 1: Resultados finales para catalán y castellano

En euskera no se dispone de la misma cantidad de datos pero se han realizado también experimentos para evaluar el grado de coincidencias en la anotación y aunque a escala más pequeña, se han obtenido unos resultados muy similares a los del castellano y catalán.

2.5. Herramientas desarrolladas

En el marco del proyecto 3LB se han desarrollado diferentes herramientas de ayuda para el etiquetado sintáctico de los corpus, ya que se ha seguido el criterio de constituyentes para el castellano y catalán, y el de dependencias para el euskera. A continuación se presentan las características más relevantes de cada una de estas herramientas.

2.5.1. Herramienta de ayuda al etiquetado sintáctico para el castellano y catalán

Como herramienta de ayuda al etiquetado sintáctico (constituyentes y funciones) de los corpora, se ha adaptado el editor de árboles TreeTrans de AGTK versión 0.92 (Cotton y Bird, 2000). Esta adaptación ha consistido, por una parte, en la modificación del formato de entrada/salida que permite la entrada en formato PennTreebank (TBF o parentizado) y XML, y, por otra parte, en la extensión del etiquetado de las palabras con sentidos y correferencias. En la figura 2 se muestra un ejemplo de la herramienta para una frase del corpus.

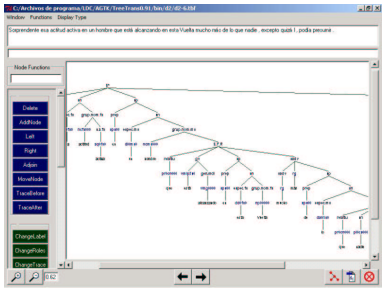

Figura 2: Ejemplo sobre la herramienta Treetrans

2.5.2. Herramienta de ayuda para el etiquetado sintáctico basado en dependencias: euskera

Para el caso del euskera se ha desarrollado 3LBAbarHitz (de Ilarraza, Garmendia, y Oronoz, 2004), una herramienta de ayuda para el etiquetado sintáctico del corpus basado en dependencias.

La utilización de 3LBAbarHitz facilita y agiliza la anotación sintáctica manual del corpus, y evita errores de anotación al controlar el número y tipo de campos descritos en cada etiqueta de dependencia. Una vez analizada, la frase es visualizada en forma de árbol, donde el anotador puede realizar distintos cambios: supresión de nodos, corrección, etc. Cuando decide que la representación es correcta, los cambios se reflejan en el análisis de la frase.

Está implementado en Java, es multiplataforma y las funciones y el idioma de la interfaz son fácilmente adaptables.

3. Etiquetado semántico

Con respecto a la anotación semántica, el principal objetivo del proyecto 3LB ha sido especificar el sentido de nombres, verbos y adjetivos en cada uno de los corpus implicados en el proyecto (siguiendo la aproximación denominada "*all words*"). Dado el carácter multilingüe del proyecto, la intención seguida en la anotación semántica ha sido buscar siempre la máxima compatibilidad entre los tres corpus. Por ello, como se expondrá en este apartado, tanto el recurso léxico utilizado para la anotación semántica como el método de trabajo seguido, así como la herramienta de anotación, son comunes a los tres corpus.

3.1. Líneas de etiquetado

Como recurso léxico para el inventario de sentidos para las tres lenguas se ha

utilizado la parte española de EuroWordNet (Vossen, 1998) así como sus extensiones para el catalán y el euskera. Con vistas a la compatibilidad entre las anotaciones se ha utilizado una versión congelada de las distintas redes (versión 1.5, diciembre 2002) Si bien no es un recurso exento de problemas, actualmente es el más utilizado en Procesamiento de Lenguaje Natural: ha sido el recurso de anotación semántica utilizado en otros corpus como SemCor (Miller, 1990) y el corpus DSO (Ng y Lee, 1996), además de ser el estándar de *Senseval*, la competición sobre desambiguación semántica automática.

EuroWordNet (de ahora en adelante EWN) presenta, además, una característica que lo hace especialmente útil para el proyecto 3LB: se trata de un recurso multilingüe. Efectivamente, en EWN están implicadas diferentes lenguas, entre ellas el catalán, el euskera y el castellano. Mediante el *Índice Interlingua*, la representación de los sentidos es la misma para todas las lenguas de EWN. Por lo tanto, los tres corpus del proyecto 3LB tienen representado el sentido de las palabras de la misma manera. Así se obtienen no sólo tres corpus etiquetados con información semántica, sino tres corpus semánticamente compatibles al estar etiquetados con el mismo recurso.

Sin embargo, no se han podido anotar todos los nombres, verbos y adjetivos, porque EWN es un recurso limitado y no contiene ni todas las palabras de cada lengua ni todos los sentidos posibles. Para marcar estos casos se han propuesto dos etiquetas nuevas, mediante las cuales se marca o bien la carencia de un sentido para una palabra concreta, o bien directamente la carencia de una palabra [1].

Sobre el método del etiquetado semántico, se pueden distinguir dos aproximaciones generales: un método lineal o "textual" (Kilgarriff, 1998), en el que el anotador etiqueta palabra tras palabra siguiendo el orden de las oraciones del corpus; o un método transversal o "léxico" (Kilgarriff, 1998), en el que el anotador etiqueta primero todas las ocurrencias en el corpus de una palabra, luego todas las ocurrencias de otra palabra, y así sucesivamente hasta finalizar el inventario total de palabras que aparecen en el corpus.

En el proyecto 3LB hemos adoptado esta segunda estrategia transversal o léxica. Desde el punto de vista del anotador, con este método, el estudio de los problemas semánticos de cada palabra se simplifica ya que se realiza una sola vez. Así, el trabajo del anotador se centra en ir contrastando la palabra en cada uno de los contextos en los que aparece en el corpus y seleccionar el sentido correcto.

Desde el punto de vista del resultado, con este método se obtiene una anotación semántica más coherente y consistente. Dado que el responsable de anotar una palabra a lo largo de todo el corpus es siempre el mismo, se mantienen exactamente los mismos criterios de anotación para todos los contextos de aparición.

Este método tiene el inconveniente de que no se obtiene una muestra del corpus etiquetado hasta que no se ha finalizado todo el proceso. Con todo, en el proyecto sólo se ha procedido al etiquetado semántico de una parte del corpus [2].

La anotación semántica de corpus es una tarea especialmente compleja. Una de las razones fundamentales es el carácter subjetivo que tiene el propio proceso de anotación, así como la segmentación del continuum del significado en unidades discretas: los diferentes sentidos o acepciones de la fuente léxica. EWN, además, se caracteriza por un grado de granularidad elevado y, a la vez, por la falta de representación de algunos sentidos básicos.

Para solventar en la medida de lo posible este problema, en el proyecto 3LB, y siguiendo el mismo procedimiento que en la anotación sintáctica, hemos dividido el proceso de anotación semántica en dos fases. En la primera fase, se ha seleccionado y etiquetado por dos anotadores un subconjunto de palabras ambiguas de diferente categoría gramatical. A partir de la comparación de esta doble anotación se han especificado los casos concretos de desacuerdo entre los anotadores y se ha elaborado una tipología de desacuerdos. Los desacuerdos se debían en la mayoría de los casos a cuestiones de ambigüedad, por lo que se ha desarrollado una guía de anotación en la que se han especificado los criterios a seguir en cada caso (Navarro et al., 2004). Como el propósito del proyecto 3LB es desarrollar una anotación semántica lo más coherente y compatible posible entre los tres corpus, todos estos criterios se han acordado

[1] Cf. sección 3.3.

[2] Cf. sección 3.2.

entre los tres equipos de anotación.

En la segunda fase se ha etiquetado el resto del corpus siguiendo los criterios de anotación semántica acordados. Estos criterios se basan en tres puntos fundamentales:

- en general, se procurará poner sólo un sentido por cada palabra;
- ante casos de duda entre dos posibles sentidos, se tenderá a anotar el sentido más general;
- únicamente en aquellos contextos en los que se vea muy claro que es posible asignar dos (o más) sentidos a la misma palabra, se asignará más de uno.

En el caso del euskera, la metodología seguida ha sido la siguiente: El corpus de partida es mayor que Eus3LB (este nuevo corpus contiene 300.000 palabras correspondientes al corpus del proyecto HIZKING21[3] en las que se encuentran incluidas las que corresponden a Eus3LB). En el etiquetado han tomado parte dos anotadores, un árbitro y un editor de EusWN (la necesidad del editor del EusWN viene dada por la oportunidad de mejorar este recurso según las necesidades de los anotadores). Se han etiquetado todas las ocurrencias de una palabra de una sola vez presentes en el corpus más amplio; las palabras a anotar se van seleccionando de entre las más frecuentes. Previamente se ha comprobado que las acepciones presentes en EusWN son correctas y completas, y en su caso se ha modificado EusWN para que así sea. Los anotadores han afrontado por separado y en paralelo el etiquetado. Al finalizar, se ha confeccionado automáticamente un informe para el árbitro, que verifica el correcto etiquetado fijándose especialmente en las discrepancias. El árbitro puede pedir al editor de EusWN que añada nuevos sentidos, en cuyo caso hacer las correcciones necesarias en la anotación. Se ha creado software adicional que asiste a los anotadores y árbitros en su tarea. A la vista de los resultados y del tiempo dedicado a la obtención de los mismos, nos hemos planteado un cambio de metodología que supondría que los anotadores trabajarán por separado textos diferentes, haciéndose posteriormente una verificación de una muestra de lo anotado.

[3]Convocatoria ETORTEK de Investigación estratégica del Gobierno Vasco

3.2. Datos

El total de palabras que se han de etiquetar con información semántica en el corpus Cast3LB son 42.291, de las cuales 20.461 son nombres, 13.471 son verbos y 8.543 adjetivos. En lo referente a Cat3LB, se han etiquetado, sobre el 10 % del corpus, 2.379 formas nominales, correspondientes a 839 nombres distintos; 1.225 formas verbales, equivalentes a 401 verbos distintos y 813 apariciones de adjetivos calificativos, que se corresponden a 377 formas distintas de adjetivos. Extrapolando los resultados conseguidos para el corpus de 300.000 palabras que hemos mencionado antes, alrededor del 10 % de las palabras de Eus3LB han sido etiquetadas; se han etiquetado por lo tanto alrededor de 5.000 formas correspondientes a 75 palabras diferentes en el que se incluyen sustantivos, verbos, adjetivos.

3.3. Herramientas

Se convino la creación de una herramienta orientada a esta tarea de anotación transversal o léxica, porque este recorrido no secuencial mejora sustancialmente el tiempo y esfuerzo invertido en la anotación. Posteriormente, un recorrido secuencial del texto anotado semánticamente, frase por frase, mediante la herramienta TreeTrans que, como ya se ha comentado, ha sido modificada para que acepte este tipo de anotación, permite la supervisión del mismo.

Así es como nace la herramienta 3LB-SAT (3LB-Semantic Annotation Tool) (Bisbal et al., 2003). Sus principales características son que está orientado a la palabra (o token), que permite introducir el corpus en diferentes formatos (TBF y XML) y que usa EWN para consultar el sentido de las palabras ya que se dispone de este lexicón para las tres lenguas objeto del proyecto (castellano, catalán, euskera)[4]. De forma simplificada, este diccionario consiste en conjuntos de sinónimos (synsets) que agrupan los sentidos de distintas palabras asociados a un único concepto. Cada uno de estos synsets tiene asociado un identificador único utilizado para anotar los sentidos con la herramienta. En el caso de cambio de identificador entre las distintas versiones de WordNet, una correspondencia básica entre versiones permite la adaptación del etiquetado. Se han

[4]La herramienta permite también el etiquetado semántico en inglés.

identificado dos posibles tipos de carencias de EuroWordNet que se desea que también sean anotadas: (a) La no existencia del sentido que representa la palabra dentro de la oración (etiquetado como C1S); (b) La no existencia de la palabra en EWN (etiquetado como C2S). Esta anotación especial nos permitirá enriquecer el diccionario con nuevos sentidos para las palabras existentes o ampliarlo a nuevas palabras.

Respecto al funcionamiento de la herramienta, durante la apertura de un fichero se anotan de forma automática todas las palabras monosémicas (aunque éstas han de ser supervisadas por si este sentido no fuera apropiado, caso que debería anotarse con la etiqueta C1S) y aquellas que no se han encontrado en EuroWordNet, a las que se les asigna la etiqueta C2S como ya se ha comentado.

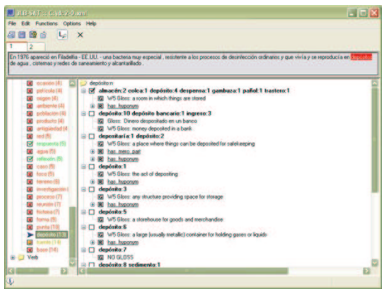

Figura 3: Herramienta 3LB-SAT

Una vez cargado el corpus, la herramienta muestra en la parte izquierda todos los lemas siguiendo un código de colores para indicar que no se ha anotado ninguna aparición del lema en el corpus (rojo), que se han anotado algunas de sus apariciones, pero no todas (naranja), o bien que todas sus apariciones han sido anotadas (verde). Además, los lemas se muestran por categorías en el orden seleccionado: ascendente/descendente por orden polisémico o por orden alfabético. Cuando se selecciona un lema, se van mostrando todas sus apariciones en la parte superior de la ventana. Si se selecciona una de ellas, se muestran todos los posibles sentidos del lema (para cada synset se pueden consultar todos los sinónimos, la glosa del propio idioma, la glosa inglesa, los hipónimos y los hiperónimos de primer nivel). Una vez se ha seleccionado una de las apariciones del lema, se anota(n) su(s) sentido(s). En la Figura 3 podemos ver que el lema *depósito* aparece dos veces en el texto y tiene trece sentidos posibles. Durante el proceso de anotación la herramienta crea un informe relativo a los cambios efectuados sobre un fichero del corpus. Esta información permite la obtención de estadísticas, comparar el proceso de anotación utilizando métodos automáticos de desambiguación o sin ellos, además de poder realizar un seguimiento del sistema.

4. *Conclusiones y líneas futuras*

Una vez finalizado el proyecto 3LB se dispone de un corpus etiquetado sintácticamente para el catalán (Cat3LB), castellano (Cast3LB) y euskera (Eus3LB). El corpus del catalán y el del castellano constan de más de 100.000 palabras etiquetadas sintácticamente a nivel de constituyentes. Cast3LB está también totalmente anotado a nivel de funciones, mientras que Cat3LB sólo tiene este nivel de anotación para 50.000 palabras. Eus3LB consta de 56.000 palabras anotadas mediante dependencias sintácticas. La anotación semántica de Cast3lb se ha completado al 100 %, mientras que el del catalán y el euskera se han anotado al 10 %.

Además de disponer de un nuevo recurso de ingeniería lingüística para estas lenguas, cabe destacar el interés que representa disponer de manuales de anotación donde se detallan los criterios que se han seguido para el etiquetado y que pueden servir de referencia para trabajos futuros.

Con el fin de garantizar la calidad del etiquetado se ha definido una metodología consistente en la anotación paralela y posterior comparación de fragmentos del corpus. Gracias a este método se ha podido llegar a niveles altos de consistencia (94 %).

En cuanto al trabajo futuro, está previsto ampliar los corpus a 500.000 palabras en las tres lenguas de 3LB e incorporar el gallego. El objetivo sería un etiquetado completo tanto sintáctico como semántico.

Bibliografía

Aduriz, I., I. Aldezabal, M.J. Aranzabe, B. Arrieta, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, K. Gojenola, M. Maritxalar, M. Oronoz, y K. Sarasola. 2003. Corpusaren etiketatze sintaktikoa analizatzailea eraikitzeko. Informe Técnico PV/EHU/LSI/TR 1-2003, University of the Basque Country. Languages and Informatic Systems.

Aranzabe, M.J., B. Arrieta, J.M. Arriola, A. Atutxa, I. Balza, y L. Uria. 2003. Guía para la anotación sintáctica manual de eus3lb (corpus del euskera anotado a nivel sintáctico, semántico y pragmático. Informe Técnico PV/EHU/LSI/TR 13-2003, University of the Basque Country. Languages and Informatic Systems.

Atserias, J. y H. Rodríguez. 1998. TACAT: TAgged Corpus Text Analyzer. Informe técnico, Software Department (LSI). Technical University of Catalonia (UPC).

Bisbal, E., A. Molina, L. Moreno, F. Pla, M. Saiz-Noeda, y E. Sanchís. 2003. 3LB-SAT: Una herramienta de anotación semántica. *Procesamiento de Lenguaje Natural, SEPLN*, (31).

Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Kl avans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, y T. Strzal kowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. En *Proceedings of the Speech and Natural Language Workshop*, páginas 306–311, Pacific Grove, CA. DARPA.

Civit, M. 2003a. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español.* Numero 3 en Colección de monografías.

Civit, M. 2003b. Guía para la anotación de las funciones sintácticas de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática. Informe Técnico X-Tract-II WP-03/02, 3LB WP 03-04, Universitat de Barcelona. disponible: http://clic.fil.ub.es/personal/civit.

Civit, M. 2003c. Guía para la anotación sintáctica de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática. Informe Técnico X-Tract-II WP-02/01, 3LB WP 02-01, Universitat de Barcelona. disponible: http://clic.fil.ub.es/personal/civit.

Civit, M., N. Bufí, y M.P. Valverde. 2004. Guia per a l'anotació de les funcions sintàctiques de Cat3LB: un corpus del català amb anotació sintàctica, semàntica y pragmàtica. Informe Técnico X-Tract-II WP-03/02, 3LB WP 03-10, Universitat de Barcelona. disponible: http://clic.fil.ub.es/personal/civit.

Cotton, S. y S. Bird. 2000. An integrated framework for treebanks and multilayer annotations. En *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

de Ilarraza, A. Díaz, A. Garmendia, y M. Oronoz. 2004. Abar-Hitz: An Annotation Tool for the Basque Dependency Treebank. forthcoming. En *Proceedings of the International Conference on Language Resources and Evaluation (LREC'04).*

Kilgarriff, A. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language. Special Use on Evaluation*, 12(4):453–472.

Miller, G. A. 1990. Wordnet: An on-line lexical database. *Intenational Journal of Lexicography*, 3(4):235–312.

Navarro, B., M. Civit, R. Marcos, B. Fernández N. Bufí, E. Pociello, y M.P. Valverde. 2004. Guía para la anotación semántica del corpus3LB. Informe técnico, Universitat de Alicante. disponible: http://gplsi.dlsi.ua.es:9998.

Ng, H. T. y H. B. Lee. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. En *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California.

Valverde, M.P., M. Civit, y N. Bufí. 2004. Guia per a l'anotació sintàctica de Cat3LB: un corpus del català amb anotació sintàctica, semàntica i pragmàtica. Informe Técnico X-Tract-II WP-03/01, 3LB WP 03-09, Universitat de Barcelona. disponible: http://clic.fil.ub.es/personal/civit.

Vossen, Piek. 1998. *A Multilingual Database with Lexical Networks.* Kluwer Academic Publishers.