

La construcción del WordNet 3.0 en español

ANA FERNÁNDEZ MONTRAVETA

Universitat Autònoma de Barcelona

GLORIA VÁZQUEZ

Universitat de Lleida

0. INTRODUCCIÓN

En este artículo presentamos los resultados de un proyecto¹ en el marco del cual se ha creado un nuevo recurso léxico para el español a partir de la adaptación de la red semántica WordNet (WN), ya existente para la lengua inglesa, en su versión 3.0.

En este recurso las unidades básicas, los llamados *synsets*, son conjuntos de palabras sinónimas que se presentan relacionadas entre si a través de la hiperonimia, por lo que el resultado final es una red semántica (Fellbaum 1998, Miller et al. 1990). Para cada *synset* se aporta la definición, compartida por los diferentes miembros del grupo de sinónimos, y, en algunos casos, se presentan también ejemplos de uso de algunos de ellos.

Este recurso, como es sabido, es un referente mundial en el campo de la lexicografía, tanto por el alcance de la macroestructura como por la estructura de red que lo caracteriza. El impacto de dicha herramienta ha sido espectacular en el campo del procesamiento del lenguaje natural (PLN), donde se utiliza como un estándar para la desambiguación semántica automática a nivel de palabra.

La novedad que presenta la versión 3.0 es que el corpus que constituyen las definiciones y los ejemplos está etiquetado a nivel morfosintáctico y semántico a nivel de palabra. Al confeccionar la versión española consideramos una oportunidad el poder construir, además de un léxico de gran alcance organizado jerárquicamente, un corpus paralelo inglés-español anotado tanto morfosintácticamente como semánticamente. Hay que tener en cuenta que, por un lado, la creación de corpus anotados, especialmente a nivel semántico, es una ardua tarea, por lo que consideramos que valía la pena reutilizar recursos ya existentes, como ya se ha hecho en otros proyectos de anotación de corpus como MultiSemCor (Ranieri et al. 2004 y Bentivogli et al. 2005). Por otro lado, existen diversos

¹ Ministerio de Educación y Ciencia (HUM2006-27968-E)

corpus paralelos inglés-español,² pero, a nuestro conocimiento, si incluyen algún tipo de anotación es sólo a nivel morfosintáctico, pero no semántico.

Al decidir confeccionar también el corpus paralelo anotado, se hizo necesario partir de las definiciones y los ejemplos del inglés³ e intentar traducir las oraciones de la forma más fiel posible, ya que cuanto mayor número de coincidencias hubiera en las estructuras de las oraciones, mayor número de enlaces entre las palabras del inglés y el español se podrían establecer, lo cual permitiría conservar también las etiquetas de la anotación.

Dichas etiquetas se han mantenido o cambiado en función de las necesidades que se han ido detectando. En el caso de la anotación morfosintáctica se han mantenido excepto cuando la categoría de la palabra traducida divergía, en cuyo caso hemos incorporado la etiqueta correspondiente, siguiendo las especificaciones propuestas en el corpus de Wall Street Journal, que es el conjunto de etiquetas utilizado en el proyecto inglés. En cuanto a la anotación semántica se han mantenido también los identificadores usados en dicho recurso (que son los del propio WN), siempre que la traducción no implicara una alteración de las características morfológicas, como un cambio de categoría o de la estructura interna de la palabra (v. ap. 2).

Actualmente el proyecto ha finalizado y se han traducido aproximadamente unas 15.000 glosas,⁴ lo cual quiere decir que están disponibles para el español aproximadamente unas 30.000 entradas léxicas (nominales y verbales). El recurso aquí presentado está actualmente disponible en la red (v. ap. 3) y se puede consultar y utilizar de manera gratuita con fines de investigación.

En este artículo vamos a presentar el proceso de creación del recurso. En primer lugar (ap. 1) vamos a describir brevemente la interfaz que se ha confeccionado para la creación del WordNet español (WNE). A continuación ilustraremos con algunos ejemplos la tipología de problemas de traducción que se han abordado en este proyecto (ap. 2). Por último, describimos la interfaz de consulta de la herramienta (ap. 3) y las conclusiones (ap. 4).

1. LA INTERFAZ DE TRADUCCIÓN Y PARALELIZACIÓN

² Algunos ejemplos de corpus paralelos inglés-español son los siguientes: el Corpus CRATER (McEnery et al. 1997), el Corpus Trilingüe Inglés-Español-Catalán GRIAL (Castellón 2005), el Corpus ACREL (Ramon 2004) y un corpus de ficción con alineación a nivel de párrafo (Gelbukh et al. 2006).

³ Nos gustaría mostrar nuestro agradecimiento a la Dra. Fellbaum por haber puesto a nuestro alcance el material que nos ha permitido desarrollar el presente trabajo.

⁴ Esta cifra se corresponde con la mitad de las glosas disponibles actualmente para el inglés.

Como ya se ha avanzado, para la creación del WNE 3.0, se ha partido del recurso en inglés. Algunos de los datos contenidos en este léxico, como las etiquetas de la anotación morfosintáctica y semántica, han sido extraídos de dicho recurso e incorporados directamente al léxico español y otros, como las palabras de las entradas, las definiciones y los ejemplos, han sido traducidos manteniendo un enlace con el original, con el fin de obtener un corpus paralelo inglés-español.

Así pues, una vez cedido el recurso inglés por la University of Princeton, se procedió a la creación de una base de datos que mantuviera la estructura y los datos del recurso original con los campos duplicados para introducir las correspondientes traducciones. En la interfaz usada para llevar a cabo esta tarea se visualiza la información completa (anotación incluida) del recurso inglés con el fin de establecer el mayor número posible de enlaces entre las palabras del texto original y traducido. Así pues, siempre que fuera posible, se debía reflejar la misma estructura de la frase. Además, también se hacía preciso prever una forma de codificar los casos en que la traducción entre la lengua origen y destino presentaba desajustes necesariamente (v. ap. 2).

En la figura 1 presentamos el aspecto de dicha interfaz. Como puede observarse, la información se divide en cuatro secciones. En la primera sección se presenta toda la información de carácter más general, como el identificador de la entrada en nuestra base de datos y el identificador del WN inglés (SK – *sense key*) de la palabra de la macroestructura. Este identificador contiene en primer lugar la palabra en cuestión (en este caso, *flying colors*), seguida del símbolo ‘%’ y de 3 dígitos separados por “:”. El primer dígito (1) indica la categoría morfológica, el segundo (04), la clase semántica a la que pertenece y el tercero (00), el número de sentido que se corresponde a esa forma. El símbolo final es “:”.

Asimismo, en esta primera sección se muestra también el campo en el que se escribirá el equivalente español (“Traducción”), el nombre del traductor y el estado de la traducción (hecho, problemático, para hacer o validado). La metodología que se ha seguido a este respecto es la siguiente: en primer lugar, un traductor propone una traducción, de forma que se almacena la información con la etiqueta “hecho”; en segundo lugar, dicha traducción es revisada por otro traductor, con lo que la entrada léxica queda “validada”. Si se da algún problema de falta de correspondencia cultural o lingüística se marca la glosa como “problemática” para que sea posteriormente revisada.

En las secciones segunda y tercera, se muestran en formato vertical las definiciones y los ejemplos, si los hubiera, respectivamente. Este formato vertical facilita la paralelización

de los dos corpus a nivel de palabra. A continuación vamos a describir el contenido de estas secciones.

En la columna “English” se muestran las palabras y signos de puntuación que forman las definiciones y los ejemplos. Todos estos elementos tienen asociado un identificador (visible en la primera columna) que hemos incorporado en nuestro proyecto para poder generar las oraciones del español y establecer de forma adecuada la paralelización de los corpus entre las dos lenguas.

Si una palabra está semánticamente anotada en el corpus inglés, dicha anotación se muestra en la columna de la izquierda, denominada como SK (*sense key*), con el identificador del sentido de WN asignado a dicha palabra. En el caso representado en la figura, todas las palabras de la definición inglesa incorporan esta información; por ejemplo “complete” está etiquetado con los identificadores “complete%3:00:00” y “éxito” con “success%1:04:00”.

SECCIÓN 1		SKey	Traducción	Desc	Traductor	Validado
34		flying_colours%1:04:00::	éxito total	wid	Ana	Validado

SECCIÓN 2									
Definición:									
ID	SK	English	Info Extra	Español	Info Extra	Orden	Multipalabra	POS	Dif. Cat.
508	complete%3:00:00::	complete		absoluta		2		JJ	[seleccione una categoría]
509	success%1:04:00::	success		consecución		1		NN	[seleccione una categoría]
510	.	:	:	:		2		:	[seleccione una categoría]

SECCIÓN 3									
Ejemplos o contextos:									
ID	SK	English	Info Extra	Español	Info Extra	Orden	Multipalabra	POS	Dif. Cat.
511	.	they				0			[seleccione una categoría]
512	.	passed		pasaron		0			[seleccione una categoría]
513	.	inspection	la	inspección		0			[seleccione una categoría]
514	.	with	con			0			[seleccione una categoría]
515	.	flying	un	éxito total		0	-515-517-		[seleccione una categoría]
517	.	colors				0			[seleccione una categoría]

SECCIÓN 4			
ID	ID Symet	Palabra	Traducción
69819	65418	flying_colours%1:04:00::	éxito total
69820	65418	flying_colours%1:04:00::	éxito total

Fig. 1. La interfaz de traducción y paralelización.

En la columna “Español”, se introduce solamente la traducción de las palabras que presentan correspondencias directas con las palabras del inglés.

El campo “Info extra” se utiliza cuando en español se da un cambio de estructura respecto a las frases en inglés y se usan elementos que no tienen un correlato directo en esta lengua. Esto es lo que ocurre en el ejemplo asociado a la palabra que estamos tratando, donde los artículos “la” y “un” se han añadido a la estructura española y no tienen una equivalencia directa en la oración inglesa.

En el caso de que se mantenga la correspondencia entre palabras pero no su orden de aparición en la frase, se codifican los cambios en la columna “Orden”. El valor de este campo es numérico y expresa el orden en el cual las palabras traducidas se han de mostrar en la frase española. Tal y como puede observarse en la figura, la disposición de las palabras que componen la definición es diferente en las dos lenguas. En inglés, el adjetivo siempre precede al nombre mientras que en español normalmente es el caso contrario. En la oración usada en el ejemplo se mantiene el mismo orden, por lo que no se ha especificado ninguna numeración.

Por lo que respecta al campo “Multipalabra”, se utiliza cuando es necesario realizar la correspondencia entre más de una palabra en inglés a una o más palabras en español. En la oración que aparece en el ejemplo de la entrada, la expresión “flying colors” no aparece en una sola casilla, sino que cada una de las palabras que componen dicha expresión se presenta separadamente. En este caso, para la traducción al español también se usa una expresión formada por dos términos (“éxito total”). Ahora bien, como el primer y el segundo término de la expresión inglesa no se corresponden con el primer y segundo término de la expresión española, ya que en el primer caso el significado no se obtiene de forma composicional, el SN del español se incorpora en una sola casilla y se asocia a las dos palabras del inglés, cuyos números de identificación aparecen en el campo correspondiente de la columna “Multipalabra”.

La anotación morfológica está codificada en la columna marcada con la etiqueta POS. Siguiendo con el ejemplo de la figura 1, “complete” está etiquetado como adjetivo (JJ) y “success” como nombre singular (NN). En ocasiones se ha tenido que cambiar la categoría de las palabras al traducirlas y para ello se ha creado la casilla (Dif. Cat. –diferente categoría). En el ejemplo que estamos examinando esta opción no ha sido necesaria pero más adelante examinaremos otros ejemplos en los que sí se ha efectuado el cambio de la categoría.

Finalmente, en la cuarta sección de la interfaz, se observa que se han incorporado también las diferentes variantes que están conectadas a la glosa y que forman parte del mismo *synset*. En este caso, como puede observarse en la columna “Palabra”, las dos variantes contempladas son de hecho dos posibles ortografías en inglés del mismo término, una con “o” (*flying colors*) y la otra con “ou” (*flying colours*). En otras ocasiones, puede tratarse de sinónimos con formas totalmente diferentes. En el caso del español, como no existe otra variante de la traducción, ambos términos del inglés se han traducido de igual forma, como se observa en la columna denominada “Traducción”. En el caso de que para el español hubiera un número mayor de variantes que en inglés, éstas se indicarían en el campo “Sinónimos”.

Por último, en las dos primeras columnas de la cuarta sección se incluyen los identificadores de las palabras que forman la macroestructura del diccionario bilingüe: en la primera columna (“ID”), dicho identificador se usa para numerar las diferentes entradas de la parte inglesa del diccionario y sus correspondencias al español y, en la segunda columna, se incluye el identificador del *synset* de WN, que es el mismo para todas las variantes del inglés relacionadas.

2. PROBLEMAS EN LA PARALELIZACIÓN

En la primera fase del proyecto se automatizó la traducción, para lo cual se creó un programa que traducía los términos del inglés al español mediante un léxico bilingüe. La calidad de la traducción resultante era pobre y se requería un extenso proceso de edición manual, lo que nos obligó a abandonar nuestra idea inicial de crear un corpus traducido automáticamente parte del cual estuviera corregido.

A continuación vamos a repasar brevemente algunos de los problemas más comunes que se han encontrado en el proceso de creación del corpus. Se trata de desajustes de traducción que tienen un reflejo o implicación en la alineación de la estructura.

2.1 Palabras funcionales

Uno de los problemas más comunes que se han encontrado consiste en que, a menudo, en una de las lenguas se necesitan palabras funcionales que no están presentes en la otra. Estos elementos han quedado sin enlazar con ninguna palabra en la lengua origen pero se ha marcado la posición indicando su orden entre los elementos de la oración.

En su mayoría se han tratado problemas relacionados con el diferente uso de los determinantes en las dos lenguas. Como puede apreciarse en los ejemplos que mostramos a continuación, este tipo de desajuste también se ha dado en el uso de los adjetivos posesivos (*su*, 1) y de preposiciones (*de*, 2). En ambos casos, dichas partículas son necesarias en español pero no en inglés.

(1) Transporting alcoholic liquor for sale illegally

Transporting (transportar) alcoholic (alcohólicas) liquor (bebidas) for (para) sale (venta) illegally (ilegal)

‘Transportar bebidas alcohólicas para *su* venta ilegal ‘

(2) Pocket-sized paperback book

Pocket-sized (tamaño de bolsillo) paperback (tapa blanda) book (libro)
'Libro *de* tapa blanda de tamaño de bolsillo'

2.2 *Problemas en el orden de las palabras*

El orden de la expresión de las palabras es otro de los más comunes que nos hemos encontrado en la construcción del corpus paralelo, y afecta sobre todo la secuencia formada por un adjetivo seguido de un nombre.

(3) The experiencing of *affective and emotional states*.

The (la) experiencing (experimentación) of (de) affective (afectivos) and (y) emocional (emocionales) status (estados).

'La experimentación de *estados afectivos y emocionales*'

En otras ocasiones, las cuestiones de orden son mucho más complicadas de resolver ya que son consecuencia de diferentes estructuras en ambas lenguas, como veremos.

2.3 *Expresiones multi-palabra*

Como ya hemos comentado, el nivel en el que se ha establecido la alineación es la palabra ya que nuestro objetivo ha sido reaprovechar la anotación del corpus inglés. A veces, la equivalencia a este nivel no ha sido posible y se ha hecho necesario establecer vínculos desde una expresión en inglés a una palabra en español, como en (4), o a la inversa (5). En otras ocasiones la correspondencia en ambas lenguas es a nivel de expresiones, como hemos visto en la oración del ejemplo de la figura 1.

(4) There was *too much* for a single person to do.

'Había *demasiado* que hacer para una persona sola'

(5) The *biggest* overturn since David beat Goliath

'El resultado *más sorprendente* desde que David ganó a Goliath'

Un problema de la misma índole que el ejemplificado en (4) es el de los pronombres clíticos cuando están conectados gráficamente con los verbos en español (infinitivo o gerundio). En estas ocasiones, para realizar la alineación se ha usado el mecanismo descrito en este apartado, es decir, se han relacionado dos palabras del inglés (el pronombre y el

verbo) con una sola palabra en español, la de la forma verbal que incorpora el clítico. Ahora bien, queda pendiente realizar un proceso de análisis morfológico posterior que analice la forma del español como una unidad compleja formada por un verbo y un pronombre.

2.4 Diferentes requerimientos gramaticales

Si tenemos en cuenta que el corpus paralelo que se ha creado se trata de un corpus de definiciones y ejemplos de uso de las palabras definidas, se hace evidente que las estructuras gramaticales a traducir son limitadas. Sin embargo, se han encontrado algunos desajustes en este nivel que vamos a revisar brevemente en este apartado.

Un tipo de desajuste muy común es el uso del gerundio en inglés en contraposición al uso del infinitivo en español (Izquierdo 2005), como se observa a continuación:

- (6) A trap for *catching* rats
'Trampa para *atrapar* ratones'

Los cambios estructurales entre ambas lenguas pueden ser más complejos al traducir una construcción de gerundio del inglés al español cuando no hay una correspondencia total a nivel de palabra, y cada lengua requiere de una construcción gramatical determinada. Por ejemplo, en vez de una construcción de infinitivo puede utilizarse una construcción subordinada en español:

- (7) The Prohibition amendment made *bootlegging* profitable
The (La) Prohibition amendment (Ley Seca) made (hizo) bootlegging
(contrabando) profitable (rentable)
'La Ley Seca hizo *que el contrabando fuese rentable*'

En estas ocasiones mantenemos la alineación de cada palabra que puede conectarse aunque su función y categoría sea diferente en cada una de las construcciones gramaticales. Cuando ello ocurre, se han codificado los cambios en la categoría gramatical, como es el caso del gerundio "bootlegging", que en español, en la traducción que hemos escogido, se expresa mediante un nombre singular ("contrabando").

En el campo "Extra-information" se han incluido las palabras adicionales que se han necesitado en la lengua española, como "que", "el" y "fuese", que no se han enlazado con ninguna palabra de la estructura inglesa.

A continuación vamos a examinar otro ejemplo de desajuste gramatical entre el inglés y el español en (8). En este último caso, se requiere un verbo español (“realiza”) en vez de un nombre (“maker”) y, en consecuencia, la estructura sintáctica resultante es bastante diferente entre las dos lenguas, puesto que el significado expresado en inglés mediante un adjetivo posesivo, en español es expresado mediante el sujeto (“que”). Además, el verbo “realiza” necesita un objeto (“lo”) que lo complemente.

(8) A miscalculation that recoils in *its maker*

A(Un) miscalculation (error de cálculo) that (que) recoils (afecta) in (a) its (su) maker (realizador)

‘Un error de cálculo que afecta al *que lo realiza*’

2.5 Falta de correspondencia léxica

En ocasiones, los *synsets* traducidos pertenecen a una realidad cultural que no necesariamente tienen un reflejo en la cultura española. No olvidemos que WN ha sido creado y diseñado en una universidad americana y, por tanto, refleja especialmente los conceptos pertenecientes a esta sociedad. En estos casos ha sido imposible una traducción literal y se ha optado por mantener la estructura en la medida de lo posible pero explicando el concepto definido, como ocurre en (9):

(9) He came all the way around on William’s hit

He (-) came (llegó) all the way around (a la meta) on (gracias al) William (William) ’s (de) hit (golpe)

‘Llegó a la meta gracias al golpe de William’

Este ejemplo pertenece al dominio del béisbol. Se trata de un deporte no muy habitual en España y, por lo tanto, las reglas del juego o la terminología del mismo suelen ser desconocidas para sus habitantes. Así, hemos optado por parafrasear el ejemplo de manera que se exprese el significado usando términos de un dominio más general. De esta manera, “llegar a la meta” es más general que “come all the way around” pero el concepto es el mismo: llegar a un punto que es el objetivo. En este sentido hemos explicado, no traducido, el significado en la medida de lo posible, manteniendo sin embargo la alineación con la anotación semántica (“hit” – “golpe” - SK: hit%1:04:03:).).

Otros ejemplos de este tipo de desajuste se encuentran en las oraciones en las que aparecen verbos y nombres deverbales que expresan manera en inglés, como “smack” (“beso

sonoro”). En este caso, los hablantes del inglés usan una palabra para designar un concepto que o no está categorizado para los hablantes del español o, si lo estuviera, no tiene asociado ningún término en esta lengua.

3. LA INTERFAZ DE CONSULTA DEL WNE 3.0

En esta sección vamos a describir la interfaz de consulta de la herramienta lexicográfica creada y que está accesible en <http://grial.uab.es/recursos/wordnet30.php>.

Como puede verse en la figura 2, en primer lugar, el usuario indica la palabra que desea consultar, que puede ser del español o del inglés. Así, se permiten búsquedas de términos ingleses o bien de términos españoles. Hay que advertir que los datos que se visualizan serán los mismos en ambos casos ya que siempre se muestra la información para ambas lenguas.

En segundo lugar, a través de un menú desplegable la herramienta nos permite visualizar más o menos información del contenido de la entrada en función de las necesidades del usuario que realice la consulta. Las opciones son las siguientes:

- a) Mostrar las variantes (sinónimos).
- b) Mostrar las definiciones.
- c) Mostrar los ejemplos.
- d) Mostrar las categorías morfosintácticas de las palabras.⁵
- e) Mostrar los lemas de las palabras.⁶
- f) Mostrar los identificadores del sentido de WordNet de las palabras de los ejemplos.
- g) Mostrar los identificadores del sentido de WordNet de las palabras de las definiciones.

Las tres primeras opciones se usan para visualizar más o menos información de la microestructura. Las otras cuatro opciones se refieren a la anotación de las palabras del corpus, ya sea morfosintáctica (d, e) o semántica (f, g).

En el ejemplo de la figura 2 se presentan los dos primeros resultados, es decir, los dos primeros sentidos asociados a la palabra consultada (“golpe”). Como se puede observar, en el resultado de cualquier búsqueda siempre se incluye un localizador interno de la base de datos

⁵ Esta información sólo está accesible en la actualidad para las definiciones de ambas lenguas, pero no para los ejemplos.

⁶ Esta información sólo está disponible actualmente para el inglés.

y el identificador de la palabra que se está presentando en cada caso. Además, en este caso, no se visualiza ningún tipo de anotación, pero sí la definición y el ejemplo, aunque en este último no aparece la palabra consultada (“golpe”). En su lugar, aparece un sinónimo en las oraciones que ejemplifican ambos sentidos: en el primer caso, “topetón” y, en el segundo, “conmoción”. El motivo es que, como ya se ha avanzado, la definición y el ejemplo de una palabra son compartidos por los miembros del mismo *synset* y dicho ejemplo se construye con uno de éstos.

- Palabra a buscar -

(Selecciona la opción que quieres cambiar)

Inglés Castellano

1

Localizador en la base de datos: { 7338681 }

SenseKey: **blow%1:11:00::**
SenseKey español: *golpe%1:11:00::*

Definición en español: impacto (como en una colisión)
Definición en inglés: an impact (as from a collision)

Ejemplo/s en español: el topetón lo botó de la bicicleta .
Ejemplo/s en inglés: the bump threw him off the bicycle .

2

Localizador en la base de datos: { 7298982 }

SenseKey: **blow%1:11:01::**
SenseKey español: *golpe%1:11:01::*

Definición en español: sorpresa desagradable o desilusionante
Definición en inglés: an unpleasant or disappointing surprise

Ejemplo/s en español: fue una conmoción saber que él fue herido .
Ejemplo/s en inglés: it came as a shock to learn that he was injured .

Fig 2. Ejemplo de resultado de búsqueda en WNE 3.0 sin visualizar la anotación de las palabras.

A modo de ejemplo, en la figura 3 se presenta el primer resultado de la consulta de la palabra “golpe” con toda la información relativa a la anotación visible.

- Palabra a buscar -

(Selecciona la opción que quieres cambiar)

Inglés Castellano

1

Localizador en la base de datos: { 7338681 }

SenseKey: blow%1:11:00::
SenseKey español: golpe%1:11:00::

Variantes en español: golpe, topetazo,
Variantes en inglés: blow, bump,

Definición en español: (DT) impacto (NN) (como (RB) en (IN) una (DT) colisión (NN))
Definición en inglés: an (DT) impact (NN) (as (RB) from (IN) a (DT) collision (NN))

Ejemplo/s en español: el topetazo le hizo caer de la bicicleta .
Ejemplo/s en inglés: the bump threw him off the bicycle .

Lemas de las palabras de la glosa:
En inglés: an (an) impact (impact%1 | impact%2) () as (as) from (from) a (a) collision (collision%1) ()

Synset Keys de las palabras de la glosa:
En inglés: an impact (impact%1:11:00::) (as from a collision (collision%1:11:00::))
En español: impacto (impacto%1:11:00::) (como en una colisión (colisión%1:11:00::))

Synset Keys de los ejemplos:
En inglés: the bump (bump%1:11:00::) threw him off the bicycle .
En español: el topetazo (topetazo%1:11:00::) le hizo caer de la bicicleta .

Fig 3. Ejemplo de resultado de búsqueda en WNE 3.0 visualizando la anotación de las palabras.

4. CONCLUSIONES

Hemos presentado un léxico creado a partir de la última versión (3.0) de la red semántica WordNet desarrollada para el inglés. Para la confección de este recurso, no sólo hemos traducido los elementos que conforman la macroestructura de WN, sino que también se han traducido las definiciones y los ejemplos asociados a las entradas.

Por lo que se refiere a la traducción de la macroestructura, en ocasiones ha sido necesario crear entradas en español que no se hubieran considerado a la hora de crear un diccionario de esta lengua sin tener como punto de partida el inglés. Esto ha ocurrido, por ejemplo, al traducir términos del inglés como “flying colours” o “smack” por “éxito total” o “beso sonoro”, respectivamente, conceptos que en español no están lexicalizados.

También se han traducido los distintos sinónimos que están asociados en el llamado *synset* (conjunto de sinónimos) en el recurso del inglés, el cual es la unidad básica en WN. Dicha traducción se ha llevado a cabo siempre que se ha identificado una posible correspondencia en español, lo cual puede no ocurrir, como por ejemplo, cuando las variantes inglesas son ortográficas. En algún caso, se han añadido otros posibles sinónimos para la

lengua española. Por tanto, es posible que el número de elementos que forman un *synset* del español sea finalmente mayor o menor que el de los correspondientes en inglés.

Respecto a las definiciones y los ejemplos del WN inglés, éstos contienen etiquetas con información morfosintáctica y semántica para las palabras que los componen. Aunque es cierto que la anotación morfosintáctica es una tarea asequible, ya que con el uso de las herramientas actuales para el español se obtienen resultados muy aceptables, la anotación semántica, en cambio, es una tarea cuya automatización no da todavía resultados tan satisfactorios. Así pues, sobre todo para la etiquetación de los sentidos de las palabras y, teniendo en cuenta que este tipo de anotación es interlingüística, se consideró adecuado aprovechar la anotación realizada para el inglés.

Esta decisión nos condicionó a traducir las definiciones y los ejemplos intentando mantener al máximo la estructura de la lengua origen, siempre que la frase resultante en español fuera adecuada, con el objetivo último de aprovechar la anotación proveniente de la lengua inglesa. Aunque es cierto que la redacción de las definiciones y los ejemplos a veces ha resultado complicada al intentar mantener al máximo la estructura de la lengua inglesa, creemos que ha valido la pena puesto que nos ha permitido obtener un corpus paralelo inglés-español y anotado morfosintácticamente y semánticamente. En la actualidad existen varios corpus paralelos entre ambas lenguas, pero son muy pocos los corpus en español con información sobre los sentidos de las palabras, cuya existencia es muy valiosa desde el punto de vista lingüístico y también en el campo del PLN.

Por lo que se refiere a la herramienta lexicográfica resultante para el español, cabe decir que dicho recurso supone un avance para la comunidad científica ya que hasta hoy no había disponible ninguna versión completamente pública para el español de WN.

REFERENCIAS BIBLIOGRÁFICAS

- Bentivogli, L., E. Pianta and M. Ranieri (2005). "MultiSemCor: an English Italian aligned corpus with a shared inventory of senses". *Proceedings of the Meaning Workshop 2005*, Trento, Italy, p. 90.
- Castellón, I., A. Fernández, G. Vázquez (2005). "Creación de un recurso textual para el aprendizaje del inglés", *NOVATICA Revista de la Asociación de técnicos de informática*, 177, p. 51-54.
- Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- Gelbukh, A., G. Sidorov and J. A. Vera-Félix (2006) "Paragraph-Level Alignment of an English-Spanish Parallel Corpus of Fiction Texts Using Bilingual Dictionaries". *Text, Speech and Dialogue*. Berlín: Springer, p. 61-67.

- Izquierdo, M. (2005). "Contrastive Analysis and Translation English-Spanish: functions of the English -ing form and its equivalents in Spanish". *Multilingua* (<http://multilingua.uib.no/marlen.page>). Acceso: 20.09.08.
- McEnery, A.M., Wilson, A, Sanchez-Leon, F. & Nieto-Serano, A. (1997) "Multilingual resources for European languages: Contributions of the Crater Project", *Literary and Linguistic Computing*, Volume 12, Issue 4, p. 219-226.
- Miller, G.A., R. Beckwith, Ch. Fellbaum, D. Gross, y K. Miller (1990). "Introduction to WordNet: An On-line Lexical Database". *International Journal of Lexicography*, 3(4), p. 235-244.
- Ramon, N. (2004) "Building an English-Spanish Parallel Corpus for Teaching and Research: The ACTRES Project". *Proceedings The sixth teaching and Language Corpora Conference*.
- Ranieri, M., E. Pianta and L. Bentivogli (2004). "Browsing Multilingual Information with the MultiSemCor Web Interface". *Proceedings of the LREC 2004 Workshop "The Amazing Utility of Parallel and Comparable Corpora"*, Lisboa, Portugal, p. 38-41.