Tratamiento de multipalabras en WORDNET 3.0

Marina Lloberes
Antoni Oliver
Salvador Climent
Universitat Oberta de Catalunya, Barcelona
Irene Castellón
Universitat de Barcelona, Barcelona

Resumen

En este artículo se describe el procedimiento seguido para la inclusión de unidades léxicas multipalabra en la construcción de la base de conocimiento léxico-semántico WordNet 3.0 para el español y el catalán. El trabajo incluye el desarrollo y aplicación de criterios con el fin de mejorar la calidad y coherencia de dichos WordNet respecto a versiones anteriores. En el trabajo se ha primado el objetivo práctico y aplicado —la creación del recurso— por encima del detalle teórico y, dada la gran proximidad tipológica de las dos lenguas tratadas, los criterios desarrollados son de aplicación a ambas.

Palabras clave: unidades léxicas multipalabra, modismos, WordNet, semántica léxica.

Abstract

This paper describes the procedure for the selection of multiword lexical units when building Spanish and Catalan WordNet 3.0. By this work several tests and criteria are developed and applied in order to improve the quality of such WordNets with respect to previous versions. This is a practical and empiric work, therefore applicability prevails over theoretical subtilities. Moreover, since Catalan and Spanish are languages typologically very close, criteria developed are assumed to be neutral, i.e. applicable to both.

Keywords: Multiwords, Idioms, WordNet, Lexical Semantics.

1. Introducción

Para Sag *et al.* (2001) los dos principales problemas del procesamiento computacional del lenguaje son la desambiguación de sentidos y el reconocimiento y producción de modismos y otras combinaciones de palabras que funcionan de forma unitaria desde el punto de vista léxico.

Por otra parte, WordNet (Fellbaum 1998) se ha convertido en el recurso estándar para el procesamiento computacional léxico-semántico. WordNet compendia unidades léxicas del idioma organizadas mediante relaciones semánticas. Sag *et al.* (2001) calculan que el 41% del léxico de WordNet 1.7 son unidades multipalabra.

2. Metodología

El presente trabajo se enmarca en la construcción de la versión 3.0 de WordNet para el español y el catalán[1]. La metodología seguida se enmarca en el modelo de expansión (Vossen 1998) por traducción del WordNet inglés (WNP3.0) mediante técnicas automáticas y posterior validación (Oliver y Climent, en prensa)[2].

En el estadio actual del trabajo, tras generar las traducciones del léxico del inglés se han aplicado diversos criterios de validación. Aquí se presentan los desarrollados para, dada una traducción que supera el límite del a palabra ortográfica, discriminar y rechazar las que no se consideran de nivel léxico ya que, dada su naturaleza intrínsecamente léxica, las cadenas composicionales no deben ser incluidas en WordNet3.0.

A causa de la gran proximidad tipológica entre las lenguas de destino, catalán y español, se ha optado por el desarrollo de criterios aplicables a ambas.

Se han diseñado filtros de dos tipos: automáticos y manuales, los cuales se aplican al léxico por este mismo orden.

El filtro automático se basa en el reconocimiento de modismos a partir de dos recursos externos: Wikipedia[3] y (sólo para el catalán) el *Diccionari de sinònims de frases fetes* (Espinal 2004).

El primer filtro, que utiliza la Wikipedia, se basa en la verificación de las unidades léxicas multipalabras presentes en el WordNet en castellano y catalán mediante la comparación de éstas con los títulos de las entradas de la Wikipedia para estas lenguas. Si una determinada unidad multipalabra existe como título de una entrada de la Wikipedia, se valida de manera automática. La revisión manual de los resultados nos ha permitido evaluar su precisión, que es del 91.4 % para el castellano y del 99.4 % para el catalán.

El uso del *Diccionari de sinònims de frases fetes* nos permite validar unidades multipalabra de manera automática, aceptando aquellas unidades del WordNet presentes en el diccionario. La precisión obtenida con este método es también muy alta, del 99.8 %.

Sobre el conjunto de cadenas no reconocidas por los filtros automáticos se aplican manualmente los filtros lingüísticos. Se trata de criterios en forma de *test* para la detección de la composicionalidad, la figuración y la fijación léxica o rigidez de las cadenas candidatas.

Su descripción, presentada a continuación, constituye el núcleo del presente trabajo.

3. Criterios lingüísticos

Se han establecido criterios lingüísticos de tres tipos: semánticos, sintácticos y léxicos, los cuales se aplican en este mismo orden.

Los *tests* semánticos (§ 3.1) se basan en la noción de unidad léxica multipalabra desarrollada en los trabajos de lexicografía y de terminología de Cabré *et al.* (1998), Grant and Bauer (2004), Lorente (2000) y Pazos (2005). Para complementarlos se ha establecido un conjunto de *tests* en los que aspectos semánticos como p.e. la figuración se relacionan con estructuras o posiciones sintácticas (§ 3.2).

Los tests semánticos y sintácticos nos han permitido discriminar la gran mayoría de casos. Sin embargo, para tratar casos específicos, especialmente terminología o nombres propios, ha sido necesario desarrollar un tercer tipo de *tests*, de carácter léxico (§ 3.3).

3.1 Tests semánticos

Pese a que el problema se ha abordado reiteradamente, es difícil encontrar en la bibliografía una definición y unos criterios claros para delimitar claramente las fronteras entre unidad léxica multipalabra y compuesto composicional, ya que la calificación de unidad léxica suele depender de diversos tipos de conocimiento, como el conocimiento del mundo, el grado de lexicalización y el grado de gramaticalización (Grant y Bauer 2004).

(1) a. Mañana recogerán la bala de heno.
b. Este chico es un bala perdida.

El rasgo principal que diferencia las cadenas "bala de heno" (1a) y "bala perdida" (1b) es la composicionalidad. Desde una perspectiva "naíf" (Grant y Bauer 2004), si cada término de una cadena multipalabra se puede sustituir por su definición lexicográfica y el significado no varía respecto al de la cadena léxica (contextualizada) la cadena es composicional —y no lo es en caso contrario—.

(2)

a. Mañana recogerán la bala de heno.
(i) bala: f. Fardo apretado de mercaderías, y en especial de los que se transportan embarcados. (*DRAE*)
(ii) heno: m. Hierba segada, seca, para alimento del ganado. (*DRAE*)

b. Este chico es un bala perdida.
(i) bala: f. Proyectil f. Proyectil de forma esférica o cilíndrico-ojival, generalmente de plomo o hierro. (*DRAE*)
(ii) perdida: adj. Que no tiene o no lleva destino determinado. (*DRAE*)
(iii) bala perdida: tarambana (persona de poco juicio). (*DRAE*)

En (2a) el significado de cada unidad léxica es interpretable independientemente del resto de significados de la cadena y el significado de la cadena es el resultado de la composición de significados. La situación contraria se da claramente en (2b) por lo que "bala perdida" es candidata a unidad léxica multipalabra.

Además, la calificación de una cadena léxica como multipalabra está vinculada a su capacidad de tener una lectura figurada (Grant y Bauer 2004). La cadena de (2a) no se puede considerar léxica ya que los significados que forman el compuesto tienen una lectura literal (2a.i) (2a.ii). En cambio, el significado de la de (2b) no se conforma a partir de los sentidos literales (2b.i) (2b.ii) sino a partir de un proceso de metaforización (2b.iii).

3.2. Tests sintácticos

Al contrario de las cadenas léxicas composicionales, las unidades léxicas multipalabra, al interpretarse como una sola unidad significativa, difícilmente aceptan transformaciones sintácticas. Por lo tanto, tienden a presentarse como estructuras fijas; presentamos diversos casos definidos a partir de la descripción de Pazos (2005).

3.2.1. Nombre + Adjetivo

Criterio de modificación: mientras que una unidad unipalabra (3a) puede ser modificada por un adjetivo (3b), los componentes de una multipalabra (4a) (4c) no admiten modificación (4b) (4d). Sin embargo, el grupo completo sí puede ser modificado.

(3) a. hacer un aterrizaje
b. hacer un aterrizaje forzoso

(4) a. tomar tierra
b. *tomar tierra forzosa
c. raíz cuadrada
d. *raíz poco cuadrada

3.2.2. Verbo + Sintagma Nominal

3.2.2.1. SN Sujeto

Lectura metafórica. En caso de que uno de los términos de la cadena tenga una lectura figurada (por ejemplo, interpretación metafórica), el conjunto se considerará unidad léxica (5).

(5) a. correr un rumor
b. estalló la guerra

3.2.2.2. SN Objeto

Lectura metafórica. El mismo criterio que en la anterior.

(6) a. acariciar una idea

Relativización. Las cadenas léxicas multipalabra no pueden estar complementadas por una subordinada relativa en la que el antecedente sea uno de los términos de la cadena (7)

(7) a. Le he echado el ojo a ese vestido
b. *El ojo que acabo de echar a ese vestido

Pasivización. Las cadenas léxicas multipalabra no permiten su transformación pasiva (8)

(8) a. El órgano fue trasplantado
b. *El bulto fue escurrido

3.2.2.3. Verbos ligeros

Un volumen importante de unidades léxicas multipalabra se crean a partir de sustantivos y verbos *ligeros*, es decir, que pierden parte de su carga semántica en este tipo de construcciones (9).

(9) a. dar clase
b. tomar una decisión
c. posar ordre
d. prendre partit

Para su detección hemos establecido la siguiente lista de verbos light: "dar", "tomar", "hacer", "poner", "tener" (español) "fer", "posar" y "prendre" (catalán).

Mientras que la cadena formada un "verbo light" seguido de un nombre se califica de unidad léxica (10a), las cadenas formadas por verbo conjugado "hacer" o "fer" e infinitivo no lo son en ningún caso, dado que "hacer" o "fer" aportan valor causal a la oración (Alsina 2000) (10b). Además, este tipo de cadenas puede tener verbos equivalentes (11).

(10) a. He dado clase durante toda la mañana.
b. He hecho callar a los alumnos.

(11) a. poner énfasis
b. enfatizar

3.2.3. Nombre / Verbo + Sintagma preposicional

Se consideran multipalabras algunas cadenas léxicas nominales y verbales concretas (12).

(12) a. llevar a cabo
b. poner en funcionamiento
c. poner de relieve

Los nombres seguidos de un sintagma preposicional y expresan unidad o grupo al que pertenece una entidad o individuo no se consideran unidad léxica (13).

(13) a. diente de ajo
b. barra de pan
c. banco de peces

3.2.4. Nombre / Verbo con combinatoria limitada

Determinadas cadenas léxicas nominales y verbales, aunque composicionales, presentan una combinatoria altamente condicionada (colocaciones estrechas) (14).

(14) a. error garrafal
b. conciliar el sueño

En estos casos hemos optado por incluir el compuesto en la base de conocimiento ya que para el procesamiento computacional funcionan igual que las unidades léxicas multipalabra.

3.3. Tests léxicos

Finalmente, definimos diferentes tipos de términos como considerados unidad léxica multipalabra:

1. Términos científicos realizados por más de una unidad léxica (*acero inoxidable* o *médula ósea*).

2. Nombres propios; de persona (*Pau Solà*), topónimos (*Vilanova y la Geltrú*), de organizaciones (*Universidad Complutense*) y títulos de libros, películas, etc. (*Lo que el viento se llevó*, *Cien años de soledad*).

3. Términos compuestos pertenecientes a otras lenguas que no se traducen enla lengua de destino (*bloody mary*, *deutsche mark*).

4. Fechas (*2 de junio de 1997*)

En cambio no se consideran unidades léxicas:

1. Expresiones y frases hechas, como *chiste de mal gusto*, *cornudo y apaleado*, etc. Se consideran unidades fraseológicas pero no de nivel léxico.

2. Los "phrasal verbs" del inglés que traducidos al castellano o al catalán resultan en verbos simples con complementos preposicionales regidos. No se consideran unidad léxica en catalán y español porque se entiende que el verbo subcategoriza los complementos y, por tanto, la preposición regida por el verbo introduce el complemento en cuestión[4].

3.4. Tests por defecto

En último término, si tras aplicar todos los filtros establecidos persiste en el desarrollador la duda sobre una determinada cadena, se la considera unidad léxica multipalabra.

4. Conclusiones y trabajo futuro

En este trabajo se ha presentado un conjunto de criterios lingüísticos para la distinción entre cadenas composicionales y cadenas léxicas multipalabra, los cuales se han recopilado a partir de trabajos previos de carácter teórico. Dichos criterios se han utilizado en el trabajo de validación manual para la inclusión o rechazo de cadenas de palabras en tanto que unidades léxicas en las bases de conocimiento semántico WordNet 3.0 del castellano y el catalán —los cuales se construyen en una primera fase mediante métodos semiautomáticos—.

Los criterios se han aplicado consecutivamente en tres niveles lingüísticos: semántico, sintáctico y léxico, después de un primer filtrado automático realizado con recursos externos de referencia.

Un aspecto pendiente de estudio es el de los verbos que adquieren una preposición cuando se realizan en un sentido determinado. Aunque de entrada se ha adoptado el criterio de no incluirlos como unidades léxicas multipalabra, la cuestión debe ser analizada con mayor profundidad.

Notas

1. Investigación financiada por los proyectos Representación del Conocimiento Semántico (SKR), TIN2009-14715-C0403 y Multilingual Central Repository 2.0: GRIAL, FFI2009-08466-E/FILO, ambos del Ministerio de Ciencia e Innovación español.

2. Versión en desarrollo consultable en: http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl.

3. http://www.wikipedia.org/.

4. Sin embargo hay algunos verbos que adquieren el sentido con la preposición: "apostar por" (confiar), "contar con" (tener en cuenta), etc. Por ello, esta cuestión está siendo revisada para poder establecer un criterio más consistente.

Referencias bibliográficas

Alsina, A. 2000. "L'infinitiu". *Gramàtica del Català Contemporani.* Eds. J. Solà, M.R. Lloret, J. Mascaró, y M. Pérez Saldanya. Empúries. 2391-2458.

Cabré, M.T., Estopà, R. y Lorente, M. 1998. "Terminología y fraseología." *Actas del V Simposio Iberoamericano de Terminología: Terminología, ciencia y tecnología,* Colegio de México y Unión Latina. 6781.

Espinal, M.T. 2004. *Diccionari de sinònims de frases fetes.* Serveis de publicacions de la Universitat Autònoma de Barcelona, Publicacions de la Universitat de València i Publicacions de l'Abadia de Montserrat.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. The MIT press.

Grant, L. y Bauer, L. 2004. "Criteria for re-dening idioms: Are we barking up the wrong tree?" *Applied Linguistics* 25(1): 3861.

Lorente, M. 2000. "Altres elements lèxics". *Gramàtica del Català Contemporani.* Eds. J. Solà, M.R. Lloret, J. Mascaró i M. Pérez Saldanya, Empúries. 833-890.

Oliver A. y Climent, S. (en prensa) "Parallel corpora for WordNet construction: machine translation vs. automatic sense tagging" *Proceedings of the13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2012).*

Pazos, J.M. 2005. *Detección automatizada de fraseologismos*. Tesis doctoral, Universidad de Granada.

Sag. I.A., Baldwin, T., Bond, F., Copestake, A. y Flickinger, D. 2002. "Multiword Expressions: A Pain in the Neck for NLP". *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002).* Lecture Notes in Computer Science 2276, Londres: Springer-Verlag. 1-15.

Vossen, P. 1998. "Introduction to EuroWordNet". *Computers and the Humanities* 32(2): 73-89.