.puntoycoma,

Boletín de los traductores españoles de las instituciones de la Unión Europea

nº 115-S noviembre/diciembre de 2009

Suplemento
VIII Jornada de AETER

En este suplemento, que puntoycoma *publica de manera excepcional, se reúnen las contribuciones presentadas en la VIII Jornada de la Asociación Española de Terminología (AETER), que se celebró el 21 de noviembre de 2008 en la Escuela Técnica Superior de Ingenieros de Caminos de la Universidad Politécnica de Madrid con el título «Modelos, recursos y aplicaciones informáticas para la terminología». En la página web de AETER <http://www.aeter.org/home.asp> se ofrece información sobre las actividades de la asociación.*

Terminología aplicada basada en corpus

XAVIER GÓMEZ GUINOVART
Universidade de Vigo
xgg@uvigo.es

1. Introducción

La orientación hacia la investigación aplicada basada en corpus textuales se ha consolidado en los últimos años como una metodología fructífera para la descripción y análisis de los fenómenos lingüísticos en prácticamente todos sus aspectos. En este artículo, presentaré una aproximación a la investigación basada en corpus en el ámbito de los trabajos terminológicos, ilustrando la aplicación de esta metodología con algunos trabajos realizados por nuestro grupo de investigación de la Universidad de Vigo en torno a la elaboración de una base de conocimientos terminológicos de la lengua gallega denominada Termoteca.

2. Lexicografía y terminografía basada en corpus

El estudio de la lengua a través de los corpus textuales permite aproximarse de una manera empírica al uso real del lenguaje en su contexto. El análisis de las unidades léxicas de un corpus textual permite observar su potencialidad semántica, su frecuencia de uso y su combinatoria de un modo muy realista y ciertamente inalcanzable desde la pura reflexión introspectiva sobre el funcionamiento del lenguaje. Del mismo modo, en el estudio del discurso lingüístico técnico o especializado, la explotación de corpus técnicos con las herramientas informáticas apropiadas facilita la tarea de identificar en los textos el repertorio utilizado de unidades léxicas con contenido terminológico y permite al mismo tiempo observar su polisemia y su sinonimia, comprobar su frecuencia en los textos, obtener ejemplos reales de uso y contextos definitorios e, incluso, descubrir las relaciones semánticas codificadas en los textos entre los términos asociados a un ámbito temático de especialidad.

Tradicionalmente, los autores de repertorios léxicos buscaban sus fuentes de información sobre los datos lingüísticos en otros repertorios léxicos, en citas selectas de textos del canon literario o en su propia intuición como hablantes de la lengua. Este método de trabajo suponía limitaciones muy considerables para la práctica lexicográfica, ya que, por una parte, las reflexiones propias de los lexicógrafos sobre el uso del léxico podían no ser ajustadas a la realidad lingüística; por otra parte, la recopilación manual de citas de obras canónicas resultaba un trabajo lento y muy poco productivo; y, por último, los diccionarios usados como fuente de inspiración solían no estar actualizados o, en el peor de los casos, podían contener errores acumulados debidos a su sucesiva reproducción a lo largo de los tiempos.

La introducción del uso de corpus textuales informatizados en la práctica lexicográfica contribuye sin duda a la superación de estas limitaciones de la metodología tradicional, facilitando la observación del léxico de una lengua en la realidad de su uso lingüístico, es decir, en su práctica textual. El primer caso de éxito en la introducción del uso de los corpus textuales informatizados para la elaboración de diccionarios le correspondió a la Universidad de Birmingham y a la editorial Collins, promotora del diccionario *Cobuild* (Sinclair 1987), cuya primera edición vio la luz en 1987. En su momento, el proyecto Cobuild fue muy innovador, ya que por vez primera se utilizaba en lexicografía un corpus representativo de textos para facilitar el análisis de los significados de las palabras, la identificación de patro-

nes sintácticos y la descripción de las colocaciones y de la fraseología de una lengua, en concreto el inglés contemporáneo. Tras el éxito del *Cobuild*, la metodología de trabajo de la lexicografía basada en corpus fue seguida por otras grandes editoriales, como Oxford University Press, Longman y Larousse (que colaboraron en la elaboración del British National Corpus) y Cambridge University Press.

En el caso del español, podemos ver ejemplos recientes de la aplicación de esta metodología en el diccionario publicado por la editorial SGEL a partir del corpus *Cumbre* (Sánchez 2001) o en el diccionario de colocaciones *Redes* (Bosque 2004) basado en un corpus periodístico de 250 millones de palabras de la editorial SM. La metodología de trabajo de la lexicografía basada en corpus se está empleando también para el catalán en la elaboración por parte del IEC del *Diccionari descriptiu de la llengua catalana* basado en el *Corpus Textual Informatitzat de la Llengua Catalana* (Rafel 1997). En Galicia, el corpus de referencia del gallego denominado *Tesouro Informatizado da Lingua Galega* (Santamarina 2003) constituye la base del dicionario de uso de la lengua gallega dirigido por el profesor Antón Santamarina, en fase de preparación; y el *Corpus CLUVI* (Gómez Guinovart 2003), elaborado en el marco de nuestro grupo de investigación de la Universidad de Vigo, es la fuente textual en la que se fundamenta el *Dicionario CLUVI inglés-galego* (Gómez Guinovart *et alii* 2008), disponible libremente en la red desde 2005 y de inminente aparición en edición impresa.

Aunque las bases teóricas para el trabajo en terminología basada en corpus son similares a las de la lexicografía basada en corpus, la terminología basada en corpus ha tardado más tiempo en afianzarse como un procedimiento de trabajo normalizado, a causa, probablemente, de la diferente naturaleza de los corpus con los que se trabaja, ya que en el caso de la lexicografía, los corpus suelen ser de amplia base y alcance general, mientras que en el caso de la terminología se trabaja con corpus más orientados a determinados dominios que muchas veces resultan de difícil obtención.

Con todo, en estos momentos, la terminología moderna (que tanto debe a los trabajos del Grupo IULATERM, liderado por Teresa Cabré en la Universidad Pompeu Fabra de Barcelona) sostiene principios teóricos y metodológicos que destacan la importancia del uso de grandes repertorios textuales para el trabajo terminográfico, debido a las facilidades que estos ofrecen para la identificación en los textos de las unidades con contenido especializado y para la extracción de la información terminológica codificada en los textos asociada con estas unidades. Como nos recuerda la Teoría Comunicativa de la Terminología (Cabré 1999), para la terminología moderna los textos son el «hábitat natural» de los términos, el medio en el que observar la verdadera naturaleza de las unidades de valor especializado. En este sentido, la teoría terminológica moderna substituye el paradigma prescriptivo de la Teoría General (o Tradicional) de la Terminología por una visión descriptiva de su objeto de estudio, una perspectiva que nuestro grupo de investigación de la Universidad de Vigo comparte y que nos ha conducido de manera natural a la adopción de una metodología basada en corpus en nuestra investigación en el campo de la terminología de la lengua gallega.

Presentaré ahora a modo de ejemplo, con suma concisión, los trabajos que está llevando a cabo nuestro grupo universitario de investigación en la construcción de la Termoteca, un banco de datos terminológico para el gallego basado en corpus especializados monolingües y paralelos.

3. La Termoteca

La Termoteca es un banco de datos terminológico basado en los textos de especialidad monolingües y paralelos recopilados, respectivamente, en el *Corpus Técnico do Galego* (Gómez Clemente / Gómez Guinovart 2006) y en el

Corpus CLUVI (Gómez Guinovart 2003). El CLUVI (Corpus Lingüístico da Universidade de Vigo) es un conjunto de corpus paralelos de unos 23 millones de palabras, formado principalmente con traducciones al gallego o del gallego, de libre consulta en la web en la dirección <http://sli.uvigo.es/CLUVI>. Por su parte, el CTG (Corpus Técnico do Galego) es una colección de corpus del gallego contemporáneo de unos 14 millones de palabras, compuesta de textos monolingües especializados en los campos del Derecho, la informática, la economía, las ciencias ambientales, la sociología y la medicina, disponible para libre consulta en <http://sli.uvigo.es/CTG/>.

La información terminológica extraída de los corpus CTG y CLUVI de manera semiautomática incluye los propios términos, junto con sus contextos, variantes formales y frecuencias de uso; su definición o definiciones, cuando se pueden documentar en los corpus; y las relaciones semánticas que establecen con otros términos del corpus, cuando aparecen explícitamente codificadas en los textos. Las técnicas utilizadas para extraer la información son de tipo lingüístico-computacional y estadístico, y sus resultados son siempre revisados y complementados por especialistas (Crespo *et alii* 2008).

El banco de datos terminológico de la Termoteca, de libre acceso en la web en la dirección <http://sli.uvigo.es/termoteca>, está mantenido por el Grupo TALG de la Universidad de Vigo y cuenta, en la actualidad, con unos 6 000 registros con información sobre más de 10 000 términos documentados en los corpus CLUVI y CTG pertenecientes a los ámbitos del Derecho (3 473 términos del gallego y del español especificados en registros bilingües y monolingües de la Termoteca), de la sociología (3 365 términos del gallego, del español, del francés y del inglés en registros tetralingües y monolingües de la Termoteca), de la economía (1 410 términos del gallego y del español en registros monolingües y bilingües de la Termoteca) y de la ecología y ciencias ambientales (1 437 términos del gallego en registros monolingües de la Termoteca). Se está trabajando en la ampliación de la base de datos terminológica a los campos de la medicina (actualmente, 1 015 términos del gallego en registros monolingües de la Termoteca) y de la informática (en estos momentos, 399 términos del gallego en registros monolingües de la Termoteca), a partir de los datos de las secciones especializadas correspondientes de los corpus CLUVI y CTG (Gómez Guinovart 2008).

Cada registro de la Termoteca incluye toda la información relativa a un concepto especializado, expresado con un término gallego documentado en los corpus, y del que se pueden recoger también en el mismo registro sus variantes documentadas, tanto intralingüísticas (términos sinónimos, variantes ortográficas o variantes dialectales) como interlingüísticas (traducciones o, con mayor propiedad, equivalencias). La información recogida en la Termoteca para cada variante (incluida la variante común o no marcada) incluye el lema del término, su categoría gramatical como conjunto, el análisis morfosintáctico de sus componentes, su definición, su frecuencia de aparición y un contexto de uso documentado en el corpus. Todos los registros de la Termoteca están catalogados, además, según su campo temático, en referencia a un árbol conceptual jerarquizado de la materia, y pueden incluir información sobre las relaciones semánticas (antonimia, hiperonimia, holonimia, etc.) que guardan con otros registros del banco de datos.

La Termoteca puede incluir también información relativa a la neología para los términos considerados neológicos, es decir, para los neónimos. Por ahora, solo se ha podido codificar la información neológica relativa a los términos de las ciencias ambientales. Para cada término neológico, analizamos su antigüedad, su frecuencia y su dispersión en distintos corpus, comprobamos la exclusión lexicográfica de sus componentes léxicos, y a partir de estos

datos derivamos un índice de neologicidad que incluimos, junto con el resto de los datos neológicos analizados, en los registros terminológicos correspondientes de la Termoteca (López Fernández 2009).

La aplicación web de consulta de la Termoteca permite realizar consultas en el banco de datos a partir de un término dado, a partir de una secuencia de caracteres y comodines (técnicamente, expresiones regulares) que definen los términos buscados, a partir del área temática de elección, o bien a partir del patrón morfosintáctico al que se desea que se ciñan los términos consultados. Una vez situados en un registro terminológico de la Termoteca, la aplicación de consulta utiliza la información temática y semántica incorporada para permitir recorrer los registros siguiendo las relaciones semánticas que se establecen entre ellos, o accediendo a todos los registros que comparten la misma rama del árbol temático que el registro consultado. De este modo, la Termoteca puede concebirse y visualizarse como una red léxico-semántica a dos niveles formada por nodos conceptuales que se interrelacionan en función de su clasificación temática y de sus relaciones semánticas.

4. Conclusiones

El manejo de corpus técnicos permite observar directamente la realidad lingüística plasmada en los textos especializados, facilitando el análisis empírico de muchos aspectos pragmáticos de la terminología que no sería posible estudiar de otra manera sin grandes dificultades (como su frecuencia de uso, su potencialidad semántica, su dispersión textual, su datación temporal o su combinatoria).

Sin embargo, el trabajo con corpus impone ciertas limitaciones de las que la investigación terminológica no se encuentra exenta. En primer lugar, hay que tener en cuenta que basar el trabajo terminográfico en corpus exige la existencia de material textual suficiente escrito en el ámbito especializado y en la lengua que se desea estudiar. Por ejemplo, la producción textual del gallego en ámbitos técnicos muy recientes o de alta especialidad, como los de la genómica, la mecánica cuántica, o la aceleración de partículas es muy limitada o prácticamente inexistente, excepto en aquellos casos en que la producción es impulsada por la Administración, por lo que la investigación terminológica basada en corpus en esos campos es impracticable. Esta limitación es aún mayor en el caso de desear realizar una aproximación plurilingüe basada en corpus. Por ejemplo, en gallego poseemos una cierta producción textual sobre el cambio climático. Sin embargo, son prácticamente inexistentes los textos paralelos inglés-gallego en este campo. La incorporación del factor traducción limita al gallego en casi todos los ámbitos especializados, con la excepción del Derecho en la combinación gallego-español, gracias al imperativo legal vigente.

Otra limitación importante derivada de la metodología de corpus se debe a que a veces, por azar o por limitaciones de la selección de los textos del corpus, términos que sospechamos que pueden ser frecuentes o normales en un determinado ámbito de especialidad no se encuentran documentados en el corpus manejado. La causa es que, por lógica estadística (no olvidemos que un corpus es una muestra de una población mayormente desconocida), lo más posible es que ningún corpus contenga todos los términos de un ámbito. Para solucionar este problema, al menos parcialmente, se puede intentar aumentar el tamaño del corpus y diversificar la variedad temática y de registros de los textos recopilados, siempre que eso sea posible.

Finalmente, aunque la extracción semiautomatizada de información terminológica de los corpus técnicos complementa con gran eficiencia el trabajo de investigación humano, de ninguna manera lo hace innecesario. Cualquier metodología de extracción automática de información terminológica aplicada a cor-

pus debe ser complementada por una larga fase de trabajo humano de ponderación, reflexión y toma de decisiones a partir de los datos obtenidos.

Bibliografía

BOSQUE, Ignacio (2004), *Diccionario Redes: Diccionario combinatorio del español contemporáneo*, Ediciones SM, Madrid.

CABRÉ, Teresa (1999), *La terminología: representación y comunicación*, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

CRESPO BASTOS, Ana / Xosé María GÓMEZ CLEMENTE / Xavier GÓMEZ GUINOVART / Susana LÓPEZ FERNÁNDEZ (2008), «XML-based Extraction of Terminological Information from Corpora», 28-39 en José Carlos RAMALHO, João CORREIA LOPES / Salvador ABREU eds. *Actas da 6ª Conferência Nacional XATA'2008*, Universidade de Évora, Évora.

GÓMEZ CLEMENTE, Xosé María / Xavier GÓMEZ GUINOVART dirs. (2006), *Corpus Técnico do Galego*, Universidade de Vigo, Vigo: <http://sli.uvigo.es/CTG/>.

GÓMEZ GUINOVART, Xavier (2008), «A investigación en lexicografía e terminoloxía no Corpus Lingüístico da Universidade de Vigo (CLUVI) e no Corpus Técnico do Galego (CTG)», 209-228 en Ernesto GONZÁLEZ SEOANE / Antón SANTAMARINA / Xavier VARELA BARREIRO eds. *A lexicografía galega moderna: Recursos e perspectivas*, Consello da Cultura Galega / Instituto da Lingua Galega, Santiago de Compostela.

GÓMEZ GUINOVART, Xavier dir. (2003), *Corpus CLUVI (Corpus Lingüístico da Universidade de Vigo)*, Universidade de Vigo, Vigo: <http://sli.uvigo.es/CLUVI/>.

GÓMEZ GUINOVART, Xavier coord. / Alberto ÁLVAREZ LUGRÍS / Eva DÍAZ RODRÍGUEZ (2008²), *Dicionario CLUVI Inglés-Galego*: <http://sli.uvigo.es/dicionario/>.

LÓPEZ FERNÁNDEZ, Susana / Xavier GÓMEZ GUINOVART / Xosé María GÓMEZ CLEMENTE / Ana CRESPO BASTOS (2009), «A avaliación da neoloxicidade en terminoloxía», en Teresa CABRÉ / O. DOMÈNECH / Rosa ESTOPÀ / Judit FREIXA eds. *Actes de CINEO 2008: Actes del I Congrès Internacional de Neologia de les Llengües Romàniques*, Universitat Pompeu Fabra, Barcelona.

RAFEL, Joaquim dir. (1997), *Corpus Textual Informatitzat de la Llengua Catalana*, Institut d'Estudis Catalans, Barcelona: <http://ctilc.iec.cat/>.

SÁNCHEZ, Aquilino dir. (2001), *Gran diccionario de uso del español basado en el Corpus lingüístico CUMBRE*, Sociedad General Española de Librería, Madrid.

SANTAMARINA FERNÁNDEZ, Antón dir. (2003), *Tesouro informatizado da lingua galega (TILG)*, Universidade de Santiago de Compostela, Santiago de Compostela: <http://www.ti.usc.es/TILG/>.

SINCLAIR, John ed. (1987), *Collins Cobuild English Language Dictionary*, Collins, Londres.

puntoycoma

Cabos sueltos: notas breves en las que se exponen argumentos o se facilitan datos para solucionar problemas concretos de traducción o terminología.

Neológica Mente: reflexiones, debates y propuestas sobre neología, en concomitancia con el foro NeoLógica.

Colaboraciones: opiniones, propuestas y debates firmados por nuestros lectores y por los miembros de la redacción cuando intervienen a título personal.

Tribuna: contribuciones especiales de personalidades del mundo de la traducción.

Buzón: foro abierto a los lectores de *puntoycoma* para que manifiesten su opinión sobre temas ya tratados.

Reseñas: crítica de obras relacionadas con los temas tratados en *puntoycoma.*

Comunicaciones: información sobre publicaciones y calendario de acontecimientos relacionados con la traducción.

(La responsabilidad de todas las colaboraciones firmadas incumbe a sus autores)

puntoycoma ISSN 1830-5415

CORRESPONDENCIA Y SUSCRIPCIONES

Alberto Rivas
Comisión Europea
JMO A3-071A
L-2920 Luxemburgo
Tel. (352) 4301-32094
dgt-puntoycoma@ec.europa.eu

ISSN 1830-541-5

REDACCIÓN

Bruselas
Isabel Carbajal, Mónica Fuentes, Pollux Hernúñez, Miguel Á. Navarrete, María Valdivieso y José Luis Vega

Luxemburgo
Josep Bonet, Victoria Carande, Loli Fernández, Alberto Rivas, Carmen Torregrosa, Xavier Valeri y Miquel Vidal

Madrid
Luis González

Secretaría: Luz Ayuso e Isabel de Miguel,
con la colaboración de Tina Salvà y May Sánchez Abulí