>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus lingüístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

Sobre la construcción de diccionarios basados en corpus

Guillermo Rojo
Universidade de Santiago de Compostela

Desde la publicación del conocido libro de Thomas S. Kuhn *The Structure of Scientific Revolutions,* se ha generalizado una visión de la historia de las disciplinas lingüísticas según la cual las épocas de desarrollo tranquilo de la actividad científica, que ocupan la mayor parte del tiempo, son interrumpidas de vez en cuando por etapas de fuerte agitación en las que dos o más paradigmas conceptuales entran en lucha y de las que solo se sale cuando uno de ellos triunfa sobre los demás. Así pues, el cambio de paradigma, que es lo que produce la revolución, se origina como consecuencia de un cambio conceptual. Esta visión, muy discutida y matizada por otros autores y también por el propio Kuhn, fue especialmente cultivada y defendida por los partidarios de la lingüística generativo-transformacional, que consideraban que, gracias al trabajo de Chomsky, se había producido un cambio de paradigma o, más bien, el salto desde una concepción precientífica, de orientación puramente descriptiva, a una configuración realmente científica, de intención explicativa.

La cuestión es realmente compleja y no es posible, por tanto, dedicarle aquí la atención que merece. Me interesa, sin embargo, dada su relevancia para lo que aquí nos ocupa, señalar que las revoluciones científicas pueden deberse a cambios conceptuales, pero pueden también ser originadas por la aparición de nuevos instrumentos o herramientas. Los cambios conceptuales consisten, usando las palabras empleadas por Freeman Dyson, en "explicar cosas antiguas de nuevas maneras" (Dyson, 1997: 50). Son, sin duda, las más llamativas (y las menos frecuentes): el paso de la física newtoniana a la einsteniana primero y a la cuántica después es uno de los casos más citados. Pero hay a su lado otras revoluciones, las que se deben al descubrimiento y aplicación de nuevas herramientas de trabajo, con las que, de repente, lo que está al alcance de los científicos experimenta una enorme ampliación, con lo que aumenta el número de fenómenos que pueden y deben ser descritos o explicados. Eso es lo que sucedió – de nuevo con un episodio mencionado con mucha frecuencia – cuando Galileo apuntó a la Luna y a Júpiter con el telescopio que acababa de construir y descubrió un universo mucho más rico y complejo que el que hasta ese momento habían podido contemplar los seres humanos. Con palabras de Dyson, "[e]l efecto de una revolución impulsada por herramientas es descubrir cosas nuevas que tienen que ser explicadas" (*ibídem*).

En la lingüística contemporánea, la generalización del uso de computadoras ha supuesto enormes cambios de los más diversos tipos, lo que nos permite considerar que estamos ante una auténtica revolución debida a la introducción de nuevos instrumentos de análisis. En efecto, las computadoras han cambiado la propia configuración de nuestra disciplina, haciendo surgir campos de trabajo inexistentes con anterioridad, como la lingüística informática, la lingüística computacional o la tecnología lingüística y, por supuesto, han modificado profundamente la forma de trabajar en los que ya existían. Existe un amplio acuerdo en la consideración de que la lexicografía es una de las disciplinas en las que este efecto ha sido más fuerte, quizá incluso la que lo ha experimentado con más claridad. En efecto, los diccionarios en formato electrónico, que son el producto visible más exitoso de esta influencia, surgen de la conjunción de dos factores muy distintos y aparentemente opuestos entre sí. Por una parte, la preocupación que siempre han mostrado los lexicógrafos por aspectos como la codificación y la jerarquización de la información contenida en las entradas aproxima los diccionarios a las bases de datos computacionales, de forma que la reconversión a formato electrónico de lo concebido previamente para ser impreso constituye un proceso bastante natural y realizable con unos costes razonables. De otra, todo el mundo sabe que la recuperación de la información contenida en los diccionarios impresos resulta rápida y cómoda cuando lo buscado encaja bien en los principios organizativos del diccionario, pero muy dificultosa o simplemente imposible cuando no

es así. Piénsese, por ejemplo, en lo sencillo que resulta en un diccionario de uso del español recuperar las palabras que presentan un determinado prefijo frente a la pesadísima tarea que sería hacer lo mismo con los sufijos, tarea que, en cambio, es muy fácil con un diccionario inverso. La razón de ello es evidente: un diccionario en papel se organiza con un único criterio o bien con una serie de criterios jerarquizados, mientras que el formato electrónico permite, por su propia esencia, acceder por vías diferentes a la misma información. En cierto modo, el formato electrónico permite tener el equivalente de varios diccionarios tradicionales en un único soporte, lo cual, unido a su comodidad de transporte y facilidad de manejo explica el enorme éxito que ha tenido este producto.

La influencia de las técnicas computacionales en lexicografía es, sin embargo, mucho más amplia. En su clásico *Manual of Lexicography*, Zgusta distinguía cuatro fases en la preparación de un diccionario: recogida del material, selección de las entradas, redacción de las entradas y control de las entradas. A ellas tenemos que añadir hoy otras dos: composición e impresión del diccionario y difusión de la obra. Pues bien, todos los proyectos lexicográficos de importancia media o alta incorporan hoy recursos computacionales en la mayoría de estas fases o incluso en todas ellas. Me refiero, por supuesto, a la utilización de recursos computacionales como auténticos elementos del proceso, no, por ejemplo, a la simple utilización de un procesador de textos para escribir las entradas.[1]

La utilización de corpus textuales en el trabajo lexicográfico incide en el bloque constituido por las tres primeras fases diferenciadas por Zgusta, esto es, la recogida de material, la selección de las entradas y la redacción. Para situar adecuadamente la cuestión hemos de tener en cuenta que en la todavía corta historia de los corpus lingüísticos (desde 1964, con la finalización del *Brown Corpus*), se ha pasado de conjuntos de un millón de formas a corpus de cientos o miles de millones o incluso a la utilización de todo lo que está contenido en la red (la línea de trabajo conocida como *Web as Corpus*). Además del aumento de tamaño, con todo lo que ello supone, los avances en los recursos computacionales permiten hoy obtener en décimas de segundo la respuesta a consultas realizadas sobre esas inmensas masas de textos.

Pues bien, con respecto al conjunto de tareas que hay que realizar en todo este bloque, la utilización de corpus lingüísticos supone un cambio radical en el modo de concebir el trabajo con respecto a la mayor parte de la lexicografía anterior. Sinclair lo expuso en toda su profundidad en la introducción a la primera edición del COBUILD, que es el primer diccionario basado en corpus en el sentido más estricto de la expresión. Lo revelador de sus palabras justifica, creo, la extensión de la cita:

> For the first time, a dictionary has been compiled by the thorough examination of a representative group of English texts, spoken and written, running to many millions or words. This means that in addition to all the tools of the conventional dictionary makers –wide reading and experience of English, other dictionaries and of course eyes and ears– this dictionary is based on hard, measurable evidence. No major uses are missed, and the number of times a use occurs has a strong influence on the way the entries are organized. Equally, the large group of texts, called the corpus, gives us reasonable grounds for omitting many uses and word-forms that do not occur in it. It is difficult for a conventional dictionary, in the absence of evidence, to decide what to leave out, and a lot of quite misleading information is thus preserved in the tradition of lexicography (Sinclair, 1987: xv).

El objetivo de un proyecto lexicográfico basado en corpus es, con toda claridad, recoger las palabras que figuran en un corpus representativo de la lengua o variedad lingüística sobre la que se trabaja y reflejar los significados realmente presentes en los textos, incorporando las marcas de uso correspondientes en cada caso. No hay, pues, de entrada, intento de reproducir lo que otros diccionarios han incorporado previamente, ni de seleccionar palabras o acepciones en función de su consideración desde criterios de tipo normativo,

1 Para una interesante y actualizada revisión del papel de los corpus en el trabajo lexicográfico, cf. Atkins & Rundell (2008, esp. cap. 3).

de adecuación a usos 'políticamente correctos', etc. Se trata de incluir en la obra lo que realmente se da en la lengua utilizando un corpus textual para acceder a ella. Conviene, en este sentido, tener en cuenta que la utilización de corpus para afinar definiciones, decidir sobre la inclusión o no de palabras o acepciones, etc. es sin duda algo positivo, pero no hace por sí misma que se pueda considerar que un diccionario está realmente basado en corpus si, por ejemplo, la selección de palabras o acepciones se hace con criterios procedentes de alguna perspectiva externa a lo que es el propio corpus. Por supuesto, tampoco es realmente un diccionario basado en corpus aquel que parte de las palabras contenidas en un corpus o de una selección de ellas basada en su frecuencia, pero incorpora después, para cada una, todas las acepciones que figuran en otros diccionarios.

Hay que reconocer que lo que acabo de señalar como objetivo de un diccionario basado en corpus no es diferente de lo que perseguido anteriormente por los diccionarios más científicos (diccionarios históricos, por dar el caso más claro).[2] Sin embargo, la diferencia con los procedimientos tradicionales es también muy fuerte, pero en este caso tiene que ver con la revolución producida por el uso de estas nuevas herramientas. En el modo de trabajo tradicional, la recogida de material –y, a continuación, la selección de las entradas y la redacción de los artículos– se llevaba a cabo mediante el despojo, más o menos sistemático, de un conjunto de textos habitualmente muy amplio y bien elegido. En realidad, no podía hacerse de otro modo. El bien conocido proceso de recolección de materiales para la primera edición del *Oxford English Dictionary* (OED), modélico en los planteamientos tradicionales, proporciona una idea exacta: cientos de colaboradores (*readers*), que leen textos (escritos) de diferentes épocas y, según las indicaciones recibidas del equipo central y su propio criterio personal, seleccionan los ejemplos de los que hacen papeletas y las envían al equipo redactor. En un proceso de este tipo es inevitable que estén sobrerrepresentados aquellos casos que, por alguna razón, llaman la atención de la persona que hace la lectura del texto, lo cual depende de sus conocimientos lingüísticos, la familiaridad con el texto o los textos de la época, etc. Como muestra de lo que quiero decir, recuérdese la conocida circular de Murray a sus colaboradores para pedirles que no olvidasen enviar ejemplos de palabras corrientes en empleos habituales, puesto que esos casos tenían que estar también registrados en el OED.[3]

La cuestión básica es aquí, por supuesto, la existencia de un filtro, de un proceso de selección. Selección primero de los textos que van a ser examinados y luego de los ejemplos que van a ser registrados e incorporados a los materiales que estarán a disposición del equipo redactor. No siempre están claros los criterios con los que se realizan estas operaciones, pero es evidente que tienen una enorme importancia debido a que, en la organización habitual del trabajo, quienes se ocupan de la redacción de la entrada solo van a ver lo que previamente haya sido seleccionado, casi siempre por otras personas.

En un diccionario basado en corpus, en cambio, el conjunto de textos examinados puede ser muchísimo mayor y es posible analizar todos los casos de las palabras que vayan a estar en el diccionario contenidos en el corpus. Para no perdernos en vías secundarias al tema de este artículo, demos por resueltas –aunque sin olvidar que constituyen un problema de gran carga teórica y práctica– todas las cuestiones relacionadas con la lematización de los textos. Supongamos también que los redactores de una entrada pueden obtener todos los ejemplos que les interesan en cada caso de forma equivalente a como lo harían mediante la consulta directa de los textos a cargo de una persona experta. Conviene tener en cuenta, en primer lugar, que el corpus sobre el que se va a construir el diccionario puede tener un volumen muchísimo mayor que el

2 Para valorar adecuadamente este punto no puede olvidarse que las palabras de Sinclair reproducidas arriba están especialmente referidas a diccionarios del tipo del COBUILD, es decir, diccionarios para aprendices de inglés.

3 Las circulares de Murray y algunos otros documentos de gran interés para la historia de la lexicografía en general pueden verse en la página electrónica del OED: http://www.oed.com/archive/ [comprobado el 13/8/2009].

conjunto de textos despojados en la mayor parte de los procesos lexicográficos. No siempre ha sido así, por supuesto. El primer diccionario basado en corpus, el COBUILD, fue proyectado con un corpus de 7,5 millones de formas, que, ante los problemas surgidos para lograr documentación adecuada, tuvo que ser ampliado a 20 millones. Esas cifras son muy inferiores a las que manejamos hoy,[4] pero no es ese, me parece, el factor crucial. Lo realmente importante es que la construcción de un corpus, que es un proceso, largo, complejo y costoso, se hace con un propósito general, no únicamente para elaborar un diccionario (ni una gramática, por supuesto). Por tanto, las posibilidades van desde la construcción de un corpus específico hasta la utilización total o parcial de un corpus general o el corpus resultante de un proyecto diferente, con lo que los costes y el tiempo necesario son mucho menores. No hace falta insistir en que esa preparación previa y ajena al proyecto lexicográfico puede incluir todo lo relacionado con la anotación y codificación de los textos.

Dado lo anterior, surgen dos problemas diferentes, contrarios entre sí en cierto modo, pero derivados ambos de la naturaleza estadística de los textos reales -y, por tanto, de los corpus-. En primer lugar, en las palabras más frecuentes, la cantidad de ejemplos puede resultar absolutamente inmanejable para el equipo redactor. Según los datos contenidos en Davies (2006), la palabra *hoja*, que tiene una frecuencia no excesivamente alta (ocupa la posición 1000) aparece 1781 veces en el corpus de veinte millones de formas que utiliza como base del diccionario de frecuencias, esto es, tiene una frecuencia de 89 por millón. Ciertamente resultaría imposible revisar todos los casos que *hoja* y todas las palabras de frecuencia similar o superior presentan en un corpus de 100 millones o más, pero no se debe deducir de ahí que las palabras de frecuencia alta no sean casos adecuados para la utilización de corpus. No es difícil conseguir por medios computacionales que los lexicógrafos puedan, en estos casos, trabajar con una selección de ejemplos que tenga el tamaño adecuado. Esa selección puede proceder de una reducción puramente aleatoria, pero también es posible, si el corpus está anotado, conseguir por procedimientos automáticos una muestra representativa de los usos de la palabra en cuestión en lo referente a los contornos sintácticos en que aparece y a los tipos de textos en los que se da. Por tanto, el número excesivo de casos es en realidad un pseudoproblema que se puede solucionar. Sin duda, conviene tener la mayor cantidad posible de ejemplos como punto de partida, puesto que la distribución de frecuencias se presenta del mismo modo por todas partes y también las palabras de alta frecuencia presentan acepciones o contextos sintácticos muy poco frecuentes.

El segundo problema es la cara opuesta del anterior: la mayoría de las palabras –y también de las acepciones– tienen una frecuencia baja o muy baja, de modo que el corpus debe tener un gran tamaño para dar garantías relativas de que vamos a poder encontrar en él todo aquello que nos interesa. Un corpus bien diseñado debe mostrar muchos casos de los fenómenos más frecuentes y muy pocos (o incluso ninguno) de los poco frecuentes. Desde este punto de vista, la selección que prima lo diferencial, lo raro, lo extraño, proporciona una visión más amplia de lo que se puede dar en una lengua. Por supuesto, las características estadísticas de los textos, con su enorme peso de algunos elementos y la escasez o ausencia de otros, tienen su cara positiva, puesto que nos dan la realidad de lo que sucede frente a la distorsión que puede suponer una selección que prima lo extraño. Aun así, conviene tener siempre en cuenta que, según las características del diccionario que se está construyendo, la consulta de fuentes complementarias puede ser muy aconsejable o incluso imprescindible. Esas mismas características estadísticas de los textos hacen que no resulte extraño que un fichero de ejemplos construido al estilo tradicional pueda resultar más rico en la documentación de los elementos más raros, que son en muchos casos aquellos a los que se ha prestado más atención en la fase de selección del material. Como contrapartida, la visión general que proporciona

4 La segunda edición del COBUILD (1995) pudo utilizar ya el *Bank of English*, que constaba en aquel momento de 200 millones de formas.

este sistema de trabajo puede distorsionar la realidad de lo que se da en una lengua, puesto que tiende a infravalorar lo muy común y sobrerrepresenta lo poco habitual.

Por otra parte, trabajar con todo lo contenido en un corpus amplio y representativo es el único modo de aproximarse a la descripción completa, exhaustiva, de las características de un fenómeno gramatical o el comportamiento de una expresión en un determinado estado de lengua, esto es, lo que Leech (1992) y Quirk (1992), entre otros, han denominado la 'total accountability'. En palabras de Leech, los datos presentes en un corpus

> are used exhaustively: there is no prior selection of data which we are meant to be accounting for and data we have decided to ignore as irrelevant to our theory. This principle of "total accountability" for the available observed data is an important strength of CCL [= *computer corpus linguistics*, G.R.] (Leech, 1992: 112).

No se trata, por tanto, del simple hecho de obtener los ejemplos de un corpus más o menos amplio en lugar de obtenerlos de un fichero, sino de enfrentarse a todo lo que contiene el corpus sobre aquello que vamos a estudiar. Para Quirk, es posible que los gramáticos (o los lexicógrafos) usen un corpus

> as a convenient source for "good examples" to put in their grammar. But that is not where the value or the challenge of a corpus will lie. If we ignore the value and evade the challenge of total accountability, our use of a corpus will be no advance on Jespersen's use of his voluminous collections of slips or Murray's use of those file boxes bursting with marked-up quotations for the *OED*. Such scholars certainly ensured that everything in their published volumes was firmly anchored in textual reality, but not that everything in their samples of textual reality was reflected in those published volumes (Quirk, 1992: 467).

El uso de un corpus adecuado es la forma más aconsejable de llevar a cabo la selección de las entradas en función de las características del diccionario que se vaya a construir. Si está debidamente codificado, el corpus da la frecuencia de los lemas que contiene, pero también proporciona la frecuencia de ese lema en cada uno de los subcorpus que puedan diferenciarse en su interior. Así, en un corpus de propósito general como el Corpus de Referencia del Español Actual (CREA)[5] es posible hallar las frecuencias absolutas y relativas de una palabra por países, tramos temporales, tipo de texto, área temática o cualquier combinación de dos o más factores, con lo que se hace posible incorporar marcas de uso a las acepciones. Añadiendo a un corpus de este tipo la información léxica, morfológica y sintáctica necesaria, las herramientas computacionales adecuadas pueden proporcionar los contextos sintácticos característicos de cada palabra, de nuevo con las frecuencias absolutas y relativas en cada bloque de textos. Incorporando información sobre la acepción, es posible cruzar todos estos datos para obtener el perfil completo de uso de un lema. Por supuesto, cada una de estas operaciones tiene un coste, pero, de nuevo, el esfuerzo necesario para llevarla a cabo no se hace solo para construir un diccionario, de modo que incluso es posible utilizar directamente corpus generales o bien corpus construidos para otros proyectos.

Por fin, los datos contenidos en un corpus adecuado son los que permiten acometer la redacción de las entradas con mayor garantía de que el resultado va a reflejar lo que sucede en la lengua sobre la que se trabaja, que es lo que constituye el objetivo básico de un proyecto de este tipo. Frente al tan frecuente problema de que los diccionarios integran de forma acrítica lo que está ya en otros, mantienen palabras o acepciones que ya no se usan y omiten otras de uso corriente, lo esperable en un diccionario basado en corpus es que aspire a incorporar todo lo que se encuentra en el corpus o subcorpus en el que se basa (quizá con un corte por frecuencia) y solo de lo que se aparece en ese corpus, objetivo que puede mantenerse con todas sus consecuencias únicamente si el corpus tiene el diseño y tamaño adecuados a la

5 Puede consultarse en la página de la Real Academia Española http://www.rae.es.

finalidad con que se elabora el diccionario.

En resumen, construir el diccionario sobre los datos procedentes de un corpus permite, entre otras cosas:

- Seleccionar las palabras y acepciones realmente usadas en la lengua o variedad lingüística a la que se refiere el diccionario, dejando a un lado todo aquello que los diccionarios contienen como herencia de obras anteriores aceptada acríticamente y, al tiempo, incorporando aquello que podría quedarse fuera por falta de documentación o por la aplicación de criterios ajenos al uso real.
- Detectar los significados que presenta realmente cada palabra en el corpus. A partir de ahí es posible tomar decisiones bien fundamentadas acerca de las demás: necesidad de documentación a partir de materiales complementarios, por ejemplo. No parece válido, en cambio, utilizar el rasgo de presencia o ausencia de una palabra en el corpus para, en caso positivo, incorporar todas las acepciones de esa palabra que figuran en todos los demás diccionarios.
- Dar información sobre la frecuencia de cada palabra. El COBUILD contiene, a partir de su segunda edición (1995) una indicación aproximada de la mayor o menor frecuencia de las palabras que figuran en él. El paso siguiente sería dar esta información para cada una de las acepciones diferenciadas.
- Identificar las características sintácticas del uso de cada palabra en cada una de sus acepciones. Esa información, debidamente reconvertida en función de los objetivos del diccionario y las características de sus destinatarios, se convierte en un elemento fundamental, sobre todo en la faceta del diccionario que muestra cómo se usa una palabra.
- Añadir las marcas técnicas adecuadas (ámbito geográfico, temático, etc.) a partir de las características que la palabra presenta en el corpus. Naturalmente, lo anterior vale también para cada una de sus acepciones.
- Dar ejemplos reales del uso de las palabras en cada una de sus acepciones. Dado que hablamos de diccionarios basados en corpus, los ejemplos no tienen la misión de documentar el uso en cuestión y justificar su inclusión en la obra, puesto que eso se da por garantizado. Los ejemplos son ilustrativos de la forma en que se emplea cada palabra.

Por supuesto, no se dice aquí que solo los diccionarios basados en corpus contengan esas informaciones. La novedad fundamental que aportan no está en el tipo de datos que contienen, sino en la fiabilidad de las fuentes y el carácter exhaustivo con que son analizadas. Es evidente que estas características se han hecho posibles solo a partir de la utilización de computadoras, que permiten construir corpus formados por cientos de millones de palabras, obtener con rapidez y comodidad la información que nos interesa en cada caso y procesarla de modo que los equipos lexicográficos puedan trabajar con todas las ventajas que proporcionan las grandes masas de documentación.

Referencias bibliográficas

Atkins, B. T. S. & M. Rundell (2008). *The Oxford Guide to Practical Lexicography*, Oxford University Press, 2008.

Davies, M. (2006). *Frequency Dictionary of Spanish*, Nueva York y Londres: Routledge, 2006.

Dyson, F. (1997). *Imagined Worlds*, Harvard University Press, 1997. Cito por la trad. esp. de Joandomènec Ros: *Mundos del futuro*, Barcelona: Crítica, 1998.

Kuhn, T. S. *The Structure of Scientific Revolutions*, University of Chicago Press, 1962. Hay trad. esp. de Agustín Contín: *La estructura de las revoluciones científicas*, México D.F.: Fondo de Cultura Económica, 1971.

Leech, G. (1992). “Corpora and Theories of Linguistic Performance”, en Svartvik (1992), 105-122.

Quirk, R. (1992). contenidos en el corpus. “On Corpus Principles and Design”, en Svartvik (1992), 457-469.

Sinclair, J. (1987). “Introduction” a *Collins Cobuild English Language Dictionary*, Londres: HarperCollins Publishers, 1987, xv-xxi.

Svartvik, J. (ed.) (1992). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (= Trends in Linguistics. Studies and Monographs, 65), Berlin: Mouton - de Gruyter, 1992.

Zgusta, L. (1970). *Manual of Lexicography*, La Haya: Mouton, 1970.