DURÁN MUÑOZ, I.2011. "Criterios específicos para la elaboración y diseño de los corpus especializados para la terminografía". En Carrió Pastor, M. L. y Candel Mora, M. A. *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora*. Actas del III Congreso Internacional de Lingüística de Corpus. Valencia: Universitat Politècnica de València. 43-50.

Criterios específicos para la elaboración y diseño de los corpus especializados en Terminografía

Resumen: La especificidad de la Terminografía basada en corpus (Meyer y Mackintosh, 1996: 258), en contraposición a la Lexicografía basada en corpus u otras aplicaciones de los corpus (traducción, enseñanza de segundas lenguas, etc.), obliga al establecimiento de una serie de requisitos o criterios específicos para el trabajo terminográfico. Algunos de ellos serán comunes a los criterios generales de la compilación y diseño de los corpus y otros, como veremos, presentarán algunas diferencias. En este trabajo, pretendemos ilustrar estas diferencias y determinar las características propias que debe presentar un corpus compilado con un objetivo terminográfico, con objeto de mejorar los resultados de cualquier trabajo terminográfico.
Palabras clave: terminografía, corpus especializado, terminografía basada en corpus, representatividad.

Abstract: Specifity in corpus-based Terminology (Meyer y Mackintosh, 1996: 258), in comparison to corpus-based Lexicography and other corpus-based studies (on translation, second-language acquisition, etc.), requires the establishment of a series of specific criteria to carry out a terminology work. Some of these criteria coincide with the criteria established to design and compile corpora in general but other are different and need to be taken into account. In this paper, we pretend to illustrate these differences and determine the own features that a corpus compiled in the framework of a terminology work should present, with the aim to obtain a corpus adequate to our terminological necessities.
Keywords: terminography, specialised corpus, corpus-based terminography, representativity.

1. LA TERMINOGRAFÍA BASADA EN CORPUS

La Lingüística de Corpus ha sido una de las disciplinas lingüísticas que más ha influido en la Terminología y, por ende, en la *Terminografía*. Tanto es así que hoy en día la idea de la contextualización de los términos a través de los corpus textuales está totalmente aceptada en la comunidad de terminógrafos, y los corpus[1] son considerados recursos indispensables para cualquier trabajo de esta naturaleza. De forma general, podemos afirmar que en *Terminografía* el uso de los corpus textuales se ve motivado por dos motivos fundamentales:

Por un lado, el trabajo terminográfico no consiste en la invención de denominaciones para unos conceptos previamente establecidos como propugnaban la

[1] En *terminografía*, el tipo de corpus utilizado se denomina *corpus especializado*, al tratarse de un tipo de corpus especial que ha sido diseñado con un propósito específico y que tiene la finalidad de ser representativo de un tipo particular de lengua, como por ejemplo un campo de especialidad o un grupo particular de hablantes.

terminología tradicional, sino en «la identificación y recopilación de los términos que los especialistas utilizan en realidad» (Cabré Castellví, 1993: 113). Por este motivo, si un terminógrafo debe estudiar los términos que los especialistas utilizan en su trabajo diario, deberá consultar directamente con los especialistas del campo de especialidad en cuestión o realizar un estudio detallado de las producciones lingüísticas que estos especialistas crean para comunicarse entre ellos o con otros actores. De estas dos opciones, la primera no es siempre posible, ya que puede resultar complicado disponer de los especialistas adecuados y, cuando se puede acceder a ellos, a menudo encuentran dificultades a la hora de explicar el significado y el uso del lenguaje que emplean, al fin y al cabo, de forma intuitiva. En palabras de Meyer y Mackintosh (1996: 264):

> While experts obviously *know* their domains, they do not all *explain* their knowledge *clearly* (whether orally or in writing), *completely* (it is up to the knowledge acquirer to make sure that all important areas in the field are covered), or *consistently* (experts often disagree with each other or change their minds).

Por ello, la segunda opción, la consulta de documentación especializada en forma de corpus textual, es más accesible, rápida y directa y, por tanto, determinante en el trabajo terminológico.

Por otro lado, se hace imprescindible el uso del corpus en *Terminografía* en la dimensión conceptual. Para poder identificar y recopilar los términos que los especialistas emplean en la realidad, que se incluirán en el recurso terminológico, los terminógrafos necesitan estudiar las estructuras de conocimiento (conceptos y sus relaciones) y familiarizarse con el tema específico de su trabajo, lo que Cabré Castellví (1999: 144) denomina «competencia cognitiva». De la misma forma que en el caso anterior, los terminógrafos pueden dirigirse a especialistas en el ámbito de especialidad en cuestión o consultar documentación especializada y, de nuevo, en la mayoría de las ocasiones, será más fácil familiarizarse con el ámbito de especialidad, con sus conceptos y estructuras a través de la documentación especializada y consultar puntualmente a los especialistas.

La necesidad de la utilización de documentación especializada en el trabajo terminográfico sistemático queda patente con estas dos razones expuestas. Y, gracias a los avances que se han producido principalmente en el ámbito de la informática que han permitido tener a disposición gran cantidad de información en formato electrónico y

poder procesarla de forma automática o semiautomática, el uso de los corpus electrónicos se ha extendido en la actualidad, convirtiéndose en la herramienta esencial para la mayoría de los trabajos terminográficos y dando lugar a lo que Leech (1992: 106) considera «a new way of thinking about language».

2. EL CORPUS EN LAS FASES DEL TRABAJO TERMINOGRÁFICO

Como hemos visto en el apartado anterior, el corpus se ha convertido en una herramienta esencial para el trabajo terminográfico, puesto que hace posible la consulta y el procesamiento de grandes cantidades de información en un tiempo muy reducido y, a menudo, de forma automática o semiautomática. Sin embargo, las ventajas que aporta el corpus en *terminografía* no se limita a la consulta de información en la fase de elaboración de la terminología, sino que se trata de una herramienta que acompaña al terminógrafo en todas las fases de su tarea, es decir, se utiliza desde la fase inicial de preparación del proyecto hasta la fase final de validación. Así pues, la documentación, y en consecuencia, los corpus y las herramientas de análisis de corpus permiten llevar a cabo:

- La adquisición de conocimiento conceptual y la familiarización del terminógrafo no experto del dominio mediante la consulta de documentos, a fin de identificar la estructura interna del dominio, las relaciones con otros campos de especialidad y las fuentes de conocimientos adecuadas.
- La identificación de unidades terminológicas dentro de un discurso de especialidad, es decir, la nomenclatura.
- El análisis y la preparación de entradas mediante la adquisición de información lingüística, pragmática y semántica extraídas del corpus, como por ejemplo definiciones, contextos o colocaciones.
- La detección y descripción de nuevos conceptos, así como la identificación de etiquetas léxicas que se están atribuyendo a dichos conceptos dentro del dominio. En algunos casos, también la propuesta de un neologismo adecuado (cuando todavía no se ha acuñado un término en una lengua).

- La estandarización de términos sinónimos o cuasisinónimos que son utilizados por diferentes expertos.
- Y la clarificación de dudas y preguntas sobre inconsistencias y otros asuntos que, de lo contrario, solo se podría llevar a cabo mediante consultas a expertos del dominio.

Llegados a este punto, podemos resumir los roles de la documentación en dos aspectos: en primer lugar, para el análisis lingüístico (extracción de términos, elaboración de definiciones, etc.), y, en segundo lugar, para la adquisición de conocimiento (conceptualización del dominio, relaciones semánticas, etc.). Para poder extraer adecuadamente la información lingüística, pragmática y conceptual de los corpus, será necesario que estos estén compilados de forma apropiada y según unos criterios generales y específicos establecidos, teniendo en cuenta que «the corpus needs to be as linguistically and conceptually rich as possible» (Meyer y Mackintosh, 1996: 266).

3. CRITERIOS PARA LA COMPILACIÓN DE CORPUS ESPECIALIZADOS

En *Terminografía*, a pesar de las ventajas reconocidas que aporta el uso de corpus electrónicos en el estudio del lenguaje en uso, esta herramienta se ha incorporado muy recientemente (Meyer y Mackintosh, 1996: 257). Asimismo, la *Terminografía basada en corpus* sigue utilizando la metodología y herramientas de la Lexicografía basada en corpus, lo que no siempre es recomendable como nos indican las autoras (ibíd.: 258):

> Hence, it is essentials that terminographers begin to specify the types of corpora they need and the corpus-analysis tools that best suit their task. […] in broad terms, the *specificity* of corpus terminography. We shall argue that terminography and lexicography are in many ways substantially different, and consequently, that the tools and techniques developed in corpus lexicography cannot be applied intact to terminography. On the contrary, some can actually be *dangerous*.

En este contexto, es esencial que, para desarrollar una *terminografía basada en corpus* que mejore la calidad de la investigación terminológica, se concierten un conjunto de

técnicas y herramientas propias que permitan a esta disciplina realizar sus estudios y trabajos sobre la Terminología de los lenguajes de especialidad y, así, abandonar las herramientas desarrolladas propiamente para la Lexicografía. Uno de los aspectos que debemos tener en cuenta en este sentido son los criterios de compilación de los corpus, de los cuales algunos coincidirán con los seguidos en lexicografía pero otros serán específicos de la *terminografía basada en corpus*.

3.1. Criterios generales

Dentro de los criterios generales de compilación, determinamos cuatro: cantidad, calidad, documentación y simplicidad. Estos criterios son comunes a cualquier rama lingüística que utilice corpus textuales, aunque presentarán unas características concretas en cada una. A continuación, indicamos sus características en el contexto terminográfico:

El criterio de la cantidad es un aspecto polémico para la compilación de *corpus especializado*s. Para algunos autores (Meyer y Mackintosh, 1996: 268), un *corpus especializado* puede ser mucho más pequeño que uno de propósito general, pero para otros no sería necesario poner un límite a la cantidad de texto que se recopile (Pearson, 1998: 57), siempre y cuando siga unos criterios establecidos previamente. En definitiva, no existe un acuerdo acerca del volumen que debe tener un *corpus especializado* para que sea representativo. Desde la lexicografía basada en corpus se recomienda que el volumen del corpus sea muy elevado, pero en *terminografía* resulta más importante la densidad terminológica que el volumen propio del corpus utilizado, por lo que el número de palabras no significa que sea más o menos representativo. En este sentido, no es posible establecer a priori el número de textos o de palabras que necesita un corpus para ser representativo, aunque sí es posible medir su *representatividad* de forma objetiva y cualitativa posteriormente mediante el uso de la aplicación informática *ReCor*, que permite estimar la *representatividad* de los corpus en función de su densidad terminológica una vez compilados que presenta un corpus electrónico (cf. Seghiri Domínguez, 2006; Corpas Pastor y Seghiri Domínguez, 2007*a*, 2007*b*).

Con respecto al criterio de calidad, los textos deberán cumplir una serie de requisitos: primero, los textos deben ser recientes y actuales, es decir, se debe compilar un corpus sincrónico formado por textos que representen el estado actual del dominio, a fin de proporcionar los términos utilizados en el campo de especialidad en un momento concreto y la información conceptual actualizada; segundo, los textos deberán estar escritos por diferentes autores expertos (ya sean especialistas, instituciones, organizaciones, etc.) en la medida de lo posible para ofrecer fidelidad en el contenido y neutralizar cualquier tipo de idiosincrasia del autor; tercero, los textos deben ser preferentemente originales y no traducciones, ya que solo de esta manera se puede asegurar que la terminología empleada es la adecuada y la original en cada ámbito especializado,[2] así como es recomendable también que los textos estén escritos por hablantes nativos, para evitar cualquier tipo de influencia de la lengua materna del autor, errores (léxico, gramaticales, etc.) y cualquier giro o uso incorrecto; cuarto, los textos incluidos en los corpus deben ser textos completos, es decir, se debe evitar el uso de fragmentos con objeto de evitar la posible descontextualización o error en la información contenida; y, por último, según Pearson (1998: 60), los textos deben haber sido publicados previamente a su inclusión, con la idea de que adquieran un valor más formal, serio y respetable dentro del ámbito especializado. Asimismo, si el trabajo terminográfico está enfocado para unos países determinados, se deberá establecer unos límites geográficos para la búsqueda de los textos y, por tanto, el terminógrafo deberá comprobar que todos los textos pertenecen o han sido publicados dentro de esos límites geográficos. Con todos estos criterios de calidad se pretende conseguir la mayor fiabilidad posible de los textos incluidos en un corpus con un objetivo terminográfico.

El criterio de documentación ocupa un lugar muy relevante en la compilación, ya que permite llevar a cabo un trabajo sistemático y homogéneo. Consiste en registrar las referencias de los textos contenidos en el corpus (autor, lugar de publicación, fecha o actualización, etc.) con objeto de realizar un seguimiento de los textos seleccionados y de las fuentes de información utilizadas. También es recomendable utilizar un código

[2] A pesar de que el corpus compilado para un trabajo terminográfico debe ser comparable (textos originales) para poder extraer la terminología adecuada del ámbito de especialidad en cada lengua de trabajo, también suele haber un subcorpus paralelo (textos originales y sus traducciones) que permite detectar y extraer los posibles equivalentes en otras lenguas.

unívoco entre el registro del texto y el propio texto para establecer una vinculación entre los datos registrados y el texto que nos permita acceder a su información de referencias.

Por último, el criterio de simplicidad hace referencia al tipo de información añadida al texto original. Esta información es principalmente morfológica, sintáctica, léxica y semántica y permite al terminógrafo utilizar el corpus de forma más eficiente y más precisa a la hora de realizar búsquedas, estudios concretos y clasificar la información contenida en los textos. En *terminografía*, las anotaciones más utilizadas son las semánticas y léxicas, que se utilizan para el estudio del discurso especializado, extracciones terminológicas o de patrones semánticos, aunque también se puede encontrar otro tipo de estudios lingüísticos.

3.2. Criterios específicos

Además de estos criterios generales expuestos anteriormente, consideramos que son útiles otros, de carácter más específico, aunque muy relacionados con los anteriores:

El criterio de delimitación de fronteras, que exige que, antes de comenzar con cualquier proyecto, los terminógrafos deben delimitar el campo de especialidad que tiene que representar el corpus de dos formas diferentes: «side boundaries», es decir, con respecto a los campos de especialidad más cercanos al objeto de estudio y «upper boudaries», con relación a los niveles de especialización que va a incluir el corpus (de más especializado a menos) (Meyer y Mackintosh, 1996). En relación con este criterio, los terminógrafos deberán compilar un corpus que represente lo mejor posible el campo de especialidad delimitado y, para ello, deberán seleccionar textos que permitan un equilibrio en todos los aspectos del dominio, cubriendo todos los subdominios y dominios relacionados de la forma más equitativa posible.

Otro criterio específico hace referencia a la apertura del corpus. En *Terminografía*, debido a la rapidez con la que se producen los cambios, el *corpus especializado* debe ser abierto, a fin de que se pueda ir actualizando, ya sea eliminando o incluyendo textos, con el paso del tiempo.

Por último, encontramos el criterio de pragmática, que se encuentra en la misma línea que el criterio de delimitación de fronteras, aunque en este caso hace referencia a la situación comunicativa para la que va destinado el producto final. Este criterio exige tener en cuenta la situación comunicativa, a saber: los receptores, el contexto, el tipo textual y el nivel de especialización que habrá en el momento en el que se emplee el recurso final. Así pues, dependiendo de quiénes sean los usuarios, para qué utilicen y cuándo utilicen el recurso final, el corpus estará formado por un tipo de textos u otros.

En nuestra opinión, estos serían los criterios que deberían tenerse en cuenta para compilar cualquier *corpus especializado* en un contexto terminográfico y, en definitiva, para garantizar la calidad del producto final.

4. CONCLUSIONES

Como hemos visto, la *terminografía basada en corpus* se trata de la *terminografía* de actualidad, fruto del cambio de paradigma sufrido en la terminología que ha supuesto el paso de la terminología tradicional de Wüster a la terminología moderna, así como de la evolución de las herramientas informática que han facilitado el procesamiento de grandes cantidades de información en formato electrónico. Sin embargo, a pesar de lo extendido de su uso, aún seguimos muy vinculados a la metodología utilizada en la lexicografía basada en corpus, lo que limita a veces el trabajo terminográfico.

Por este motivo, necesitamos tener en cuenta las especificaciones de la *terminografía* para establecer una metodología propia que nos permita alcanzar los mejores resultados en nuestros proyectos y en las diferentes fases del trabajo terminográfico. En esta línea, hemos propuesto unos criterios generales y específicos dirigidos a la compilación de *corpus especializado*s para el trabajo terminográfico que nos permita extraer los mejores beneficios para el objetivo de nuestro proyecto, que a menudo difiere de los objetivos perseguidos en la lexicografía basada en corpus.

5. BIBLIOGRAFÍA

Cabré Castellví, M. T. **(**1993). *La terminología. Teoría, metodología, aplicaciones.* Barcelona: Antártida/Empúries.

Cabré Castellví, M. T. (1999). Hacia una teoría comunicativa de la terminología: aspectos metodológicos. En M. T. Cabré. 2000 (Ed.). *La Terminología: Representación y Comunicación. Elementos para una teoría de base comunicativa y otros artículos* (pp. 129-150). Barcelona: IULA. Universidad Pompeu Fabra.

Corpas Pastor, G. y Seghiri Domínguez, M. (2007a). Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor. *Procesamiento del lenguaje natural*, *39*, 165-172.

Corpas Pastor, G. y Seghiri Domínguez, M. (2007b). Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. *Translation Journal*, *11*(3). Disponible en http://www.translationjournal.net/journal/41corpus.htm

Leech, G. (1992). Corpora and Theories of Linguistic Performance. En J. Svartvik (Ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* (pp. 105-134). Berlín/Nueva York: Mouton de Gruyter.

Meyer, I. y Mackintosh, K. (1996). The Corpus from a Terminographer's Viewpoint. *International Journal of Corpus Linguistics*, *1/2*, 257-285.

Seghiri Domínguez, M. (2006). *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. Tesis Doctoral. Málaga: Servicio de Publicaciones de la Universidad de Málaga.