DISEÑO DE UN CORPUS ESPECIALIZADO CON FINES TERMINOGRÁFICOS: EL CORPUS DE LA PIEDRA NATURAL

Chelo Vargas Sierra (Universidad de Alicante, España)

RESUMEN:
EL OBJETIVO DE ESTE ARTÍCULO ES EL DE PRESENTAR LAS ESPECIFICACIONES Y EL DISEÑO DEL CORPUS DE LA PIEDRA NATURAL (CPN) CON EL FIN DE QUE SIRVA DE MODELO O GUÍA PARA TERMINÓLOGOS QUE PRETENDEN CONSTRUIR APLICACIONES TERMINOLÓGICAS DIRIGIDAS AL TRADUCTOR DE TEXTOS DE ESPECIALIDAD. PARA ELLO, EN LA PRIMERA PARTE, NOS CENTRAREMOS EN UNA BREVE REVISIÓN DE LOS CONCEPTOS Y CRITERIOS QUE AYUDARÁN A ENMARCAR ESTE TRABAJO EN EL PLANO TEÓRICO. EN LA SEGUNDA PARTE, TRAS PRESENTAR EL MARCO INVESTIGADOR DEL CPN, PROPORCIONAREMOS LAS FASES DISEÑADAS PARA SU CONSTRUCCIÓN, LA TIPOLOGÍA DE CÓRPORA POSIBLES Y, FINALMENTE, LOS CRITERIOS Y PARÁMETROS QUE HEMOS ADOPTADO PARA CONCEBIR EL CPN Y QUE, AL TIEMPO, LO SINGULARIZAN. DICHOS CRITERIOS SE DESARROLLARÁN DE UN MODO SECUENCIAL HASTA CONCLUIR CON UNA VISIÓN SINTÉTICA Y RESUMIDA DE LOS MISMOS.

PALABRAS CLAVE:
TERMINOGRAFÍA; CORPUS ESPECIALIZADO; DISEÑO DE CORPUS

RESUMO:
O OBJETIVO DESTE ARTIGO É O DE APRESENTAR AS ESPECIFICAÇÕES E O DESENHO DO CORPUS DA PEDRA NATURAL (CPN) COM O FIM DE SERVIR DE MODELO OU GUIA PARA TERMINÓLOGOS QUE PRETENDEM CONSTRUIR APLICAÇÕES TERMINOLÓGICAS DIRIGIDAS AO TRADUTOR DE TEXTOS DE ESPECIALIDADE. PARA ISSO, NA PRIMEIRA PARTE, CENTRAREMOS-NOS NUMA BREVE REVISÃO DOS CONCEITOS E CRITÉRIOS QUE AJUDARÃO A EMOLDURAR ESTE TRABALHO NO PLANO TEÓRICO. NA SEGUNDA PARTE, DEPOIS DE APRESENTAR O ÂMBITO INVESTIGADOR DO CPN, PROPORCIONAREMOS AS FASES PROPOSTAS PARA SUA CONSTRUÇÃO, A TIPOLOGIA DE CÓRPORA POSSÍVEIS E, FINALMENTE, OS CRITÉRIOS E PARÂMETROS QUE ADOTAMOS PARA CONCEBER O CPN E QUE, AO MESMO TEMPO, SINGULARIZAM-NO. DITOS CRITÉRIOS SE DESENVOLVERÃO DE UM MODO SEQÜENCIAL ATÉ CONCLUIR COM UMA VISÃO SINTÉTICA E RESUMIDA DOS MESMOS.

PALAVRAS-CHAVE:
TERMINOGRAFIA; CORPUS ESPECIALIZADO; CONSTRUÇAO DE CORPUS

1. INTRODUCCIÓN

La compilación de córpora se ha convertido en una actividad en la que se involucran cada vez más investigadores, especialmente cuando estos últimos llevan a cabo o están integrados en proyectos dedicados al estudio de la lengua en uso. En el campo concreto de los lenguajes profesionales y académicos (LPA), la observación del lenguaje objeto de estudio tal y como es utilizado en sus contextos específicos es vista como una necesidad.

Proliferan en la actualidad los archivos de textos electrónicos y las bibliotecas virtuales. Estos repositorios ofrecen diferentes tipos de textos electrónicos que resultan, de este modo, muy fáciles de obtener. Como consecuencia de esta gran oferta y disponibilidad textual, podría parecer que la compilación de un corpus es una labor de poca envergadura, pues para construirlo bastaría con hacer acopio de una colección de textos extraídos de una fuente de datos electrónica (Internet, CD-ROM, disquetes…) e incorporar estos datos de forma inmediata al corpus. Sin embargo, nada más lejos de la realidad. La compilación rigurosa de un corpus electrónico implica tomar una serie de decisiones que, en ocasiones, resultan verdaderamente complejas. Dichas decisiones están directamente relacionadas con el propósito concreto del corpus y se refieren, entre otras, a consideraciones como el canal de producción de los textos, el tamaño final, la distribución textual, el contenido, el tamaño de las muestras, el método de muestreo, etc.

Con este trabajo pretendemos presentar las especificaciones y el diseño del corpus de la piedra natural (CPN) con el fin de servir de modelo o guía para terminólogos que quieran construir recursos lingüísticos para el traductor técnico. Para ello, en primer lugar, revisaremos brevemente los conceptos y criterios que ayudan a enmarcar este trabajo en el plano teórico. En la segunda parte, tras presentar el marco investigador del CPN, proporcionaremos las fases diseñadas para su construcción, la tipología de córpora posibles y, finalmente, los criterios y parámetros que hemos adoptado para concebir nuestro corpus y que, al tiempo, lo singularizan.

2. MARCO DE REFERENCIA

El marco en el que actualmente tiene lugar la investigación y el desarrollo científico cuyo objeto de estudio son los lenguajes especializados y, dentro de los mismos, la terminología, se sostiene sobre la base de principios teóricos y metodológicos que destacan la importancia del uso de córpora lingüísticos, pues un corpus se concibe como «a more powerful methodology from the point of view of the scientific method, as it is open to objective verification of results» (MCENERY y WILSON, 1996: 13). Así, la terminología y, más concretamente, su actividad práctica, esto es, la terminografía, se beneficia de la información que puede aportar un corpus a la hora de elaborar aplicaciones terminográficas, pues a partir de este material se facilita la tarea de descubrir nuevas palabras o palabras en desuso, observar los distintos significados de una misma forma léxica, detectar colocaciones o combinaciones terminológicas, obtener ejemplos reales de uso, definiciones, etc. Partimos de la hipótesis de que una aplicación terminológica (diccionarios, bases de datos, tesauros, etc.) debe crearse a partir de un corpus de textos profesionales y académicos específicos del ámbito técnico objeto de estudio. Su utilidad se fundamenta en que lo entendemos como un "ecosistema", siendo las diferentes unidades léxicas y otros elementos las especies que habitan en él. Las palabras pueden estudiarse también *in vitro*, pero si se quiere conocer el modo en que se comportan se deben observar *in vivo*, en su entorno natural, y el entorno natural de las unidades léxicas son los textos.

La lingüística de corpus es un campo de investigación caracterizado en la actualidad por dos aspectos básicos y fundamentales: *a)* estudiar de forma empírica la lengua en uso con el fin de describirla; y *b)* utilizar el ordenador para el almacenamiento y el análisis de los datos. De estos dos aspectos surgen a la vez dos enfoques bien diferenciados. Desde el primero, los lingüistas tienen la finalidad de investigar la lengua a partir de un corpus. De este modo, las herramientas informáticas son meros instrumentos que hacen posible acceder y estudiar mejor y con más rigor la lengua. En el segundo enfoque, encontramos a los informáticos, interesados por el procesamiento del lenguaje natural (PLN) para quienes lo que importa es el desarrollo, la implementación o la prueba de una aplicación informática determinada (programas de concordancias, extractores automáticos de terminología, etiquetadores [morfosintácticos, sintácticos, discursivos, semánticos…], etc.).

La compilación de un corpus no es un fenómeno nuevo, sino que para determinadas disciplinas lingüísticas, como la lexicografía y la enseñanza de lenguas, ha sido una práctica común. El corpus ha sido el modo más eficaz de hacer acopio de datos para la descripción del lenguaje. En la actualidad, sin embargo, la idea de corpus, así como el proceso de creación y explotación del mismo, ha cambiado debido a la revolución tecnológica que empezó experimentarse a finales de los años setenta y que se extiende hasta principios de los ochenta, momento en que se popularizan los ordenadores personales. Esta situación hizo posible el acceso de un mayor número de investigadores al PLN. En los años ochenta se extendió el uso de córpora, pero ahora vinculados de forma determinante a las herramientas informáticas necesarias para su procesamiento y explotación. Este nuevo contexto tecnológico y social contribuyó de manera decisiva al resurgimiento y fortalecimiento de la investigación lingüística basada en corpus:

The resurgence of corpus linguistics can be measured in terms of the increasing power of computers and of the exponentially increasing size of corpora, viewed simplistically as large bodies of computer-readable text (LEECH, 1991: 9-10).

De este modo, el maridaje indisoluble entre el ordenador y la compilación y creación de grandes córpora ha sido un elemento clave para superar las críticas teóricas y prácticas que se vertían sobre la lingüística de corpus. Esta vinculación ha hecho, asimismo, que hoy en día el término *corpus* contenga en sí la característica de «legible por el ordenador» (MCENERY y WILSON, 1996: 23). Además de la popularización del uso del ordenador y de que el corpus adquiriera como característica inherente su estado en formato electrónico, existen otros aspectos que favorecieron el renacer de esta disciplina aplicada, como son: *a)* la mayor disponibilidad de infraestructura tecnológica en manos privadas e institucionales (*ibid.*); *b)* la posibilidad de acceder a córpora en línea (BIBER *et al.*, 1998: IX); *c)* el desarrollo notable de tecnología y programas informáticos específicos (LEECH, 1991: 10); y *d)* el eclecticismo que, a partir de la década de los ochenta, empiezan a adoptar los lingüistas de corpus, que defiende la coexistencia de metodologías cuantitativas con metodologías cualitativas (BIBER *et al.*, 1998; FILLMORE, 1992; STUBBS, 1996, entre otros).

Los primeros córpora especializados fueron construidos con una finalidad didáctica, es decir, se pretendía crear un recurso con el que poder enseñar y aprender los usos específicos que se hacen del lenguaje en un determinado ámbito de especialidad (QI-BO, 1989; ROE, 1977). Otros proyectos de corpus nacieron con el objetivo de estudiar algún aspecto lingüístico concreto de uno o varios lenguajes de especialidad (BACH *et al*, 1997; FABER, 2002). También han proliferado los destinados a desarrollar

herramientas para el PLN (corpus CRATER). Y otros, como el nuestro, emergen con el propósito de elaborar aplicaciones terminológicas para un usuario concreto: el traductor (GÓMEZ y VARGAS, 2004).

En el ámbito concreto de la terminología, los córpora son, por naturaleza, especializados y, dentro de esta clasificación, se trata, más concretamente, de córpora de un ámbito científico, técnico y/o profesional. Meyer y Mackintosh (1996) señalan que la terminología ha sido una disciplina en el que los córpora se han incorporado de forma tardía. Parece evidente que en la investigación del uso de los términos (terminología descriptiva), contrariamente a lo que ocurre en la normalización terminológica (terminología prescriptiva), un corpus sea un recurso de gran utilidad. Sin embargo, el corpus es una tarea que lleva tiempo y, como bien señalan las autoras citadas (*ibid*.: 258):

Terminographers do not have the luxury of waiting years or even months for a corpus to be built. Because of the nature of their jobs, terminographers are overwhelmingly preoccupied with lexical 'newness'; the object of their study is lexical items that very rapidly appear, change, and become obsolete.

A pesar de la celeridad que comporta el trabajo de construcción de corpus para un terminógrafo, los córpora electrónicos, una vez compilados, se convierten, a nuestro entender, en un medio privilegiado e indispensable para observar la estructura de la lengua en uso, para probar hipótesis lingüísticas, conocer los distintos significados que las unidades léxicas, especializadas o generales, adquieren en contexto y descubrir las combinaciones terminológicas más prototípicas de un término dado.

Seguidamente presentamos el marco investigador en el que surge el Corpus de la Piedra Natural (CPN), pues ello nos permitirá comprender las distintas decisiones que tomamos en su construcción y que detallaremos más adelante.

3. EL CORPUS DE LA PIEDRA NATURAL

El CPN surge dentro del marco de un proyecto de investigación subvencionado por el Ministerio de Educación y Cultura denominado *Creación de una base terminológica de algunos sectores industriales de la Comunidad Valenciana* (PB98-0963). El objetivo del aludido proyecto consistía en crear repertorios terminológicos contextualizados de ciertos sectores económicos de la aludida Comunidad (piedra natural, calzado, juguete y textil, principalmente). Estos repertorios se regían por ciertos parámetros concebidos para facilitar la caracterización, el significado, el uso y la traducción de los términos empleados en los citados sectores industriales. El propósito era ofrecer a los sectores nombrados unas herramientas lingüísticas que favorezcan la apertura de mercados internacionales, al tiempo que se proporcionaba a los traductores repertorios especializados fiables y útiles.

Desde los años noventa, el Departamento de Filología Inglesa de la Universidad de Alicante ha venido confeccionado varios diccionarios especializados, publicados por Editorial Ariel de Barcelona, de lo que se conoce con el nombre de *inglés profesional y académico* (IPA) (ALCARAZ, 2000). Entre los diccionarios publicados destacan el de términos jurídicos (1993), el de términos económicos, financieros y comerciales (1995), el de términos de marketing, publicidad y medios de comunicación (1998), el de términos de turismo y ocio (2000), el de la bolsa (2003), el de la propiedad industrial (2003), el del ámbito de los seguros (2003), el de la piedra natural e industrias afines (2005), y el del calzado e industrias afines (2006).

Es claro, por tanto, el interés que muestra el aludido Departamento por la investigación y elaboración de aplicaciones terminológicas bilingües (diccionarios y bases de datos, principalmente) destinadas al traductor de textos de especialidad. Dado el usuario prototípico de nuestras aplicaciones, éstas deben mostrar el uso real de los términos y, para ello, se aporta junto con el equivalente de la lengua meta, los contextos reales en donde se hallan insertos, las combinaciones terminológicas más comunes, definiciones, notas de uso, etc. Todos estos datos lingüísticos, necesarios y útiles durante el proceso de traducción, se extraen de los córpora que se elaboran para cada ámbito especializado que es objeto de estudio.

Dentro del grupo de investigación consolidado IPA de la Universidad de Alicante se han construido y se están construyendo diferentes córpora de textos escritos procedentes de diversos ámbitos especializados. En lo referente al CPN, éste comenzó a ser compilado a principios del año 2001 y nació con la idea de convertirse en una muestra representativa y equilibrada conceptual y pragmáticamente del inglés y el español profesional y académico en modo escrito del sector industrial de la piedra natural. Este corpus se crea para fines de investigación y, más concretamente, para extraer la terminología propia del sector aludido, por lo que los usuarios de dicho corpus eran los terminógrafos que trabajábamos en el proyecto.

Con el objetivo de diseñar y construir el CPN, fueron necesarias algunas decisiones preliminares para fijar las bases de las fases sucesivas. Consideramos que el corpus, que era nuestra materia prima y que de él dependían en gran medida los resultados de la aplicación que íbamos a elaborar, debía ser configurado en distintas etapas, tema que abordaremos a continuación.

4. FASES EN LA CONSTRUCCIÓN DE UN CORPUS TERMINOGRÁFICO

En la etapa de planificación de un corpus se han de definir toda una serie de especificaciones divididas en dos ejes principales: el diseño lingüístico del corpus y la planificación del proyecto en su totalidad (ATKINS *et al*. 1992: 2). Como refieren estos autores, en la planificación del proyecto se consideran cuestiones relativas a: (1) el presupuesto, en donde se han de prever los gastos que ocasionará cada una de las tareas que se realicen en cada una de las fases; (2) el tiempo necesario para llevarlo a cabo; y (3) la materialización o implementación del diseño.

Las etapas o fases previstas que diseñamos para configurar el CPN se dividieron en:

a) Las especificaciones y diseño;
b) El soporte lógico y físico (*hardware* y *software*);
c) La adquisición textual;
d) El procesamiento informático de los textos.

El gráfico siguiente pretende ofrecer una visión integral y sintética de las distintas fases que suponen nuestra propuesta (VARGAS, 2005b) y que, en definitiva, articulan un trabajo terminográfico basado en corpus, fases a partir de las cuales se despliegan una serie de subfases, como se podrá apreciar:

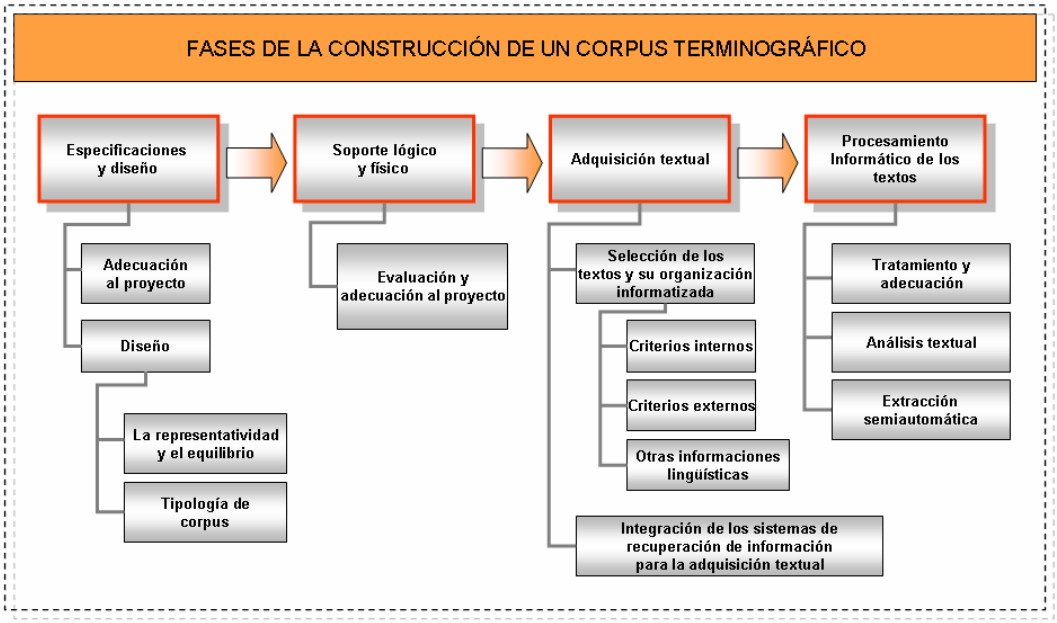

Figura 1: *Fases de la construcción de un corpus terminográfico (Vargas, 2005b)*

Dada la extensión que ocuparía la descripción pormenorizada de las distintas fases y subfases que configuramos para construir nuestro corpus, en este artículo únicamente nos proponemos tratar el proceso referente a las especificaciones y el diseño de un corpus terminográfico.

5. LAS ESPECIFICACIONES Y EL DISEÑO

El diseño de un corpus puede concebirse desde distintos puntos de vista. En cada caso, este diseño debe responder a dos parámetros esenciales (SÁNCHEZ *et al*, 1995: 25):

- los objetivos que se pretenden alcanzar;
- los medios disponibles para llevar a cabo el proyecto.

En la concepción del corpus, se debe tener presente que éste debe servir al propósito específico que se marque el investigador, pues, de otro modo, los resultados y su interpretación podrían resultar defectuosos. La determinación de cuál es la función específica de un corpus subyace a la metodología que se emplee a la hora de recoger las muestras textuales. Hemos de ser capaces de justificar el corpus en términos lingüísticos, por lo que las muestras que lo componen han de seleccionarse en consonancia con unos criterios explícitos concebidos para capturar las regularidades de una lengua, de una variedad o de un sublenguaje, como es nuestro caso. Por tanto, para una investigación como la nuestra, que se marcaba como objetivo extraer la terminología característica de un sector industrial determinado, se hizo necesario configurar un corpus especializado mediante un conjunto de muestras textuales representativas del lenguaje objeto de estudio.

5.1 LA ADECUACIÓN AL PROYECTO

Es claro que la calidad de los resultados de un trabajo con corpus depende en gran medida de la calidad del corpus. Lo anterior implica tener en cuenta multitud factores, si bien podemos agruparlos en dos dimensiones: una referida a aspectos de tipo lingüístico —las muestras textuales que componen el corpus— y otra relativa a las herramientas informáticas utilizadas en su explotación. Por tanto, el corpus debe ser adecuado al proyecto concreto, tanto en su dimensión textual, como en lo que atañe a su dimensión informática.

La adecuación textual del corpus al proyecto concreto implica tener en cuenta aspectos como:

1) que el dominio, tema o tipo de textos que contenga el corpus sea bien definido y bien delimitado;
2) que los textos sean suficientemente representativos y que el conjunto de los mismos resulte en una muestra adecuadamente equilibrada para, con todo ello, fundamentar las conclusiones que se deriven;
3) que la organización y el contenido del corpus favorezcan su explotación;
4) que los textos sean adecuados en tamaño y formato, de modo que no surjan problemas de compatibilidad con las diferentes herramientas informáticas que utilicemos en su tratamiento y explotación (SÁNCHEZ *et al.*, 1995: 41).

Todo ello hace necesario cierto grado de especialización de las técnicas de construcción y explotación de un corpus según el proyecto en cuestión, así como una buena planificación de todo el proceso en su conjunto.

5.2 EL DISEÑO DE UN CORPUS ESPECIALIZADO: LA REPRESENTATIVIDAD Y EL EQUILIBRIO

El diseño del corpus en su dimensión lingüística necesita primero definir qué tipo de corpus se va a crear. Se trata, en definitiva, de ir decidiendo si el contenido del corpus va a ser general o específico, qué tamaño va a alcanzar, si va a contener algún nivel de anotación o, por el contrario, va a ser un corpus no anotado, y todo este tipo de cuestiones.

En la fase de diseño será necesario precisar de forma pormenorizada aquellos parámetros que garanticen el equilibrio y la representatividad del material lingüístico contenido en el corpus. Asimismo, el terminólogo ha de considerar otros aspectos de naturaleza más operativa y de gestión, y que tienen que ver con la selección de los atributos textuales que el compilador considere necesarios registrar y que se relacionan con la documentación de cada uno de los textos.

En el marco de nuestra investigación, se concretó el proyecto y la tipología que iba a caracterizar al CPN. Desde el principio, el proyecto del corpus se identificó, según su contenido (véase Tabla 1), como perteneciente a la categoría de «especializado por la temática». Definimos este tipo de corpus como un repertorio de muestras textuales de un ámbito de conocimiento o actividad, ordenado según unos determinados parámetros y suficientemente extenso de modo que los datos que de él se obtengan sean representativos y susceptibles de generalización en lo referente al uso lingüístico específico del ámbito. Las muestras textuales que van a formar parte del corpus se almacenan como una colección de ficheros en el ordenador, a fin de tener fácil acceso a los datos lingüísticos y de manipularlos para obtener resultados de diferente naturaleza para la descripción y el análisis. Se observará que esta definición es una adaptación de la propuesta por Sánchez *et al.* (1995: 8-9).

Detallamos a continuación los parámetros que caracterizan al CPN y que tienen que ver con su representatividad y equilibrio, su tipología y los atributos textuales para la selección de los textos.

La definición de la representatividad es un punto crucial en la creación de un corpus, pero también resulta ser un aspecto controvertido entre los especialistas, sobre todo dada la ambigüedad inherente en su uso por las connotaciones cualitativas y cuantitativas que se entremezclan. Mientras que para algunos investigadores los córpora de cientos de millones de palabras deben plasmar una diferenciación mínima en los diversos tipos de textos representados en el corpus; para otros, es una condición básica realizar una diferenciación muy extensa de las distintas variedades para poder llevar a cabo algún tipo de generalización sobre la muestra. En nuestra opinión, un corpus es una muestra finita de una población infinita y, consiguientemente, nunca se alcanza el fin de este conjunto de acontecimientos lingüísticos. Por muy extensivo que sea el muestreo, éste se convertirá en una operación simplificada a la luz de la complejidad del fenómeno objeto de estudio. Incluso llevando a cabo selecciones aleatorias en la construcción del corpus, nos parece que en la transición de la muestra a la generalización deben proporcionarse ciertos grados de aproximación, permitiendo así una máximo de flexibilidad y dinamismo en el modelo propuesto.

En vista de estos problemas de naturaleza epistemológica que se plantean en la planificación de un corpus para que éste pueda ser definido, sin lugar a dudas, como representativo de un sublenguaje, se decidió proceder reconociendo los límites inherentes del proyecto mismo e identificando determinados parámetros que acabaran por equilibrar estos límites. Se definieron algunos criterios de identificación para los parámetros de referencia que permitieron la creación de una colección de subcórpora que incluyera las variedades conceptuales y pragmáticas más significativas de los textos escritos del ámbito en las dos lenguas de trabajo. De este modo fue como obtuvimos un modelo de creación dinámico y adaptable que satisfizo nuestras necesidades e hipótesis de trabajo, al tiempo que respetaba los criterios de construcción de un corpus, establecidos en la bibliografía especializada.

La calidad de un proyecto terminológico o terminográfico está directamente relacionada con la calidad de los textos que constituyen el corpus especializado en que se basa (BOWKER, 1996: 42), al igual que ocurre con uno de contenido general. La calidad y representatividad de los textos que van a constituir el corpus especializado resulta un factor de más importancia que la cantidad de textos que han de componerlos (MEYER y MACKINTOSH, 1996: 268). Por *calidad* hacemos alusión a una de las cuatro características apuntadas por Sinclair (1996: 7), esto es, auténtico: «all the material is gathered from the genuine communications of people going about their normal business».

En los córpora generales, los pequeños desequilibrios de calidad y representatividad que puedan presentar parte de sus textos acaban por ser neutralizados, precisamente por la gran cantidad de datos que contienen (100 millones de palabras en el caso del BNC). En contrapartida, difícilmente se consigue construir un corpus especializado de dimensiones muy extensas por varias razones, pero sobre todo por cuestiones relativas a la disponibilidad de los textos, al *copyright* y a la cantidad de tiempo del que se dispone. Respetando las cuestiones de calidad y representatividad de los textos podremos asegurar una mayor fiabilidad en los resultados.

En el diseño de un corpus general se contempla una gran diversidad de áreas del conocimiento, de temas, de tipos de textos con el objeto de construir una muestra representativa de la lengua que está siendo objeto de estudio. Del mismo modo que ocurre en la lengua general, los discursos científicos y técnicos están lejos de ser homogéneos:

Un simple análisis de la comunicación especializada real en situaciones profesionales de distinto signo muestra una multiplicidad importante de registros, en los que, sin abandonar el carácter especializado del conocimiento y su transmisión, se ponen de manifiesto una serie de características que coinciden con las que muestran otras unidades utilizadas en otros tipos de situación comunicativa (CABRÉ, 1999: 118).

Así, tal y como ocurre con los córpora generales (LOB, COBUILD, Brown, LLC), los especializados deben contemplar las diferentes situaciones comunicativas del discurso que representan, teniendo en cuenta, además, que una identificación adecuada de los distintos términos y de sus niveles de especialización depende de forma crucial de la tipificación rigurosa de los textos que se van a seleccionar. Es más, si el objetivo es construir una aplicación terminológica dirigida al traductor los distintos niveles de especialización de los textos del ámbito deben considerarse, dado que cada nivel proporcionará, a su vez, distintos tipos de datos lingüísticos, como son: términos con distintos grados de especialización, definiciones, contextos, sinónimos, etc. Con todo, todo corpus es limitado, en el sentido de que únicamente puede representar algunos subconjuntos de la lengua, ya sea o no especializada, y no el conjunto en su totalidad. Es prácticamente imposible que un corpus contenga todos y cada uno de los tipos de textos que se pueden producir en todas y cada una de las situaciones comunicativas que pueden

darse en una lengua o en una variedad de ésta. Por tanto, para compilar un corpus, es aconsejable elegir de forma explícita los usos lingüísticos que van a ser el centro de atención.

Asimismo, es primordial que el corpus sea representativo con respecto a los tipos de textos que se dan en el ámbito en cuestión. Se hace, asimismo, necesario en la fase de diseño del corpus considerar las diferentes situaciones académicas o profesionales en las que los especialistas del ámbito objeto de estudio producen sus discursos para abarcar, de este modo, los diferentes tipos de textos representativos del campo profesional en cuestión.

Por otra parte, pensamos que un corpus que es equilibrado en términos de tipos textuales podrá satisfacer las diferentes necesidades conceptuales, pragmáticas y lingüísticas que se le presentan al terminógrafo. En el plano conceptual, un corpus que refleje, a través de la concepción y adopción de una tipología textual pragmática (*cf.* VARGAS, 2005), como ha sido nuestro caso, los diferentes grados de especialización del discurso resulta muy útil al terminógrafo que se introduce en un ámbito nuevo que le es ajeno para adquirir un cierto grado de conocimiento. Un corpus creado de este modo permitirá al terminógrafo comenzar a adquirir más fácilmente este conocimiento a partir de textos dirigidos a semiexpertos o legos, muy ricos en definiciones, explicaciones y sinónimos, e ir progresando paulatinamente con textos dirigidos a expertos, con el grado más elevado de especialización, que contienen una densidad importante de términos técnicos, pero son pobres, por lo general, en definiciones, explicaciones y uso de sinónimos o perífrasis. En el plano pragmático, los índices de frecuencia y los contextos de uso de los términos guían al terminólogo sobre lo que es oportuno recoger o no en la aplicación terminológica que esté elaborando, siempre en función también del usuario prototípico de dicha aplicación. En el plano lingüístico, la inclusión de textos con diferentes niveles de especialización proporciona al terminógrafo una imagen más completa de la diversidad terminológica existente en el ámbito objeto de investigación (BOWKER, 1996).

En el contexto de los córpora especializados y con la finalidad de conseguir el referido equilibrio, creemos que resulta conveniente establecer dos tipos de equilibrio: uno conceptual y otro pragmático. El equilibrio conceptual es una noción introducida por Bowker (*ibid.*), quien recomienda recoger en el corpus todos los subcampos en los que se divide una parcela del conocimiento o campo de actividad, así como todos los ámbitos que guarden relación con ésta cuando tiene un carácter multidisciplinar. El árbol de campo o estructuración conceptual del ámbito se convierte, en este sentido y para este fin, en una herramienta de trabajo nada desdeñable para el terminógrafo, dado que éste pretende abarcar del modo más exhaustivo que sea posible la terminología que se usa en un ámbito de especialidad concreto. Creemos que esta configuración conceptual de un campo de actividad ayuda a alcanzar, si no completamente, sí en gran medida, la exhaustividad pretendida.

En el CPN, y una vez establecidos los limites temáticos a través del árbol de campo, trabajamos sobre nuestro corpus intentando establecer una representación macro y microtemática lo más real posible del ámbito delimitado, siendo conscientes, no obstante, de sus posibles imperfecciones. Para ello, y en consonancia con lo que propugnan Meyer y Mackintosh (1996: 270), elegimos un conjunto de textos que nos proporcionaba un alcance equilibrado de todos los aspectos temáticos del dominio, abarcando subdominios (petrología, diseño de canteras, maquinaria, etc.) y ámbitos directamente relacionados (construcción y transporte, por ejemplo) en una proporción equilibrada y ponderada, según su importancia, dentro del ámbito explorado.

No sólo es importante intentar obtener un equilibrio conceptual del dominio que intentamos representar en nuestro corpus, sino que, a nuestro modo de ver, propugnamos que es todavía más relevante dirigir nuestros esfuerzos hacia la obtención de un «equilibrio pragmático» (AHMAD, 1995: 61). Para llevarlo a cabo deberemos, en primer lugar, considerar las diferentes situaciones académicas y/o profesionales de producción y recepción de textos que nacen y circulan en el ámbito especializado objeto de estudio. De este modo, contemplaremos los diferentes tipos de textos representativos del campo profesional en cuestión.

En la construcción del CPN y con el fin de alcanzar, o intentarlo al menos, el equilibrio pragmático, debíamos considerar, en primer lugar, las diferentes situaciones de comunicación de nuestro ámbito y, a continuación, proceder con la recopilación de una proporción equilibrada y ponderada de textos de las situaciones contempladas. En este sentido, la *distribución textual* de nuestro corpus podemos valorarla de equilibrada, puesto que las muestras textuales que lo componen están distribuidas en proporciones semejantes.

En el marco de nuestra investigación, construimos un modelo tipológico textual que se regía por principios pragmáticos (VARGAS, 2005). Dicho modelo se organizó, de acuerdo con una tipología de varios niveles, en dos criterios generales de análisis: (1) situacional (campo, tenor y modo); (2) funcional (funciones comunicativas). Estos criterios intentaban abarcar aspectos que establecían la relación de los participantes de la comunicación especializada y las funciones comunicativas. A partir de la aplicación de estos dos criterios se definió de manera más acotada los tipos textuales o géneros resultantes.

A continuación se presenta la Figura 2, en la que se muestran los porcentajes del contenido de nuestro corpus según las situaciones comunicativas y las funciones resultantes que han sido previstas. Seguidamente, se comenta el gráfico.

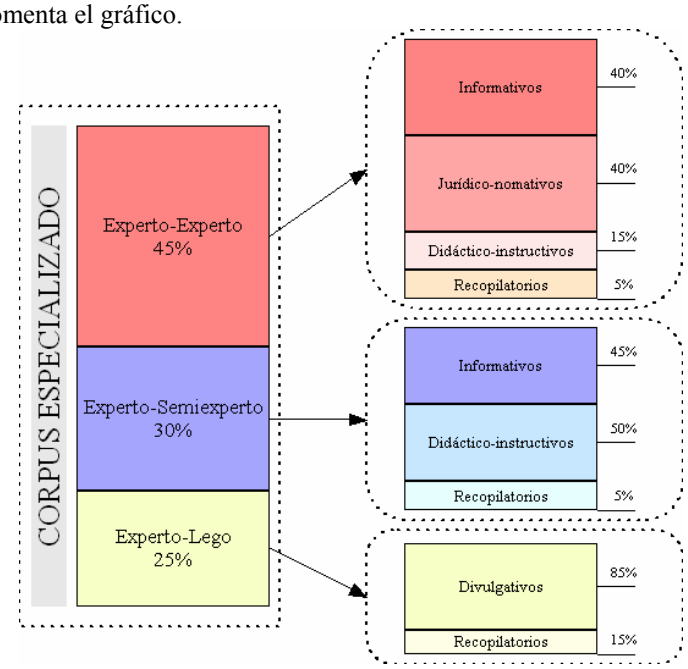

Figura 2: *Porcentajes del contenido del CPN*

En nuestra opinión, el gráfico anterior representa una muestra global del CPN equilibrada en términos pragmáticos. Los porcentajes que se observan creemos que así lo constatan. Como ya hemos comentado anteriormente, y atendiendo a los criterios de características del producto que pretendíamos construir (un diccionario) y de acuerdo con el usuario a quien va destinada la aludida aplicación (el traductor), el equilibrio pragmático de nuestro corpus viene representado, en primer lugar, por la distribución de las relaciones del productor-receptor de los textos o tenor. En segundo lugar, por las funciones dimanantes de dicha relación. En la primera distribución, el tenor (experto→experto) supone un 45% del contenido del CPN. Dicha cantidad creemos que se justifica por la producción de textos mayor que se genera en dicha relación y por la densidad terminológica más elevada que presentan los textos resultantes de esta relación. En el siguiente tenor (experto→semiexperto), el CPN ha contemplado un tercio de su total, quedando una cuarta parte reservada al último tenor de nuestro modelo. Con respecto a la distribución en funciones de cada uno de los tenores, no hemos considerado oportuno el dividir nuestro corpus de forma igualitaria en cada una de ellas. Lo anterior se justifica si atendemos a criterios de ponderación. Dicha ponderación viene determinada, en nuestro caso, por la relevancia de las funciones contempladas. En consecuencia, se ha considerado como equilibrio funcional la distribución representada en el gráfico anterior y que ofrece como característica más importante que los mayores porcentajes se incluyan en aquellas funciones cuya densidad terminológica también es mayor, quedando relegados a representaciones casi testimoniales las funciones recopilatorias.

5.3 TIPOLOGÍA DE CÓRPORA

Un corpus se diseña siempre teniendo en mente un propósito definido y concreto y el tipo de corpus dependerá en gran medida de este objetivo. La tipología de corpus se precisa y se especifica de acuerdo con diferentes parámetros que lo caracterizan, tales como el tamaño, las lenguas, el contenido, etc. La siguiente tabla es una recopilación de las distintas clasificaciones realizadas en la bibliografía especializada sobre corpus. En ella se muestran, de manera sintética, los criterios a partir de los que se puede clasificar un corpus dado, los tipos resultantes según el anterior parámetro y una breve descripción de cada tipo.

Criterios	Tipos	Descripción
Canal de producción	escrito	textos escritos
	oral	fragmentos de habla transcritos
	mixto	textos escritos y hablados
Cantidad de palabras	grande	10 millones o más
	mediano-grande	un millón a 10 millones

Criterios	Tipos		Descripción	
	mediano		250.000 a un millón	
	pequeño-mediano		80.000 a 250.000	
	pequeño		menos de 80.000	
Distribución textual	equilibrado		textos distribuidos en proporciones semejantes	
	piramidal		textos distribuidos en *n* número de niveles	
Contenido	general		Su propósito principal es reflejar el comportamiento de la lengua general. Abarca una amplia gama de tipos de textos diferentes, es equilibrado y muy extenso para representar todas las variedades de la lengua y el vocabulario característico	
	especial		diseñados con el objetivo de representar un uso lingüístico que se aleja del lenguaje corriente de una comunidad de hablantes específica	
		genérico	sólo contiene un tipo de género concreto	
		canónico	formado por todas las obras escritas por un determinado autor	
		ámbito o tema	textos con contenido sobre un ámbito o tema concretos	
Tamaño de muestras	de textos completos		sus muestras corresponden a textos íntegros	
	de fragmentos		las muestras son fragmentos con un tamaño específico	
Codificación/anotación	simple		muestras en formato ASCII	
	etiquetado		marcado estructuralmente	con etiquetas descriptivas de elementos constitutivos
			anotado lingüísticamente	con etiquetas analíticas de aspectos lingüísticos
Documentación	documentado		cada documento lleva asociado un archivo DTD	
	no documentado		los textos no disponen de ningún apartado o archivo donde estén descritos sus elementos o su filiación	
Período de tiempo	sincrónico		textos de un período específico de tiempo	
	histórico o diacrónico		textos de uno o varios períodos de un tiempo pasado	
	contemporáneo		textos actuales	
Capacidad de actualización	cerrado		colección finita de textos	
	abierto o monitor		constantemente alimentado con textos nuevos	
Finalidad	fines generales		fuente de información textual y de referencia para fines diversos	
	fines específicos		pretenden dar respuesta a un propósito concreto	
Selección de las muestras	por muestreo de conveniencia		las muestras textuales se seleccionan por ser fáciles de obtener	
	por muestreo proporcional		de un grupo de personas se hace acopio de todo el material escrito y hablado que produzcan y reciban durante un determinado espacio de tiempo	
	por muestreo estratificado		se divide la población en estratos; en cada estrato se recoge un número de muestras proporcional al peso relativo del estrato en el total de la población	
Idiomas	monolingüe		textos en un idioma	
	bilingüe		en dos lenguas	comparable
				paralelo → alineado

Criterios	Tipos	Descripción	
	multilingüe	en más de dos lenguas	comparable
			paralelo → alineado

Tabla 1: *Tipos de córpora*

En la tabla anterior creemos que quedan suficientemente explícitos los criterios que condicionan el tipo de corpus que se crea. La relación de dichos criterios, así como sus correspondientes variables se convierten, al mismo tiempo, en una suerte de formulario o protocolo que da cuenta del conjunto de decisiones generales y primeras que se necesitan tomar a la hora de diseñar un corpus (véase Tabla 2). Se trata, en definitiva, de ir decidiendo si el contenido del corpus va a ser general o específico, qué tamaño va a alcanzar, si va a contener algún nivel de anotación o, por el contrario, va a ser un corpus no anotado, etc. La compleción del protocolo nos ofrecerá de forma definida y explícita el tipo de corpus que pretendemos confeccionar.

A continuación especificamos cada uno de los criterios referidos en la anterior tabla con el fin de describir el corpus creado dentro de nuestro grupo de investigación, a excepción de la distribución textual y del contenido, parámetros a los que ya hemos hecho referencia anteriormente (§ 5.2).

5.3.1 EL MODO O CANAL DE PRODUCCIÓN

La decisión que se ha de tomar respecto al modo o canal de producción es si el corpus va a contener textos escritos, textos orales transcritos o bien una combinación de estos dos modos. Decidirse por si se quiere compilar un corpus escrito, oral o mixto dependerá del tipo de estudio que se pretenda realizar. No obstante, es necesario destacar que es más sencillo y rápido construir un corpus escrito, dado que el proceso de trascripción de un texto oral es muy laborioso y lleva un tiempo considerable.

En nuestro proyecto, no contemplamos la posibilidad de que el CPN contuviese muestras orales, pues desde el principio se dio prioridad a los textos escritos y se centró exclusivamente en ellos. Lo anterior no quiere decir que no consideremos interesante estudiar la utilidad de discursos orales especializados y valorar este tipo de discurso en una ampliación futura del proyecto.

5.3.2 EL TAMAÑO

Podemos afirmar que no existe un criterio único y consensuado sobre el tamaño ideal que debe alcanzar el corpus concreto que nos propongamos compilar. Así, es el compilador quien decide el tamaño de su corpus atendiendo a variables como las necesidades y el propósito del proyecto en cuestión, la disponibilidad de los textos y la cantidad de tiempo del que se dispone. Es importante tener en cuenta, asimismo, que la falta de recursos, en ocasiones, condiciona el diseño y el tamaño deseado en un principio.

Un análisis de los córpora disponibles en la actualidad refleja que no es posible hacer referencia explícita a un tamaño estándar. El desarrollo acuciante y ampliamente difundido que ha caracterizado, particularmente en años recientes, tanto la disponibilidad a bajo coste del *hardware* así como la producción de *software* más eficaz y fácil de utilizar ha transformado de manera radical los criterios de creación más recientes de córpora, si los comparamos con aquéllos de la primera y segunda generación[1]. Mientras que los criterios sobre los que se basaron la primera generación de córpora (Brown, LLC, LOB, etc.) estuvieron influidos de manera decisiva por la potencialidad de las tecnologías de la información del momento, la tecnología actual ya no establece limites a las decisiones del investigador, quien puede extender y ampliar el corpus para incluir las variedades que considere como relevantes para su análisis y, dentro de éstas, realizar las selecciones apropiadas de otras subvariedades de textos representativos.

Los avances en las tecnologías de la información que se han sucedido durante los últimos años, la velocidad actual de procesamiento del material y los bajos costes de las unidades de almacenamiento masivo se traducen en que es posible crear córpora de cientos de millones de palabras, como el BNC y el Bank of English. Parece que, especialmente en los que se refiere a la lengua escrita, el estándar de un millón de palabras ha dejado paso al de cientos de millones de palabras. Sin embargo, cualquier generalización en este sentido es discutible, así como lo es cualquier definición de un límite obligatorio. El corpus Brown, compuesto de un millón de palabras, con 500 muestras textuales escritas y 2.000 palabras en cada una, que representa, además, en una medida similar los principales tipos textuales, es todavía considerado por muchos investigadores como un modelo válido. Uno de los córpora más recientes en inglés, el Longman Spoken and Written English Corpus —LSWE Corpus— creado por académicos como Biber, Johnson, Leech, Conrad y Finegan, consiste en 40 millones de palabras y contiene 37.244 textos. Según los autores, estos textos varían en tamaño según el registro.

Es generalizada la opinión de que los córpora que sirven para estudios sobre el lenguaje general han de ser más extensos que aquéllos necesarios para el trabajo con lenguajes especializados (BOWKER y PEARSON, 2002: 48; ENGWALL, 1994: 50) o con propósitos específicos (*cf.* GHADESSY *et al.*

2001). Con respecto al tamaño que debe alcanzar un corpus terminográfico, éste dependerá del lenguaje concreto que esté siendo objeto de estudio:

For an LGP dictionary, a corpus to the tune of 50 million text words is usually considered necessary. […]. No corresponding figures can be given for LSP corpora, where size depends on the LSP in question. A corpus compiled for the lexicographical description of American gene-technology usage consisted of approximately 500,000 text words, which turned out to be sufficient for the purpose. (BERGENHOLTZ y TARP, 1995: 95)

Así, si el ámbito que queremos estudiar es muy delimitado y circunscrito, como era el caso del ejemplo en la cita anterior (ingeniería genética, inglés americano), bastará con contar con un corpus de tamaño pequeño a mediano (de 80.000 a un millón de palabras), pero bien diseñado, pues para los córpora terminográficos la cantidad es un aspecto menor en importancia que la calidad.

En algunos casos, según el propósito, incluso convendría no poner límites, como recomienda Pearson (1998: 59), e incluir todos los textos que cumplan los requisitos explícitos del proyecto. Ahmad y Rogers (2001: 736) exponen varios argumentos que justifican un tamaño de corpus menor para el trabajo terminológico:

- El estatus de las diferentes unidades contenidas en el lenguaje especializado (incluyendo los términos) proviene de fuentes más o menos fidedignas (expertos en el ámbito), mientras que en la lengua general la noción de autoridad es más problemática;
- La variación léxica y gramatical en los textos especializados tiende a ser menor, por lo que en estos sublenguajes restringidos semánticamente los patrones lingüísticos pueden resultar más evidentes a partir de muestras más pequeñas;
- La noción de ámbito o dominio, si bien no está bien definida, está mejor delimitada y restringida que el concepto un tanto vago de, por ejemplo, «inglés británico contemporáneo»;
- Los textos especializados tienden a ser más densos léxicamente que los textos generales[2]. El foco de atención lo constituyen las palabras de clase abierta (verbos, adjetivos, sustantivos y adverbios) y frases que son potencialmente términos;
- Los terminólogos no están interesados por todas las palabras contenidas en los textos especializados, sino solo por aquéllas específicas del ámbito objeto de estudio. Se toman en consideración otras palabras en tanto en cuanto ilustren patrones fraseológicos;
- El propósito específico de la terminología que está siendo compilada puede limitar la selección del tipo de texto o género.

En un principio, para el CPN se decidió proceder con la planificación de un corpus cuyo tamaño no quedaba predeterminado, sino que se establecería en función de aspectos como la disponibilidad de los textos y el tiempo que nos llevase su recuperación y procesamiento. Finalmente, construimos un corpus que contiene, en inglés, 196 muestras textuales y alcanza un tamaño de 738.147 palabras. El corpus español se compone de 148 textos y su tamaño es de 612.799 palabras. Por tanto, según la clasificación del tamaño que realiza Berber (2000: 346) el corpus bilingüe alcanza un tamaño mediano-grande (de un millón a diez millones), concretamente, 1.350.946 palabras, mientras que separando el corpus por lenguas obtendríamos para cada uno un tamaño mediano (de 250.000 palabras a un millón).

El volumen de nuestro corpus bilingüe puede parecer modesto en comparación con el número de palabras que se utiliza en la actualidad en la lingüística de corpus (cientos de millones de palabras). Sin embargo, hay que tener en cuenta que los textos especializados son más densos en términos léxicos que los textos de la lengua general (AHMAD y ROGERS, 2001: 726). Asimismo, los textos con un elevado grado de tecnicidad presentan lo que se conoce como *densidad terminológica* o, dicho de otro modo, un número elevado de unidades que trasmiten conocimiento especializado. Pensamos, por tanto, que aunque no se maneje en el corpus especializado un número tan elevado de palabras como ocurre con un corpus de la lengua general se pueden obtener de él resultados satisfactorios, si conseguimos, por una parte, el equilibrio y la representatividad ya apuntados y, por la otra, si contamos con una muestra considerable de textos con un elevado grado de tecnicidad.

5.3.3 EL TAMAÑO DE LAS MUESTRAS

En el contexto de la lingüística de corpus, uno de los criterios básicos aceptados por todos los proyectos es que los textos seleccionados deben ser auténticos y de uso común en la interacción social. Sin embargo, no hay consenso sobre si incluir textos en su integridad o fragmentos que se definen como

representativos. En contrapartida, parece haber consenso en la bibliografía que aborda el tema de los córpora en terminología (AHMAD, 1995: 61; BOWKER, 1996: 43; MEYER y MACKINTOSH, 1996: 268) y se recomienda que las muestras textuales que los compongan sean textos completos. Así, el protocolo establecido por algunos córpora generales y que recomienda recoger fragmentos textuales y limitar la muestra a un número determinado de palabras —2.000 en el caso del Brown y el LOB; y 40.000 en el caso del LLC— no debe aplicarse a los córpora especializados. La integridad de un texto es necesaria para garantizar su valor conceptual (MEYER y MACKINTOSH, 1996: 268). Los terminógrafos no pueden realizar una selección arbitraria de los fragmentos que se han de incluir, pues, además de que los términos pueden aparecer en cualquier parte del texto, se corre el riesgo de omitir algún tipo de descripción conceptual.

En el CPN, así como en todos los demás córpora que han sido construidos por nuestro grupo de investigación, las muestras textuales son textos íntegros por las razones que acabamos de apuntar.

5.3.4 LA ANOTACIÓN

Una vez que está en formato electrónico, el corpus puede adquirir dos formas: simple o anotado. El corpus simple o no anotado es aquél que está sin etiquetar, es decir, no contiene información lingüística, sino las palabras "en bruto". Un corpus anotado, por su parte, además de contar con etiquetas descriptivas de los elementos constitutivos de cada texto (marcado estructural), incorpora etiquetas de diferente naturaleza, que van desde símbolos sencillos para marcar rasgos fonéticos o prosódicos hasta etiquetas morfológicas complejas que vinculan cada palabra con su categoría gramatical, marcas sintácticas e incluso anotaciones semánticas, pragmáticas o discursivas. La anotación de un corpus es un proceso que requiere una gran inversión de tiempo, esfuerzos e infraestructura y no es, en absoluto, un proceso sencillo, por mucho que en la bibliografía especializada se pase por alto la importancia que merece esta cuestión (SÁNCHEZ *et al.*, 1995: 145).

Decidirse por un corpus simple o anotado depende, entre otros aspectos, de los objetivos del corpus, de sus usuarios, de una formación tecnológica avanzada de los terminólogos, y del tiempo del que se dispone para la ejecución completa del proyecto terminológico. Otra cuestión relevante para tomar esta decisión es tener presente que cada nuevo proyecto terminológico implica crear un nuevo corpus, a diferencia de lo que ocurre en un proyecto lexicográfico, en donde el corpus construido puede ser reutilizado y alimentado con nuevos datos. Si, además, se trata de un proyecto en donde se trabaja con varias lenguas la practicidad de un corpus etiquetado se ve reducida. Para poder emplear un etiquetador morfológico basado en reglas (p.ej., Brill's tagger) tendremos que buscar, probar y elegir uno para cada lengua de trabajo. Si queremos hacer uso de un etiquetador estadístico (p.ej., QTAG) éste tendrá que ser adaptado para que trabaje con las distintas lenguas del proyecto.

Asimismo, hay que tener en cuenta que la anotación de un corpus no es un proceso completamente automático, en el sentido de que necesita siempre de la intervención humana para ser etiquetado al 100%. Por otra parte, la automatización de este proceso es aún menor cuando lo que queremos anotar es un corpus que contiene muestras textuales de un ámbito especializado, muy ricas en términos y neologismos no contenidos en los diccionarios o lexicones sobre los que operan los etiquetadores y lematizadores. Estas aplicaciones necesitan ser entrenadas para trabajar con lenguajes especializados de un determinado ámbito, pues el diccionario no reconocerá algunas formas técnicas y responderá con la etiqueta <desconocido>, a modo de ejemplo.

El corpus con el que se contó para nuestra investigación terminológica era no anotado. Esta decisión se justificó, en nuestro caso, a partir de los dos parámetros esenciales ya citados anteriormente. Nos referimos a los medios de los que disponíamos para llevar a cabo nuestro proyecto y a los objetivos que se pretendían alcanzar.

El corpus no anotado nos ha servido como material para encontrar, por medio de los programas de concordancias, las frecuencias de aparición de palabras simples y compuestas, así como para estudiar el comportamiento de estas palabras o combinaciones de palabras a través de los análisis de concordancias. También hemos podido extraer otro tipo de información lingüística (definiciones y contextos) y estudiar las relaciones conceptuales entre los términos (sinonimia, hiponimia, meronimia, etc.).

5.3.5 DOCUMENTADO O NO DOCUMENTADO

Un corpus documentado es aquél en el que cada documento que compone el corpus lleva asociado un archivo DTD (Document Type Definition) o cabecera de SGML. Un corpus no documentado es el que los textos que lo componen no disponen de ningún apartado o archivo relacionado donde estén descritos sus elementos o su filiación.

Como indicamos en el apartado anterior, nuestro corpus no lleva ningún tipo de anotación de ningún nivel, ni siquiera estructural. Así, los textos que lo constituyen no contienen archivo alguno o

apartado donde se describan sus elementos constituyentes. En este sentido, por tanto, se trata de un corpus no documentado, si bien todos los atributos textuales de cada muestra sí que fueron registrados, como explicamos a continuación.

En la selección y clasificación de los textos que van a formar parte de un corpus que va a ser analizado lingüísticamente es necesario considerar, según la bibliografía especializada (ATKINS *et al.*, 1992; PEARSON, 1998; SINCLAIR, 1996), dos tipos de criterios: los internos o puramente lingüísticos y los externos o extralingüísticos (dónde/cuándo se ha producido el texto, por o para quien, sobre qué trata, etc.).

Tanto los aspectos internos como los externos afectan y caracterizan a los textos que nos proponemos recoger. Se trata, en definitiva, de los atributos básicos y elementales que son recomendables registrar para una organización y gestión óptima del corpus. Para organizar, registrar y poder recuperar los textos en razón de distintos criterios diseñamos un Sistema Gestor de Bases de Datos Relacional (SGBDR) en *Access* (versión *XP*) del paquete *Microsoft Office* y que denominamos *GesCorpus.* Se trata de una aplicación de distribución gratuita[3] en donde se recogen los atributos textuales de diseño seleccionados para nuestro corpus.

5.3.6 PERÍODO DE TIEMPO

Este criterio resulta del proyecto de investigación concreto y del tipo de corpus que se pretende construir (sincrónico, diacrónico o histórico y contemporáneo). Dado que se trataba de un corpus especializado con fines terminográficos no se nos planteaba el problema de tener que elegir entre compilar bien un corpus diacrónico, o bien uno contemporáneo. La bibliografía sobre terminología recomienda seleccionar documentación que haya sido recientemente publicada o puesta al día. Tanto en el plano conceptual, como en el lingüístico, los terminógrafos se interesan por textos actuales, pues en ellos es donde se puede encontrar los términos más novedosos, así como la información conceptual actualizada (MEYER y MACKINTOSH, 1996: 271). Desde el principio, por tanto, se vio claro que debíamos seleccionar textos contemporáneos, con una antigüedad preferiblemente no superior a diez años, que facilitaran la descripción generalizada del uso terminológico actual de nuestro ámbito de especialidad en dos idiomas. La actualidad de los textos es un parámetro también determinante para alcanzar el equilibrio pragmático de un corpus especializado, pues los neologismos difícilmente los encontraremos en textos que no sean actuales.

5.3.7 LA CAPACIDAD DE ACTUALIZACIÓN

En este punto conviene decidir si vamos a construir un corpus cerrado, formado por una colección finita de textos o un corpus abierto o monitor. Este último tipo se refiere a aquél que se alimenta de forma constante con nuevos datos a fin de dar cuenta de las transformaciones por las que pasa una lengua hasta configurarse en su estado actual y reciente.

Meyer y Mackintosh (1996: 269) conciben un corpus terminográfico como abierto: «because of the speed with which terminology changes, the terminographical corpus should be open». Sin embargo, si el fin es construir un diccionario terminológico, como ha sido nuestro caso, creemos que es necesario cerrar el corpus en algún momento para proceder con la elaboración de los listados a partir de los que extraemos la terminología y otros datos lingüísticos. Y de este tipo es el CPN, es decir, cerrado. Ello no implica que una vez emitidas las listas y elaborado el diccionario el corpus siga siendo alimentado con nuevos textos, aspecto que beneficia, sin duda alguna, cualquier fin de investigación posterior que se pueda realizar sobre el corpus construido. Para completar un corpus especializado se necesita emplear muchos esfuerzos y recursos, así como una gran inversión del total del tiempo del que se dispone. Por lo tanto, es lógico que una vez construido se desee, sobre todo para futuras investigaciones, que no quede obsoleto por no haber sido alimentado con nuevos textos.

5.3.8 LA FINALIDAD

Un corpus se diseña siempre teniendo en mente un propósito definido y concreto. Sin embargo, podemos establecer diferentes propósitos y agruparlos en más general, o bien más concreto y circunscrito. Concebimos un propósito general del corpus cuando éste pretende ser de utilidad para investigaciones de distinta naturaleza y para una población de usuarios más extensa y también diversa. Por el contrario, la finalidad del corpus es más concreta y específica que la anterior cuando el corpus diseñado únicamente pretende servir a un fin restringido al grupo de investigación que lo crea, si bien es cierto que posteriormente puede resultar de interés para otros investigadores por su diseño y composición.

El fin de nuestro corpus es específico, esto es, pretendía dar respuesta a las necesidades de nuestro grupo de investigación y, más concretamente, se diseñó, como ya hemos venido diciendo, para elaborar un diccionario especializado por la temática.

5.3.9 EL MÉTODO DE MUESTREO

Una de las cuestiones principales que es necesario considerar cuando se compila un corpus es el método de muestreo que se va a seguir para hacer acopio de una muestra representativa de la población objeto de estudio. En este contexto, y a tenor de la bibliografía consultada, podemos elegir tres formas o, por supuesto, nada nos impide llevar a cabo una combinación de éstas. Nos referimos a que el método de muestreo puede ser por conveniencia, proporcional y/o estratificado. En el muestreo de conveniencia, las muestras textuales que compondrán el corpus se seleccionan por ser fáciles de obtener (a través de Internet, CD-ROM, los especialistas, etc.). Resulta ser un método al que habitualmente se recurre, pues en ocasiones es difícil, por no decir imposible, obtener el material deseado. En el corpus donde el muestreo es proporcional, se selecciona un grupo de personas y se hace acopio de todo el material escrito y hablado que produzcan y reciban durante un determinado espacio de tiempo. En el corpus por muestreo estratificado se divide la población —los textos— en estratos, esto es, en categorías de textos, proporcionales al peso relativo del estrato en el total de la población.

En el CPN el método de muestreo fue de dos tipos: por conveniencia y estratificado. Con respecto al primero, parte de nuestro corpus contiene textos que fueron recuperados a través de Internet y de los especialistas colaboradores en el proyecto. Los textos así seleccionados, cumplían con los criterios de calidad y de representatividad que establecimos. Dicho de otro modo, se trataba de textos auténticos, elaborados por especialistas dentro de su comunidad epistemológica, y su contenido era representativo del ámbito y adecuado a los diferentes subdominios articulados a través del árbol de campo. En cualquier caso, es cierto que resulta mucho más sencillo seleccionar los textos que van a formar parte de un corpus especializado que en el caso de córpora generales, como bien afirman Atkins *et al* (1992: 5): «The more highly specialized the language to be sampled in the corpus, the fewer will be the problems in defining the texts to be sampled». Es claro, no obstante, que un corpus no puede basarse exclusivamente en un muestreo por conveniencia, recopilando todo el material electrónico que se pueda. Sin embargo, también es cierto que un proyecto con un presupuesto y tiempo limitados, en muchas ocasiones e ineludiblemente, tendrá que optar, atendiendo a cuestiones pragmáticas, por este tipo de material fácil de obtener.

En nuestro proyecto, junto con el de conveniencia que acabamos de exponer, adoptamos el procedimiento de muestreo estratificado, en el que los tipos de textos que se van a compilar, así como sus proporciones, son definidos de antemano. Así, se intentó representar el material publicado de una gama suficientemente extensa de temas, de tipos textuales y de niveles de especialización de los textos. Por medio de los artículos especializados, informes de proyecto, artículos sectoriales, manuales, instrucciones de uso, normas de ensayo, monográficos, anuarios, folletos publicitarios, catálogos, etc., nuestro objetivo consistía en asegurarnos de que las formas más representativas del lenguaje especializado de nuestro ámbito en modo escrito estuvieran presentes en nuestro corpus, en mayor o menor medida, según su coeficiente de ponderación.

5.10 LAS LENGUAS

Un corpus, según las lenguas de trabajo que contemple, puede ser monolingüe, bilingüe o multilingüe. Dentro de estos dos últimos tipos, un corpus puede ser comparable o paralelo. Se entiende que los córpora son comparables cuando los textos que los componen presentan determinados rasgos que comparten, tales como el tema, el tipo de texto, el periodo de tiempo en que se redactaron los textos, la función comunicativa, el grado de especialización, etc. Se dice que un corpus es paralelo cuando contiene textos redactados en una lengua (la original) junto con las traducciones de éstos a otras lenguas.

El CPN es un corpus bilingüe (inglés-español), compuesto en un 55% de textos en inglés y el resto (45%) en español (Figura 3).

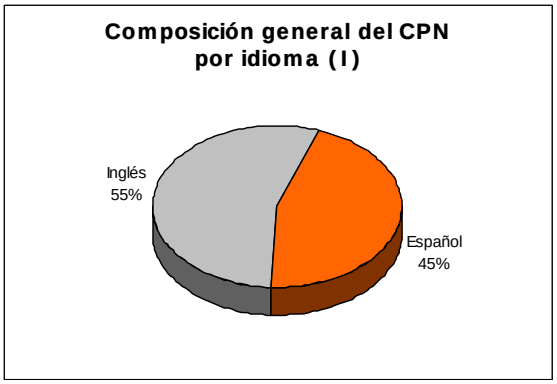

Figura 3: *Composición del CPN por idiomas (inglés-español)*

Es también un corpus compuesto de subconjuntos textuales que son comparables. Es lógico pensar que no todos los textos en su globalidad cumplen con los criterios de comparabilidad necesarios; esto es, todos los textos no abordan un único tema, ni pertenecen todos al mismo género, ni todos tienen la misma función comunicativa ni el mismo grado de especialización. No afirmamos, por tanto, que se trata de un corpus comparable en términos homogéneos. Sin embargo, tomado desde una perspectiva heterogénea, y concebido como un conjunto de subconjuntos más pequeños sí que podemos afirmar que cada uno de estos subconjuntos cumple con alguno de los criterios que se pueden establecer para determinar que un corpus bilingüe es, de algún modo, comparable. Así, podemos obtener pequeñas muestras que son comparables en razón de los criterios que apliquemos para construir estos subconjuntos. Éste sería el caso de la creación de subcórpora comparables, por ejemplo, por la temática. Es claro que a mayor número de características compartidas por los subconjuntos mayor grado de comparabilidad se dará entre ellos. Para poder llevar a cabo cualquier segmentación del corpus (p.ej., por tipos textuales, por grado de especialización, por tema, etc.) se hará necesario recoger todos los atributos textuales de cada una de las muestras que van a componer el corpus en una base de datos. Como ya hemos comentado, dichos atributos se registraban en *GesCorpus*.

Los córpora paralelos, por su parte, presentan muchas ventajas en distintos ámbitos lingüísticos, como la terminografía, la didáctica de lenguas extranjeras, la traducción y la lingüística computacional. En terminografía, su utilidad principal se manifiesta en la extracción de terminología y fraseología especializada gracias a los programas de concordancias bilingües (como, por ejemplo, *ParaConc* y *MultiConcord*). Los profesores y estudiantes de lenguas extranjeras, así como los traductores, pueden utilizar un corpus paralelo a modo de diccionario contextualizado, o bien para observar de forma contrastiva la combinatoria de determinadas palabras, para encontrar equivalentes, etc. Los lingüistas computacionales utilizan estos recursos para probar y desarrollar alineadores, programas de traducción automática, entre otros.

Sea cual sea su uso, y a fin de explotar su potencialidad al máximo, este tipo de córpora necesita ser etiquetado, bien vinculando las frases del TO con las frases del TM, bien los párrafos o bien las palabras, dependiendo del programa que se utilice para ello. Este proceso se denomina *alineación automática de textos paralelos*. Una vez alineados, la principal ventaja de este tipo de córpora es que se pueden explotar para obtener concordancias bilingües. En el marco de nuestro proyecto, la alineación de los córpora paralelos se llevó a cabo empleando un paquete que contiene dos aplicaciones informáticas: *Minimark* y *MultiConcord* (*cf.* GÓMEZ y VARGAS, 2003).

Una pequeña parte de nuestro corpus bilingüe está compuesto por textos paralelos (Figura 4). Del 55% de los textos en inglés que lo componen, el 10% corresponde a textos en este idioma que cuentan con un paralelo al español (concretamente, se trata de 46 textos en inglés que ascienden a un total de palabras de 131.643). Y del 45% de los textos en español, el 10% cuenta con un paralelo al inglés (46 textos y un total de palabras de 135.960)[4].

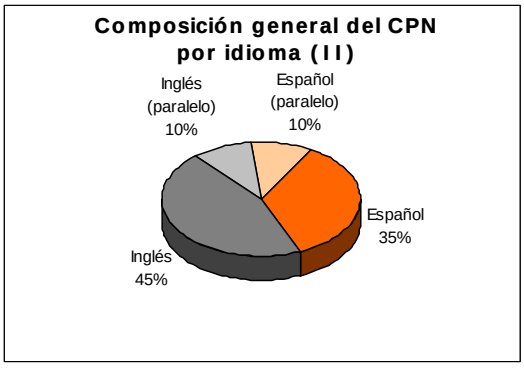

Figura 4: *Composición del CPN por idioma y textos en paralelo*

En la siguiente tabla (Tabla 2) proporcionamos cumplimentado el formulario que elaboramos para ir definiendo de forma pormenorizada las características del CPN. Su inclusión en este trabajo responde a dos objetivos: *a)* ofrecer una síntesis de las características que hemos ido desplegando en los apartados anteriores; y *b)* servir de guia para otros proyectos de investigación terminológica semejantes al nuestro.

Criterios	Tipos	Descripción
Canal de producción	☒ escrito	*textos escritos*

Criterios	Tipos	Descripción
	☐ oral	*fragmentos de habla transcritos*
	☐ mixto	*textos escritos y hablados*
Cantidad de palabras	☐ grande	*10 millones o más*
	☒ mediano-grande	*un millón a 10 millones*
	☐ mediano	*250.000 a un millón*
	☐ pequeño-mediano	*80.000 a 250.000*
	☐ pequeño	*menos de 80.000*
Distribución textual	☒ equilibrado	*textos distribuidos en proporciones semejantes*
	☐ piramidal	*textos distribuidos en n número de niveles*
Contenido	☐ general	*Su propósito principal es reflejar el comportamiento de la lengua general. Abarca una amplia gama de tipos de textos diferentes, es equilibrado y muy extenso para representar todas las variedades de la lengua y el vocabulario característico*
	☒ especial	*diseñados con el objetivo de representar un uso lingüístico que se aleja del lenguaje corriente de una comunidad de hablantes específica*
	☐ genérico	*sólo contiene un tipo de género concreto*
	☐ canónico	*formado por todas las obras escritas por un determinado autor*
	☒ ámbito o tema	*textos con contenido sobre un ámbito o tema concretos*
	☐ otros	*(especificar)*
Tamaño de muestras	☒ de textos completos	*sus muestras corresponden a textos íntegros*
	☐ de fragmentos	*las muestras son fragmentos con un tamaño específico*
Codificación/anotación	☒ simple	*muestras en formato ASCII*
	☐ etiquetado — ☐ marcado estructuralmente	*con etiquetas descriptivas de elementos constitutivos*
	☐ etiquetado — ☐ anotado lingüísticamente	*con etiquetas analíticas de aspectos lingüísticos*
Documentación	☐ documentado	*cada documento lleva asociado un archivo DTD*
	☒ no documentado	*los textos no disponen de ningún apartado o archivo donde estén descritos sus elementos o su filiación*
Período de tiempo	☐ sincrónico	*textos de un período específico de tiempo*
	☐ histórico o diacrónico	*textos de uno o varios períodos de un tiempo pasado*
	☒ contemporáneo	*textos actuales*
Capacidad de actualización	☒ cerrado	*colección finita de textos*
	☐ abierto o monitor	*constantemente alimentado con textos nuevos*
Finalidad	☐ fines generales	*fuente de información textual y de referencia para fines diversos*
	☒ fines específicos	*pretenden dar respuesta a un propósito concreto*
Selección de las muestras	☒ conveniencia	*las muestras textuales se seleccionan por ser fáciles de obtener*
	☐ proporcional	*de un grupo de personas se hace acopio de todo el material escrito y hablado que produzcan y reciban durante un determinado espacio de tiempo*

Criterios	Tipos	Descripción	
	☒ estratificado	*se divide la población en estratos; en cada estrato se recoge un número de muestras proporcional al peso relativo del estrato en el total de la población*	
Idiomas	☐ monolingüe	*textos en un idioma*	
	☒ bilingüe	*en dos lenguas*	☒ comparable
			☒ paralelo → alineado
	☐ multilingüe	*en más de dos lenguas*	☐ comparable
			☐ paralelo → alineado

Tabla 2: *Formulario cumplimentado para determinar la tipología del CPN*

6. CONCLUSIONES

Este artículo ha sido consagrado a cuestiones sobre las especificaciones y el diseño para compilar córpora creados con un fin terminográfico y específicos de ámbitos profesionales y académicos. Para ello, hemos tratado los aspectos directamente relacionados con el diseño de córpora para fines específicos, que hemos ejemplificado recurriendo al corpus creado en el marco de nuestro grupo de investigación. Su meta ha sido servir de modelo o guía para el diseño y la construcción de un corpus de un ámbito especializado. Hemos comenzado con una revisión del marco de referencia en el que se encuadra el presente trabajo. A continuación, hemos definido el CPN, que comenzó a ser compilado con la idea de convertirse en una muestra representativa y equilibrada conceptual y pragmáticamente del inglés y el español profesional y académico en modo escrito del sector industrial de la piedra natural. Se trata, en definitiva, de una colección auténtica de textos del ámbito explorado en formato electrónico y sus fines son de investigación terminológica.

En nuestra opinión, la compilación de un corpus merece un periodo de reflexión en el que se vaya configurando y matizando su diseño. Así, el corpus debe ser considerado por el terminógrafo en distintas etapas, pues es nuestra materia prima y de este recurso depende, en gran medida, los resultados y la calidad de la aplicación terminológica.

NOTAS:

1. Los proyectos de corpus de primera generación son: el Survey of English Usage (SEU), el Brown University Corpus of American English (Brown Corpus), el London-Lund Corpus (LLC), y el Lancaster-Oslo/Bergen (LOB). Se asemejan, principalmente, en tres aspectos: *a)* su tamaño no supera prácticamente el millón de palabras; *b)* las iniciativas para su creación y desarrollo son de naturaleza académica; y *c)* sus fines originales son de investigación. La segunda generación de córpora electrónicos se sitúa en los años ochenta y, según Leech (1991), está representada por dos proyectos: The Birmingham Collection of English Text y el Longman/Lancaster English Language Corpus. Esta generación de córpora se caracteriza por superar el millón de palabras y por estar involucrados agentes económicos privados en su compilación, dado que estos proyectos tienen, además de un propósito académico y de investigación, un fin comercial.

2. Ahmad y Rogers hacen referencia a un estudio elaborado por Halliday (1993). Este último autor observa que la densidad léxica, medida como la proporción de palabras de clase abierta que aparece en un texto, es mucho menor en textos generales que en textos especializados: en el primer tipo, la densidad léxica se establece en un porcentaje que oscila entre el 20% y el 30%; mientras que en los textos especializados está densidad asciende a un porcentaje que se sitúa entre el 50% y el 60% (HALLIDAY, 1993: 76).

3. *GesCorpus* es una base de datos adaptada a los propósitos específicos de nuestro proyecto concreto, esto es, compilar un corpus especializado bilingüe con fines terminográficos. Su diseño, por tanto, se configura para responder a nuestras necesidades y a los atributos textuales definidos. Ello no implica que no pueda ser adaptada a otros proyectos de corpus; más bien lo contrario, pues se trata de un sistema abierto y flexible que puede ser útil si se realizan las modificaciones pertinentes en cada caso. Se puede obtener una copia solicitándola a Chelo.Vargas@ua.es.

4. No nos es posible precisar con rigor la direccionalidad de cada uno de estos 46 textos, es decir, cuáles son originales en inglés o en español y cuáles son traducciones. En ocasiones, sí que podíamos conocer este dato y registrarlo en el SGBD donde se recogían todos los atributos textuales (*GesCorpus*); en otras, sin embargo, carecíamos de la información precisa al respecto.

ABSTRACT:
THE OBJECTIVE OF THIS ARTICLE IS TO DESCRIBE AND ILLUSTRATE THE SPECIFICATIONS AND DESIGN OF THE NATURAL STONE CORPUS (NSC) TO SERVE AS A MODEL FOR TERMINOLOGISTS AIMING TO BUILD TRANSLATOR-ORIENTED TERMINOLOGY RESOURCES. FIRST OF ALL, WE WILL FOCUS ON A SHORT REVIEW OF CONCEPTS AND CRITERIA USEFUL TO ESTABLISH THE THEORETICAL FRAMEWORK. SECONDLY, AFTER HAVING INTRODUCED THE NSC RESEARCH FRAMEWORK, WE WILL PROVIDE THE PHASES DEVELOPED TO BUILD THE NSC, THE POSSIBLE CORPORA TYPOLOGY, AND, FINALLY, THE CRITERIA AND PARAMETERS WE HAVE ADOPTED TO BUILD THE NSC, AND WHICH, IN SHORT, CHARACTERIZED OUR CORPUS. THESE CRITERIA WILL BE DEVELOPED SEQUENTIALLY TO END WITH A SUMMARY OF THEM.

KEYWORDS:
TERMINOGRAPHY; SPECIAL-PURPOSE CORPUS; CORPUS DESIGN

REFERENCIAS BIBLIOGRÁFICAS:

Ahmad, K. Pragmatics of Specialist Terms: The Acquisition and Representation of Terminology. En Steffens, P. (ed.). *Machine Translation and the Lexicon. Proceedings of the 3rd. International EAMT Workshop*, Berlin/NewYork: Springer Verlag, 1995, p. 51-76.

Ahmad, K. y Rogers, M. Corpus Linguistics and Terminology Extraction. En Wright, S. E. y Budin, G. (eds.). *Handbook of Terminology Management*. Vol.2, Amsterdam/Philadelphia: John Benjamins, 2001, p.725-760.

Alcaraz Varó, E. *El inglés profesional y académico*. Madrid: Alianza Editorial, 2000.

Atkins, B.T.S. Clear, J. y Ostler, N. Corpus Design Criteria. En *Literary and Linguistic Computing*, vol.7, n. 1, 1992, p.1-16.

Bach, C., Saurí, R., Vivaldi, J. y Cabré, M.T. *El corpus del IULA: descripció*. Serie Informes, 17, 1997.

Berber Sardinha, A. P. Lingüística de Corpus: histórico e problemática. En *D.E.L.T.A. 16* (2), 2000, p. 323-367.

Bergenholt, H. y Tarp, S. *Manual of Specialised Lexicography: the Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins, 1995.

Biber, D. Conrad, S. & Reppen, R. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

Bowker, L. Towards a Corpus-Based Approach to Terminography. En *Terminology*, 3(1), 1996, p. 27-52.

Bowker, L. y Pearson, J. *Working with Specialized Language. A practical guide to using corpora*. London/New York: Routledge, 2002.

Cabré, M.T. *La terminología: representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 1999.

Engwall, G. Not Chance, but Choice: Criteria in Corpus Creation. En Atkins, B.T.S. y Zampolli, A. (eds.). *Computational Approaches to the Lexicon*. Oxford: OUP, 1994, p. 49-82.

Faber, P. ONCOTERM: Sistema bilingüe de información y recursos oncológicos. En Alcina Caudet, A. y Gamero Pérez, S. (eds.). *La traducción científico-técnica y la terminología en la sociedad de la información*. Castellón de la Plana: Publicacions de la Universitat Jaume I, 2002, p. 177-188.

Fillmore, Ch. J. "Corpus linguistics" or "Computer-aided armchair linguistics". En Svartvik, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82.* Berlin/New York: Mouton de Gruyter, 1992, p.35-60.

Ghadessy, M. Henry A. y Roseberry, R.L. (eds.). *Small Corpus Studies and ELT. Theory and practice*. Amsterdam/Philadelphia: John Benjamins, 2001.

Gómez, A. y Vargas, Ch. Una herramienta de traducción asistida: la aplicación *Multiconcord* en la extracción de terminología bilingüe. En Gallardo San Salvador, N. (ed.). *Terminología y traducción: un bosquejo de su evolución*. Granada: Atrio, 2003, p. 227-241

Gómez, A. y Vargas, Ch. Aspectos metodológicos para la elaboración de diccionarios especializados bilingües destinados al traductor. En *Actas del II Congreso «El español, lengua de traducción»*. Bruselas: ESLEtRA, 2004, p. 365-398.

Halliday, M. A. K. Some Grammatical Problems in Scientific English. En Halliday, M.A.K. y Martin, J.R. (eds.). *Writing Science: Literary and Discursive Power*, 1993, p.69-85.

Leech, G. The state of the art in corpus linguistics. En Aijmer K. y Altenberg B. (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 1991, p 8-29.

McEnery, T. y A. Wilson. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

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

Pearson, J. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins, 1998.

Qi-bo, Z. A quantitative look at the Guangzhou Petroleum English Corpus. En *ICAME Journal*, 13, 1989, p.28-38.

Roe, P. *Scientific Discourse Analysis*. Monographs No 4. ELR Birmingham University, 1977.

Sánchez, A., Sarmiento, R., Cantos, P. y Simón, J. *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL, 1995.

Sinclair, J. *Preliminary recommendations on Corpus Typology*. EAG-TCWG-CTYP/P. Pisa: EAGLES. Versión de mayo 1996.

Stubbs, M. *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. Oxford/Cambridge (MA): Blackwell Publsihers, 1996.

Vargas, Ch. A pragmatic model of text classification for the compilation of special-purpose corpora. En Mateo, J. y Yus, F. (eds.). *Thistles. A homage to Brian Hughes. Essays in Memorian*, vol. 2, 2005, p. 295-315.

Vargas, Ch. *Aproximación terminográfica al lenguaje de la piedra natural. Propuesta de sistematización para la elaboración de un diccionario traductológico*. Tesis doctoral dirigida por E. Alcaraz Varó. Universidad de Alicante: Alicante, 2005.

Chelo Vargas Sierra es licenciada y doctora en Traducción e Interpretación por la Universidad de Alicante (UA). Comenzó su trayectoria profesional universitaria como becaria de Formación de Profesorado Universitario de los proyectos de investigación subvencionados por el MEC: «Creación de una base terminológica de algunos sectores industriales de la Comunidad Valenciana» (PB98-0963) y «El lenguaje industrial del sector textil y de los juguetes de la Comunidad Valenciana: estudio contrastivo (español-inglés) de su terminología y sus estrategias comunicativas» (BFF2002-01457).

En la actualidad, trabaja como profesora en el Departamento de Filología Inglesa y participa también en el Máster Universitario de Traducción Inglesa de la UA y en el Curso de Posgrado en Tecnologías de la Traducción y la Localización de la Universitat Jaume I de Castellón. Es investigadora del grupo de investigación consolidado *El Inglés Profesional y Académico* de la UA y miembro fundador del Instituto Interuniversitario de Lenguas Modernas Aplicadas (IULMA). Las líneas de investigación desarrolladas hasta la fecha se centran, por un lado, en el campo de la lingüística aplicada con la reciente puesta en marcha del IULMA y, por otros, en la terminología (principalmente terminografía y terminótica), así como la Lingüística de corpus y las tecnologías de Traducción Asistida por Ordenador,

temas sobre los cuales ha presentado varias comunicaciones a congresos y ha publicado diversos artículos en revistas y monografías.

Entre las publicaciones más relevantes destacan el *Diccionario de términos de la piedra natural e industrias afines* (inglés-español; español-inglés) (Ed. Ariel, 2005) y el *Diccionario de términos del calzado e industrias afines* (2006).