Extracción de candidatos a términos de un corpus de la lengua general

Gilberto Anguiano Peña *
Catalina Naumis Peña **

Artículo recibido:
21 de octubre de 2013.
Artículo aceptado:
9 de octubre de 2014.

Resumen

Entre los objetos de estudio de la Bibliotecología e Información se incluyen los fenómenos lingüísticos asociados al análisis de contenido documental tanto para organizar la información como para recuperarla. Para ello, se deben rescatar los términos usados en el lenguaje científico y técnico, estudiar su ámbito de dominio y comportamiento. A través de la lengua se controla y se excluye el conocimiento que una población pueda obtener. El análisis documental del contenido, en este caso de los textos de difusión científica, permite obtener un conocimiento de las unidades léxicas, sus aplicaciones significativas y separar los términos de

* El Colegio de México, México. ganguia@colmex.mx
** Instituto de Investigaciones Bibliotecológicas y de la Información de la UNAM, México. naumis@unam.mx

la lengua general para crear lenguajes de indización. Es así que por medio del análisis de contenido documental en un corpus de lengua general marcado con los métodos de la lexicografía se obtienen y caracterizan los componentes que permiten extraer unidades léxicas del lenguaje especializado mediante las técnicas propuestas en el presente trabajo.

Palabras clave: Análisis de contenido; Extracción de términos; Lenguaje científico; Corpus de lengua general.

Abstract

Extraction of candidate terms from a corpus of non-specialized, general language

Gilberto Anguiano-Peña and *Catalina Naumis-Peña*

Linguistic phenomena associated with the analysis of document content and employed for the purpose of organization and retrieval are well-visited objects of study in the field of library and information science. Language often acts as a gatekeeper, admitting or excluding people from gaining access to knowledge. As such, the terms used in the scientific and technical language of research need to be kept up and their behavior within the domain examined. Documental content analysis of scientific texts provides knowledge of specialized lexicons and their specific applications, while differentiating them from common use in order to establish indexing languages. Thus, as proposed herein, the application of lexicographic techniques to documental content analysis of non-specialized language yields the components needed to describe and extract lexical units of the specialized language.

Keywords: Content Analysis; Term Extraction; Scientific Language; Corpus of General Language.

Introducción

El objetivo general de este trabajo es determinar metodologías y estrategias para procesar un corpus lingüístico de la lengua general, con la fina-

lidad de obtener los términos especializados de una disciplina y los términos compartidos en varias disciplinas con la finalidad de separarlos automáticamente de la masa de unidades léxicas de la lengua general. Este tipo de trabajo permite obtener términos y esclarecer posteriormente su significado, conocer los usos en los textos donde aparecen y utilizarlos en la construcción de lenguajes documentales.

¿Es el lenguaje de la ciencia un lenguaje secreto? En un periódico digital de Murcia, España, se presenta un trabajo sobre difusión de la ciencia titulado *El lenguaje secreto de la ciencia*, cuya autora sostiene que "desde el álgebra hasta la geometría pasando por la aerodinámica, las matemáticas están presentes en todas las ramas científicas y son básicas en nuestra vida diaria" (Moreno, 2011: s. pág.), otorgando así a las matemáticas la secrecía del lenguaje porque permea las ciencias en general. La investigación sobre matemáticas básicas que presenta justifica esta aseveración y le otorga garantía a través, entre otros, del caso de Manuel Saorín, catedrático de matemáticas y grupo de excelencia, quien afirma: "Sacar dinero de un cajero o mandar un correo electrónico no sería posible sin el álgebra" (Moreno, 2011: s. pág.). Sin duda, no se refiere a la claridad del lenguaje utilizado en las ciencias a través del uso de términos para transferir información, pero se observa una posición de complacencia acerca de un lenguaje escondido en otros de las ciencias.

Hay que reflexionar, pues, sobre el hecho de que en la actualidad la humanidad cuenta con valiosos recursos y nuevas tecnologías para obtener y difundir la información, entre los que destacan las telecomunicaciones, la radio, la televisión, la telefonía y la transmisión de datos, el libro digital interactivo, entre otros; de entre todos estos recursos sobresalen el Internet y las redes sociales como Facebook y Twitter. Con el Internet los seres humanos pueden tener acceso al conocimiento global desde teléfonos, tabletas, computadoras fijas o portátiles, televisores, etcétera.

Este acceso ha sido posible desde hace ya muchos años, ¿pero entonces? ¿Cómo es que las ciencias, las técnicas y el conocimiento se incrementan y difunden con amplitud, sin que la población mundial lo haya aprovechado para incrementar su bienestar? La respuesta es que hay un tipo de analfabetismo funcional[1] que no permite a buena parte de los seres humanos decodi-

1 Una definición para este término la encontramos en Jiménez del Castillo: "El analfabeto funcional sería aquella persona que ante una información (o conocimiento en codificación alfabética) es incapaz de operativizarla en acciones consecuentes y, en este sentido, diremos que no posee la habilidad de procesar dicha información de una forma esperada por la sociedad a la que pertenece" (2005: 290). A esto mismo se puede agregar un aspecto que aclara la Wikipedia: "El analfabetismo funcional también limita seriamente la interacción de la persona con las tecnologías de la información y la comunicación, puesto que tiene dificultades para usar un ordenador personal, trabajar con un procesador de texto o con una hoja de cálculo y utilizar un navegador web o un teléfono móvil de manera eficiente" ("Analfabetismo funcional", 2013: s. pág.).

ficar el significado de muchos de los mensajes de difusión científica, aunado al uso de términos desconocidos por la población, lo cual debe contrarrestarse con acciones como las propuestas por López-Barajas (2009) respecto a la alfabetización virtual.

En esencia el problema se hace patente cuando la población requiere entender las palabras, frases u oraciones de los textos de difusión científica para poder interpretarlos en cuanto a su significado y no lo logran. Por lo mismo no se pueden beneficiar del conocimiento científico y tecnológico; esto produce que a la mayoría de las personas el código de las ciencias les resulte opaco, oscuro, ajeno, y refuercen su idea de que el lenguaje de la ciencia es casi un lenguaje secreto.

Lo anterior en realidad es muy normal porque en una comunicación científica las personas deben poseer los elementos básicos para codificar, transmitir, decodificar e interpretar los significados científicos o técnicos, y quienes no poseen estos elementos quedan excluidos automáticamente de la comunicación especializada que sostienen emisores y receptores de la ciencia y de la tecnología, como puede consultarse en *El proceso de comunicación* de Sánchez González (2010).

El problema es que la ausencia de claridad parecería repetirse en los textos de difusión que participan en la conformación de los corpus lingüísticos de la lengua general. El *Corpus del Español Mexicano Contemporáneo, 1921-1974* (*CEMC*) es utilizado como base para desarrollar el presente estudio; en esta obra se integraron textos de difusión científica y textos de educación formal a nivel licenciatura, en los cuales aparecen términos del lenguaje especializado que se rescatan de la masa del corpus para analizar. En este trabajo no se tratarán aspectos semánticos de los términos, únicamente la metodología para aislarlos del corpus de la lengua general.

El enfoque sobre la comunicación y otros aspectos del texto

La Bibliotecología e Información actualmente consideran, entre otras cosas, que para ayudar a los usuarios a acceder a la información, el punto de partida es dejar claro que la idea principal de la comunicación de un mensaje de difusión de conocimiento es que se entienda su significado, con lo cual es necesario observar lo que ocurre con el signo lingüístico y sus componentes: significante, referente y significado, para que exista una comunicación efectiva.

Cuando se pretende trabajar con textos, como ocurre en general en la Bibliotecología e Información, es necesario establecer que existen varios aspectos inherentes a la necesidad de comunicar algo, como lo aclaran autores como

Luis Fernando Lara (1977, 1984, 1996, 1999, 2001; y Jetta Zahn, 1973), Ana María Cardero García (1998; 2003; 2004; 2005 y 2009) y Catalina Naumis Peña (1997; 1999; 2000 y 2003) en México, al igual que especialistas internacionales como Juan Carlos Sager (1993), Juana Marinkovich (2008), Rosa Estopà (1998), Rita Temmerman (2000) y su teoría sociocognitiva de la terminología, María Teresa Cabré y Rosa Estopà (2002) y María Teresa Cabré (1999a, 1999b, 2002), con sus dos propuestas sobre la Teoría de las puertas y la Teoría Comunicativa de la Terminología (TCT). Estos especialistas argumentan que hay que tomar en cuenta el contexto en que se hace uso de una unidad léxica y su correlación con el resto de la lengua para entender su significado verdadero, el cual a su vez está marcado por el consenso común de los hablantes.

Si se consideran pertinentes las propuestas teóricas de estos especialistas, habría que recurrir también a la sociolingüística respecto al contexto de situación, campo, tenor y modo (Halliday, 1979) y a la sociología cuantitativa urbana o variacionismo,[2] que señala la posición socioeconómica del hablante así como su formación cultural. Con esto se aclarará que el acto comunicativo del lenguaje científico comprende las circunstancias espaciales y temporales en que se desarrolla y esto obliga a que, al estudiar el objeto llamado texto, se tome en cuenta también al *contexto lingüístico*, que se refiere a los factores vinculados a la producción de un enunciado; este mismo contexto afecta la interpretación, la adecuación y el significado del mensaje (por medio de la gramática, la sintaxis, el léxico y el contexto). También hay que tomar en cuenta al contexto o situación extralingüística, que es el conjunto de los partícipes potenciales en la comunicación, tales como el lugar, tipo de registro y momento en que se concreta un acto lingüístico.

El estudio y mantenimiento de registros lingüísticos es sumamente importante para esclarecer los términos, pues incluye el conjunto de variables contextuales y sociolingüísticas que condicionan el modo en que una lengua es usada en un contexto socioeconómico concreto. Es decir, al analizar un registro lingüístico se define si lo comunicado está ubicado con un uso de la lengua estándar, no-estándar, culta, subculta o pertenece a una comunicación formal o informal, entre otras situaciones, como se estableció en la estratificación del *Corpus del Español Mexicano Contemporáneo, 1921-1974* (*CEMC*) (Lara y Ham Chande, 1979: 7-39), del cual se obtuvieron los resultados que utilizamos en la elaboración de este artículo.

En este mismo sentido, cuando se analizan textos científicos es importante indicar que la ciencia es un tipo de comunicación basada en registros de

2 "*Sociolingüística cuantitativa urbana* o *variacionismo* (esta rama estudia la variación lingüística asociada a factores sociales que se da en un hablante o en una comunidad de hablantes)" (*DTCE*, 2014: s. pág. Cursivas desde el original).

uso y situaciones formales donde el emisor selecciona los recursos lingüísticos adecuados, en los registros especializados, destinados a un receptor cuyo nexo común es el interés en una actividad especializada o profesional específica. Estas características ayudan a diferenciarlo e identificarlo de los registros pertenecientes a otros contextos socioculturales como el estudiado en este caso. Las situaciones profesionales se caracterizan por utilizar un vocabulario técnico propio del área de interés y el uso de expresiones con un significado especial. Los mensajes que se transmiten son regularmente por escrito. Sin embargo, los autores científicos en la vida real no pueden comunicar su mensaje tal como lo propone Wüster (2003) en la Teoría General de la Terminología (TGT), esto es, con las unidades terminológicas exclusivas de su disciplina, ya que también necesitan usar unidades léxicas de la lengua general e incluso unidades léxicas especializadas que se emplean en otras disciplinas.

Hay aspectos importantes a tomar en cuenta cuando se selecciona a un autor para analizar las unidades léxicas usadas en una de sus obras o textos, pues éste en realidad funge como una autoridad digna de ser seguida y respetada, por lo cual hay que seleccionar a uno productivo y de los más citados en su campo. También hay que considerar otros aspectos, como su lugar de nacimiento, estrato socioeconómico, vivencias individuales, cultura, ideología, religión, postura política, tradición verbal, idioma, formación profesional, investigaciones previas individuales y en equipo, experiencia, libertad de expresión, intereses individuales, actualización, especialización científica y el tipo de documentos o textos que produce, pues pueden ser de tan distinta índole como los siguientes: cartas, comunicados, informes, tesis, reportes de investigación, artículos, libros, ponencias, dictámenes, normas, leyes, reglamentos o documentos de divulgación.

Para situar la producción de los términos que se quieren analizar documentalmente es necesario identificar el tipo de documento o texto científico, definir si proviene de una autoridad en la materia, si corresponde a una comunicación oral o escrita, si fue elaborado con premura o fue preparado con tiempo o si fue un tema libre o dirigido, por mencionar algunos aspectos. También hay que considerar el uso de las expresiones especializadas porque en general los autores científicos son meticulosos en la elección de las unidades léxicas que utilizan en sus textos, con la finalidad de disminuir al mínimo las ambigüedades en la comunicación científica y técnica. Sin embargo, un autor puede tener o no éxito al seleccionar las palabras más precisas para conseguir su objetivo comunicativo, pues en su mente puede haber infinidad de razonamientos que guían la elección de las unidades léxicas y de las unidades terminológicas de su discurso, como pudieran ser la situación misma de la elaboración del discurso, el idioma en que se produce y el uso correcto de no-

menclaturas, nombres propios, abreviaturas, acrónimos, siglas, expresiones fijas, claves, contraseñas, conceptos, números escritos y en cifras, símbolos, fórmulas, convenciones, etc., lo cual favorece o no la materialización de una unidad terminológica en los textos especializados.

Como se puede ver hay una gran cantidad de factores que pueden influir en la selección de unidades léxicas y terminológicas por parte de un autor, pues además hay formas simples, sintagmas, expresiones fijas o frases terminológicas pluriverbales, y a todo esto se puede añadir otro tipo de información, de mayores proporciones, que se hace presente frecuentemente en los textos académicos, técnicos y especializados. Esta información es el uso de citas y transcripciones, que se estudian para identificar la existencia de una gran cantidad de menciones de lo que otros han dicho, ya sea en forma de pensamientos o de comprobaciones científicas (Cunha, 2004). Estos datos muchas veces aparecen en el idioma original en que se produjeron, como el latín, el griego, el inglés, el francés, etc., y aparecen acompañados del aparato crítico correspondiente.

El análisis de contenido

Es oportuno situarnos en la idea de que, para solucionar los problemas científicos, tanto las ciencias como las técnicas en sus búsquedas del conocimiento utilizan el método de análisis para efectuar sus investigaciones, y que en cada disciplina o campo del conocimiento humano se utiliza un tipo de análisis particular y coherente con ella. Para la Bibliotecología e Información hay algunas técnicas que resultan dignas de ser tomadas en cuenta por usar métodos afines y/o complementarios de análisis. Por supuesto que hay infinidad de disciplinas de las que se pudiera tomar conocimiento útil para esta materia, pero en los hechos tradicionalmente existen áreas cercanas que complementan el conocimiento correspondiente, tales como la Lingüística, la Lingüística Aplicada en todas sus vertientes y la Computación, entre muchas más.

En estas disciplinas se desarrollan análisis que son susceptibles de ser utilizados en estudios multidisciplinarios, por ejemplo: análisis de contenido, análisis del discurso, análisis gramatical, análisis cualitativo, análisis cuantitativo, análisis de definiciones analíticas, análisis de fraseología contrastiva, análisis lexicológico, análisis de documentos, análisis de las relaciones conceptuales, análisis de textos, análisis de unidades sintagmáticas, análisis y diseño de corpus lingüísticos, análisis de términos y por último el método del análisis documental de contenido usado para transferir información.

En la introducción del libro *La ciencia del texto: un enfoque interdisciplinario*, Teun A. Van Dijk explica cómo el análisis del discurso se estudia desde diferentes disciplinas científicas y qué tanto existe una "conexión transversal" interdisciplinaria. Van Dijk parte del supuesto de que en el uso de la lengua, la comunicación y la interacción se producen a través de textos o discursos. La lingüística estudia una parte del uso de la lengua, pero otras ciencias también lo hacen: la sociolingüística, la comunicación, la psicología cognitiva, la pedagogía, la jurisprudencia, la ciencia política, la sociología y, por supuesto, la bibliotecología. Las relaciones textuales o discursivas se dan entre distintos tipos de textos, las estructuras textuales subyacentes, sus diferentes condiciones y funciones, los contenidos y los efectos que producen en los hablantes (Van Dijk, 1992: 9-10).

Los distintos tipos de textos, las relaciones entre ellos y con la sociedad tienen conexiones de diversa índole que se analizan desde puntos de vista distintos, de acuerdo al campo disciplinario desde donde se realice. Las ciencias del texto se interesan por profundizar en las propiedades y características comunes del uso de la lengua en el espectro de disciplinas que abarcan las ciencias sociales y humanas.

El área de análisis y sistematización de la información que integra la disciplina Bibliotecología e Información se concreta a describir los tipos de textos, los datos y los contenidos informativos que lleven a su localización en los sistemas. Sin embargo, la utilización de procesos comunes con otras disciplinas es innegable, entre ellos el análisis terminológico y lexicográfico usado en este trabajo.

La documentación en la lexicografía

El análisis documental como lo presenta Rubio Liniers (2004) también se aplica en la lexicografía, pues es parte del proceso para elaborar diccionarios; de hecho, como también lo destaca Gómez González-Jover (2005), es el método imprescindible para representar el contenido de los documentos que conforman el corpus donde se incluyen las unidades lexicográficas definidas en el diccionario. La representación de los contenidos que se lleva a cabo permite la consulta y recuperación por diferentes puntos de acceso; además, casi siempre se pueden generar, con la información resultante de este tipo de análisis, nuevos productos para satisfacer las necesidades de información léxica, como son las concordancias, datos estadísticos, índices y diccionarios.

El análisis documental de contenido ayuda a la decodificación de los mensajes y a la recuperación de la información pertinente para los usuarios del sis-

tema documental del proyecto *Diccionario del español de México* (*DEM*).[3] Lo anterior se sustenta en que el autor ya hizo su mensaje y está contenido en un soporte documental, por lo general en textos escritos que pertenecen a una especialidad. Por lo tanto, corresponde a los centros informativos hacer que los contenidos de esos documentos, como pueden ser los candidatos a términos, sean de fácil consulta y recuperación por parte de sus usuarios.

La elaboración del *DEM* está sustentada en la Lingüística de Corpus, misma que establece pautas para mantener y ofrecer una gran capacidad y versatilidad en el manejo de la información contenida. Al igual que cualquier otro sistema de información, la Lingüística de Corpus define las entradas y puntos de acceso que se deben incluir. Aunque existen en estos días corpus de tipo multimodal (voz, imagen, texto, etc.), los corpus utilizados hasta hace poco tiempo por las ciencias y las técnicas tienen como objetivo, de manera general, analizar bajo sus distintas modalidades y características las palabras o unidades léxicas contenidas en textos de lengua general o de lenguajes especializados, y en el caso estudiado está aplicado a sustentar comunicaciones de difusión científica.

El proceso documental en la lexicografía básicamente requiere cumplir ciertas etapas para aplicarse, como las siguientes:

- Actividades de planeación, como el establecimiento de metas, objetivos, organización y metodologías a implementar.
- Se establece como comienzo del proceso propiamente dicho la selección y adquisición de los documentos. Para el caso de las grabaciones con informantes, se les transcribe.
- Se efectúa el tratamiento documental en el aspecto externo, que implica preparar físicamente el material y así obtener el archivo correspondiente para analizarlo posteriormente.
- Se procede a efectuar la descripción bibliográfica del documento, resaltando los puntos de acceso que permitirán su identificación en relación con los otros documentos. Para los textos impresos se efectúa su descripción, que incluye: autor, título, pie de imprenta y descripción física del material. Aunado a esto, en la lexicografía se incluyen datos externos al documento de interés para la sociolingüística, la pragmática y la semiótica; estos datos en general corresponden a la unidad de comunicación analizada, en la que se destaca al emisor, la situación en que se generó la comunicación y el canal utilizado en la misma. Tam-

3 Este proyecto inició sus actividades en 1973 y desde un principio, como lo menciona Barcala Rodríguez (2010) para otros corpus, el *DEM* estructuró su sistema de recuperación de información lexicográfica tomando como base la Lingüística de Corpus.

bién se añade un contexto extralingüístico o registro de habla, con el que se puede identificar la formalidad o informalidad con la que se escribieron los documentos, también si el texto fue destinado a una audiencia general o a una especializada. De estos registros en su conjunto dependerá la posterior identificación situacional y temática respecto al uso de las unidades léxicas, en consecuencia esto ayudará a que los usuarios del sistema asignen significado a las unidades léxicas de la información recuperada.

- Respecto al texto o contexto estrictamente lingüístico, los textos escritos en una disciplina científica en general deberán contar con los componentes del signo lingüístico (significante, significado y referente), siendo el texto más pequeño el equivalente a un párrafo separado por un punto y aparte, o un ítem. Los textos se analizan por medio de programas y algoritmos previamente determinados con el fin de obtener la información contenida en el documento. En general, del análisis se extraen las formas gráficas de las palabras o unidades léxicas, tal y como se encuentran en los textos del lenguaje natural, ya sean del lenguaje común o de lenguajes especializados.

En la Bibliotecología e Información, cuando se indiza con lenguaje natural el término es aislado de su contexto. El método de trabajo es el análisis textual[4] del documento científico, para después efectuar el análisis documental de contenido, teniendo como principal objetivo la indización por lenguaje natural; aquí se aprovecha el mismo texto para extraer los términos de indización. Derivado de esto se obtienen las listas de significantes o unidades léxicas, separadas de sus significados y referentes; de esta forma queda fragmentado el signo lingüístico, lo que produce que al usuario se le complique la recuperación de la información que necesita y por eso requiere ser apoyado en sus búsquedas.

A diferencia de este método, cuando en la lexicografía se extraen términos para conformar corpus lingüísticos se obtienen distintas listas, que pueden ser de palabras simples o compuestas, con su categoría gramatical, por su morfología, según su estructura interna, según su número de sílabas, o también como colocaciones, unidades fraseológicas, sintagmas compuestos, enunciados fraseológicos, palabras significativas, palabras clave, palabras vacías, tecnicismos, neologismos o candidatos a términos.

De manera general, las unidades léxicas que se obtienen de la Lingüística de Corpus están acompañadas de datos cuantitativos (rango y frecuencia) y

4 Puede hacerse un corpus *ad hoc*, o pueden usarse programas comerciales de análisis de textos como *WordSmith*, *AntConc, Notepad*, *Atlas.ti*, *Sketch Engine*, entre otros.

se puede reconocer el ámbito de su origen mediante el registro de su uso, esto si principalmente pertenecen a los lenguajes especializados.

Exclusión terminológica mediante subconjuntos del lenguaje general

Cuando se busca extraer candidatos a términos de textos generales o especializados, resulta de la mayor utilidad tener en cuenta la información previa que existe sobre el léxico en general proveniente de estudios métricos de la información, como la informetría, la bibliometría, la cienciometría y la lexicometría, así como los cortes de Luhn y la obtención de pesos TF-IDF (Blázquez, 2013), esto con el objetivo de hacer filtros para excluir la lengua común y recuperar especialmente los candidatos a términos.

Por otra parte, en este artículo se plantea que además de los indicadores enunciados arriba se pueden usar otros muy parecidos basados en el lenguaje natural para excluir subconjuntos del lenguaje general, entre ellos el vocabulario fundamental (parecido al índice de mayor frecuencia y al modelo de Zipf), el léxico común (basado en el índice de dispersión) y la lista de palabras gramaticales (equivalente a las palabras vacías), con el objeto de aislar al máximo las unidades especializadas que se buscan en el texto. Es decir, se puede reutilizar el conocimiento lexicográfico, en este caso el producido por el proyecto *Diccionario del español de México* (*DEM*) y su *Corpus del Español Mexicano Contemporáneo, 1921-1974* (*CEMC,* 1975), con el objeto de simplificar la información que se pretende analizar.

En estas páginas se utilizan algunos resultados del análisis de contenido efectuado en el *CEMC*, el cual se estructuró con cerca de dos millones de palabras etiquetadas gramaticalmente; de este corpus se obtuvo a su vez un producto lexicográfico, que es propiamente un índice estadístico de lenguaje natural con información léxica, gramatical, sociolingüística, registros de uso de la lengua y datos cuantitativos denominado *Diccionario estadístico del español de México* (*DEEM*, 2005).

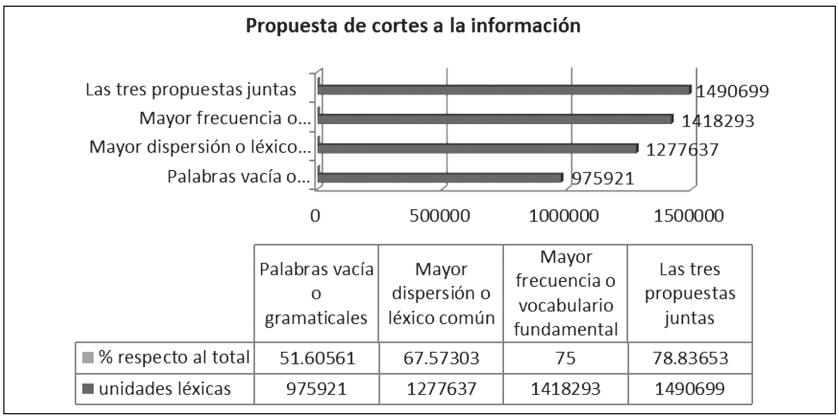
Los resultados obtenidos del *DEEM* respecto a las palabras vacías, mayor dispersión y mayor frecuencia fueron los siguientes:

1) Unidades léxicas gramaticales o *palabras vacías*. Son principalmente artículos, preposiciones, interjecciones, pronombres, etc. Equivalen a 292 lemas que son el 51.60 % del total de información del corpus. Este es el tercer grupo de términos que se excluyen cuando se busca extraer términos científicos y técnicos.

2) Las unidades léxicas con mayor dispersión o *léxico común* (Anguiano Peña, 2013a). Estas unidades son 994 lemas distintos que correspondieron al 67.57 % del total de la información del corpus. Cuando se hace la búsqueda de términos especializados, este tipo de unidades léxicas suelen separase del análisis de contenido documental.
3) Las unidades léxicas con la mayor frecuencia o *vocabulario fundamental*, las cuales presentó Lara (2007). En este rubro hay que considerar que fue a partir de estudios de lexicometría, informetría, el modelo de Zipf (Zipf, 1949), entre otros, como ha sido posible comprender que existe un fenómeno económico en el uso del lenguaje, denominado del "menor esfuerzo", que básicamente describe cómo es que el ser humano utiliza una enorme cantidad de palabras gráficas que corresponden a una muy pequeña cantidad de lemas, lo que da por resultado que haya un número muy reducido de unidades léxicas con una frecuencia muy elevada. Siguiendo este razonamiento se comprende que el vocabulario fundamental o el de mayor índice de frecuencia sea el más usado en los textos y discursos, como en el *CEMC* en el que apenas 861 lemas tienen el 75 % del total de información del corpus. Se sugiere que este tipo de unidades léxicas también sean eliminadas.

En la *Figura 1* y la *Tabla 1* se muestran los resultados obtenidos sobre estos tres rubros, con lo que se explica su exclusión del análisis por significar un drástico ahorro.

Figura 1. Propuesta de cortes: palabras vacías, mayor frecuencia, mayor dispersión y las tres juntas, respecto a 1 891 058 unidades léxicas (%)



	Palabras vacía o gramaticales	Mayor dispersión o léxico común	Mayor frecuencia o vocabulario fundamental	Las tres propuestas juntas
% respecto al total	51.60561	67.57303	75	78.83653
unidades léxicas	975921	1277637	1418293	1490699

Fuente*:* elaboración de Gilberto Anguiano Peña para su investigación de doctorado (2015)

Tabla 1. Resumen de palabras vacías, mayor dispersión, mayor frecuencia y las tres juntas

Concepto	Palabras gráficas	% respecto al total
Palabras vacía o gramaticales	975 921	51.60561
Mayor dispersión o léxico común	1 277 637	67.57303
Mayor frecuencia o vocabulario fundamental	1 418 293	75
Las tres propuestas juntas	1 490 699	78.83653

Fuente: elaboración de Gilberto Anguiano Peña para su investigación de doctorado (2015)

Como se puede observar, los tres subgrupos integrados no son la suma de ellos mismos, esto porque hay unidades léxicas que se repiten en dos subgrupos o incluso en los tres. Si se considerara que en conjunto pueden alcanzar hasta un 78.83 % del total de la información analizada, resulta entonces de interés para la recuperación de información elaborar filtros con la información del lenguaje general antes del análisis de contenido, con lo que se ahorraría cerca del 80% de la recuperación de los candidatos a términos (esto coincide con lo calculado en otros estudios de recuperación de información). Para hacer más eficiente el trabajo de recuperar términos científicos y técnicos, además, se establece un mínimo de apariciones válidas de las unidades léxicas para evitar la filtración de aquéllas con muy baja frecuencia, ya que pueden aparecer términos cuyo significado no tiene una garantía literaria.

Proceso documental para desambiguar significado y definir el uso de los candidatos a términos

La forma que se plantea en este trabajo para recuperar el texto de interés para los usuarios es que una vez que se obtenga el índice de significantes, equivalentes a la lista de candidatos a ser unidades terminológicas, se proceda a simplificarlas y lematizarlas; después hay que recuperar cada unidad por el registro de uso temático al que pertenece el o los documentos analizados donde se documentó, lo cual se convertiría en la práctica en algo parecido a señalar el léxico disponible del texto.[5] Con esto se ayudará al usuario "a desambiguar el significado y a encontrar el uso adecuado de ciertas voces" (Estopà, 1998: 360) y podrá solicitar posteriormente al sistema de recuperación de información el referente que más se acerque al que busca, simplificándose

5 Para López Morales (2013: s. pág. Cursivas desde el original):"El *léxico disponible* es el conjunto de palabras que los hablantes tienen en el lexicón mental y cuyo uso está condicionado por el tema concreto de la comunicación. Lo que se pretende es descubrir qué palabras sería capaz de usar un hablante en determinados temas de comunicación".

las búsquedas al mínimo. Sin embargo, y no obstante cualquier esfuerzo, el verdadero significado será siempre una interpretación del lector.

Al igual que en el proceso de la indización en la bibliotecología, los candidatos a términos o palabras clave pueden ser adecuados a un lenguaje controlado para mejorar la recuperación del contenido; esto se puede efectuar mediante la utilización de encabezamientos de materia o tesauros. Se realiza así la conversión de palabras del lenguaje natural obtenido de la indización a expresiones y conceptos de un lenguaje controlado.

Al final del proceso documental se difunde la información para hacerla llegar a los usuarios con el objeto de que se apropien de la misma. Para el caso de los proyectos lexicográficos se cuenta con distintos productos informativos derivados del análisis documental, que son destinados a los usuarios internos y externos. Estos pueden estar por separado o en conjunto como un sistema. Los componentes pueden ser la base de datos de las concordancias, parecidos a los KWIC (*Key Word in Context*), la información cuantitativa, los ficheros documentales, el propio diccionario que se elabora o las distintas interfaces generadas para consultar la información lexicográfica.

Pues bien, como parte de los resultados del largo proceso de análisis documental de contenido de corte lexicográfico de los textos, lo que se espera obtener al concluir la indización o clasificación por lenguaje natural es una lista de unidades léxicas significantes de la lengua general, pero también de las ciencias y de las técnicas con base en la presencia en textos relacionados con este ámbito de trabajo.

El aprovechamiento de las marcas de uso provenientes de la documentación lexicográfica

Fue con base en los resultados del *DEEM* que resultó posible conformar otra base de datos, el *Modelo sociolingüístico del léxico del español usado en México* (Anguiano Peña, 2006); después de asignar a las unidades léxicas del *DEEM* una indización semiautomatizada se pudo conseguir la suma de los resultados parciales que mostraba la anterior base y, una vez con los datos completos, se pudieron identificar los resultados totales de las unidades léxicas utilizadas en el lenguaje general usado en México por medio de sus registros sociolingüísticos (*Tabla 2*).

Tabla 2. Ejemplo de los registro de uso en la identificación de candidatos a términos en el Modelo sociolingüístico

Lemas	Cat. Gram.	Frecuencia total	% total	Uso del español	Nivel de lengua	Registros de habla	Mayor frecuencia	Mejor distribución	Clave de texto	Registro de uso 1	Registro de uso 2	Registro de uso 3
action	nom	4	0.00021	estándar	lengua culta							
actitud	s	259	0.01370	estándar			vocabulario fundamental					
activación	s	14	0.00074	estándar	lengua culta	ciencias			420, 427, 428, 454, 469, 473, 477, 478	Química	Medicina y veterinaria	Medicina humana
activado	adj	2	0.00011	estándar	lengua culta	ciencias			389, 478	Electrónica y electricidad	Medicina humana	
activamente	adv	12	0.00063	estándar	lengua culta							
actividad	s	511	0.02701	estándar			vocabulario fundamental					
activista	adj; s	6	0.00031	estándar	lengua culta							
acto	s	308	0.01629	estándar			vocabulario fundamental					
actor	adj; s	133	0.00704	estándar								

Fuente: elaboración de Gilberto Anguiano Peña para su investigación de doctorado (2015)

Propuesta para acotar los candidatos a términos

Para la búsqueda y recuperación de información especializada se propone eliminar, previamente al análisis de contenido documental de textos generales y especializados, los siguientes datos provenientes de los datos cuantitativos y de las marcas de uso del lenguaje general:

- Las unidades léxicas de mayor frecuencia.
- Las unidades léxicas de mayor dispersión.
- Las unidades léxicas pertenecientes al grupo de las palabras vacías.
- Las unidades léxicas que sean de la lengua no estándar.
- Las unidades léxicas que sean de la lengua subculta.

Si se eliminan del análisis las unidades enlistadas de tipo cuantitativo y sociolingüístico se podrá economizar sustancialmente en la recuperación de la información de candidatos a términos, pero lo importante es que después de obtener la lista de tales elementos se podrán comparar los registros de uso de la lengua que ya existen en este mismo *Modelo sociolingüístico del léxico del español de México*, comparación que ayudaría tanto a los usuarios de la información como a los profesionales de la Bibliotecología e Información en la reconstrucción del significado del signo lingüístico y la elaboración de un lenguaje controlado.

En esta nueva confrontación se podrán encontrar candidatos a términos que son exclusivos de uso de una disciplina, con lo que confirmarían primero que son palabras clave y, después de la validación de un experto, podrían llegar a ser términos en sentido estricto. Derivado de esto podrán reconocerse los candidatos que tienen uso en dos o más disciplinas, lo que indicaría que son términos en sentido lato y que incluso tienen polisemia, de forma que para la lexicografía son tecnicismos. También se podrá encontrar que existen candidatos que pertenecen a las ciencias pero también a las técnicas, con lo que podrían considerarse también tecnicismos, pero que pueden llevar la marca "Científ." en los diccionarios, es decir, que pertenecen al lenguaje científico.

Lo que también se propone en este artículo es la reutilización de los procesos lexicográficos para diferenciar las unidades léxicas y extraerlas mediante el análisis de contenido de los textos especializados, esto al utilizar las marcas de uso o registros de habla, como lo planteó Josette Rey-Debove (1971) cuando consideró tres aspectos fundamentales para lograr esta meta:

1) El conjunto de palabras (unidades léxicas) que pertenecen a una lengua o idioma.
2) La información sociolingüística de las unidades léxicas.
3) Las marcas de uso consensuadas por la propia comunidad de hablantes.

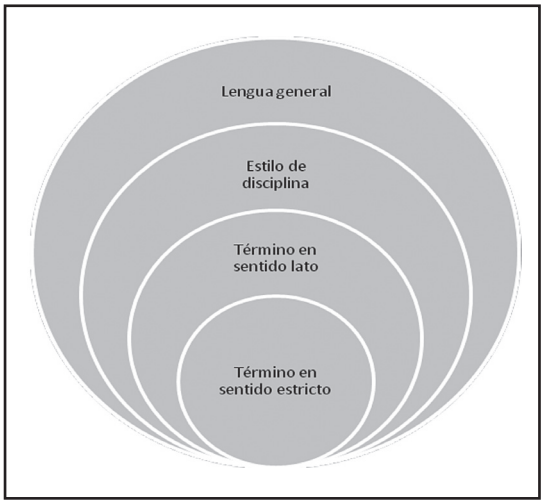
Al incorporar estos lineamientos en el análisis de la información se busca que sean las mismas unidades léxicas de la lengua general, identificadas por consenso, las que por contraste con el lenguaje especializado sirvan para clasificar primero los tecnicismos[6] y, por comportarse éstos de forma muy parecida a las unidades terminológicas, también se puedan designar los candidatos a términos.

La búsqueda de unidades terminológicas en los textos

Para obtener términos especializados con la ayuda de un corpus de la lengua general como el Modelo citado, primero se separan los candidatos a términos que tienen un registro de habla relacionado con un texto especializado. En esta etapa del proceso de la búsqueda de términos es normal encontrar, en los listados producidos por el análisis automatizado, unidades léxicas que pertenecen al uso de una disciplina en sus distintos niveles de comunicación, aunque todas estas unidades pertenezcan a la lengua estándar, a la lengua culta y a una ciencia o técnica. Esto mismo significa que en el análisis de contenido se pueden obtener las siguientes unidades léxicas de un texto general o científico: 1. unidades que pertenecen a la lengua general; 2. unidades que pertenecen al estilo de la disciplina analizada; 3. candidatos a términos en sentido lato y 4. candidatos a términos en sentido estricto, como lo muestra la *Figura 2:*

6 Al respecto, desde un enfoque lingüístico, a los términos se les puede llamar también tecnicismos, como se enuncia en la siguiente definición: "Tecnicismos. *m.* 1 Término que posee un sentido concreto y determinado dentro del lenguaje propio de un oficio, ciencia, arte o industria: la palabra ''algoritmo" es un tecnicismo de las matemáticas". (*DMLE*, 2007)

Figura 2. Unidades léxicas en el análisis de contenido documental de un texto



Fuente*:* elaboración de Gilberto Anguiano Peña para su investigación de doctorado (2015)

Para ahondar en lo planteado en el anterior párrafo, a continuación se explica con más detalle lo referente a las unidades a encontrar en los textos:

1) Unidades léxicas que pertenecen al lenguaje general y que aparecen en textos de las ciencias y las técnicas, pero que también son identificadas sociolingüísticamente como pertenecientes a la lengua general estándar, la lengua no estándar, la lengua subculta y a la lengua culta, es decir, que no son exclusivas de las ciencias ni de las técnicas. Se recomienda excluirlas del listado de candidatos a términos.
2) Unidades léxicas que corresponden al estilo de redacción característico de la disciplina que se analiza. Estas unidades son en general unidades léxicas pertenecientes a la tradición verbal de la disciplina, frases fijas y locuciones. Su aparición corresponde a un índice de frecuencia muy bajo respecto a un texto analizado; sin embargo, son características de ciertas disciplinas científicas por lo cual no es oportuno eliminarlas por anticipado del análisis de contenido. Aquí podemos encontrar locuciones, unidades fraseológicas, latinismos, etcétera.
3) Unidades léxicas especializadas o tecnicismos. Son de uso y significado propio de la disciplina a la que corresponde el texto analizado, si bien pueden tener un mismo significante en la lengua general e incluso en otras disciplinas, es decir, pueden llegar a tener sinónimos.

Este tipo de unidades léxicas son registradas en los diccionarios de lengua general,[7] y de hecho estas unidades son los términos en sentido lato.[8] Las formas de palabras gráficas, tal y como aparecen en el texto original, por lo general son pocas: femenino, masculino, singular y plural; tienen un índice de frecuencia en la lengua común muy bajo en el análisis de contenido documental, pero ya como unidades léxicas lematizadas (palabras agrupadas bajo su forma canónica) adquieren un porcentaje por demás elevado respecto del total de la muestra analizada; en otras palabras, un reducido número de unidades léxicas se agrupan en una cantidad elevada de lemas. En cuanto a su índice de dispersión en el *DEEM*, se observó que aunque pueden estar concentradas por su uso en una disciplina, puede ocurrir que también tengan apariciones en otras disciplinas de las ciencias, de las técnicas o pertenezcan al lenguaje científico, que abarca ambas áreas del conocimiento. Se les puede reconocer entre otras cosas porque aun teniendo un significante conocido, tienen un significado distinto al de la lengua natural, por eso el lector común no entiende su significado y le resulta un tanto secreto. Estas unidades pueden presentarse en una forma simple o como una forma pluriverbal, como sintagmas, frases hechas o como unidades fraseológicas.

4) Los candidatos a ser unidades terminológicas de la disciplina analizada. Éstos son muy parecidos en su comportamiento documental a los tecnicismos pero no tienen sinónimos y presuponen un significado unívoco. Estas unidades pertenecen a la lengua estándar, son de la lengua culta, son usadas exclusivamente en las ciencias o las técnicas, tienen un registro de habla que hace que se sitúen en una forma de comunicación formal y son usadas exclusivamente en un lenguaje especializado, por lo que no tienen significado ni equivalencia en la lengua común. Estos candidatos pueden tener la forma de unidades léxicas simples o unidades compuestas por varias palabras. Los candidatos pueden en principio ser considerados como palabras clave; después de ser validados por un especialista de la información pueden llegar a formar parte del lenguaje documental, y en el mejor de

7 Como en el *DRAE* (2001) o el *DEM* (2012).

8 En este estudio se utiliza como término en sentido lato lo propuesto por Cardero (2004: 42-43) en un trabajo dedicado al control de satélites, donde argumentó que los tecnicismos son "[…] designaciones de la lengua general que especializan su significado o designaciones que son comunes a varias áreas de conocimiento […]". Esto correspondería a un significado no frecuente con un significante frecuente.

los casos pueden ser términos en sentido estricto[9] de alguna disciplina. Su frecuencia de aparición es baja en el análisis de textos pero cuando se agrupan las unidades léxicas, en relación al total del análisis, resultan tener un porcentaje elevado de lemas. Carecen de dispersión pues sus datos están concentrados en una sola disciplina.

Al considerar todo lo anterior, también se puede esperar que en cualquier análisis documental de contenido de un texto, ya sea general o de la ciencia o de la tecnología, y teniendo en cuenta lo propuesto por Cardero (2004: 37), lo más probable es que las unidades léxicas analizadas de un texto científico o técnico tengan características que se presentan en la *Tabla 3* en lo que concierne al significante, el significado y al tipo de comunicación al que pertenecen.

Tabla 3. Características de las unidades léxicas analizadas

Significante*	Significado**	Tipo de lenguaje
Un significante común	y un significado común	forman parte de la lengua general.
Un significante no común	y un significado común	sería un tecnicismo de significante, por ejemplo, *close up*, *stock shot*, *feidear*.
Un significante común	con un significado no común	es un tecnicismo en sentido lato, por ejemplo, embobinadora, óptica, cámara.
Un significante no común	y un significado no común	sería un tecnicismo en sentido estricto, por ejemplo, borradora magnética, lámpara de proyección, sistema de pantalla translúcida, técnica de animación.

* Significante es el que señala algo, en este estudio una palabra o unidad léxica que se le da a una persona, animal, cosa o concepto tangible o intangible, concreto o abstracto, para distinguirlo de otros.
** Significado es lo señalado, y para nuestros intereses, la representación o concepto mental de algo.
Fuente: elaboración de Gilberto Anguiano Peña para su investigación de doctorado (2015)

A pesar de la coexistencia de unidades léxicas y unidades terminológicas en un texto científico, es posible diferenciarlas si se verifica su registro de habla, constatando si está ubicado en una forma de comunicación o en un texto que pertenezca exclusivamente a un lenguaje especializado; es decir, si se constata que son producto de una comunicación formal utilizada por los especialistas de alguna disciplina para asegurarse una comunicación efectiva entre ellos.

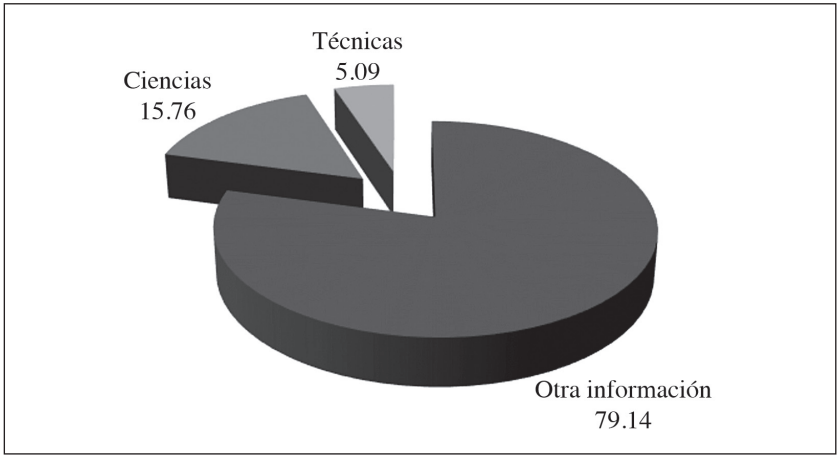
Como se puede observar en la descripción del proceso llevado a cabo y descrito en los párrafo anteriores, las unidades léxicas analizadas parten de un estudio empírico desarrollado por la lexicografía, el cual muestra que con los textos provenientes del lenguaje especializado ocurre algo parecido a lo

9 Se toma también lo propuesto por Cardero (2004: 43), quien considera a los términos en sentido estricto como "[…] las designaciones exclusivas de control de satélites […]", o sea que pertenecen a un sola disciplina, y serían de un significado y un significante no frecuentes.

que pasa con cualquier texto de lengua general, pues ambos tipos de textos están compuestos, en mayor o en menor proporción, con unidades léxicas del lenguaje general y no sólo de unidades especializadas de las ciencias o de las técnicas. Aunque parezca ser lo contrario, estas diferencias son en verdad útiles en la recuperación de la información, pues los términos que se pretenden extraer de los textos no son propios de la lengua común.

Ejemplo del tipo de análisis efectuado con el Modelo

Siguiendo los pasos propuestos en este mismo artículo y haciendo el aislamiento de los lemas correspondientes a las ciencias y a las técnicas contenidos en el Modelo, se consiguieron los siguientes resultados:



Gráfica 1. De un total de 30 899 lemas (palabras que podrían encabezar la entrada de un diccionario, acompañada de la definición) asignados en el *CEMC*, fueron recuperados como candidatos a términos 4 871 lemas en ciencias y 1 574 lemas en técnicas

Esta gráfica se desprende de 30 899 lemas de uso general en la lengua. Para obtener del corpus el 15.76 % de términos de uso exclusivo en la ciencia, que corresponden a 4 871 lemas, y el 5.09 % relacionado con 1 574 lemas en técnicas, se restringieron los lemas dos veces (véase inciso 4 del apartado "La búsqueda de unidades terminológicas en los textos"). Primero, los 30 899 lemas generales se redujeron a 16 296 lemas generales en el ámbito de las ciencias y las técnicas; de esta última agrupación se extrajeron los lemas exclusivos de las ciencias y de las técnicas. Los 6 450 lemas usados exclusivamente en estas áreas alcanzaron el 20.85 % del total de lemas del corpus.

Consideraciones finales

El Modelo expuesto aquí, u otros recursos lexicográficos con similares características, pueden ser útiles en un futuro cercano para la indización asistida por computadora o como corpus monitores respecto a nuevos análisis de textos o corpus especializados. Su utilización facilitaría la rápida generación de listas de significantes candidatos a términos, los cuales además de ser útiles para representar y recuperar el contenido del texto original, también serán de gran valía en la etapa del desarrollo del lenguaje controlado cuando se trabajen los términos, unitérminos, encabezamientos de materia o descriptores que conformen la terminología de alguna disciplina analizada de esta forma.

Hay que considerar asimismo que el lenguaje natural y el lenguaje especializado están en constante evolución, de lo que resulta consecuente que existan dificultades para controlar y recuperar los lenguajes especializados y sus terminologías, pero por esto mismo se hace más necesaria la presencia y el desarrollo de la Bibliotecología e Información con el fin de que ayuden a los usuarios y a los lectores a decodificar el lenguaje de la ciencia.

Referencias

Aguilar, C. A.; Alarcón, Rodrigo; Rodríguez, Carlos y Sierra Martínez, Gerardo (2006), "Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados", en María Teresa Cabré, Rosa Estopà y Carles Tebé Soriano (eds.), *La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad*, Barcelona, Institut Universitari de Lingüística Aplicada-Universidad Pompeu Fabra, pp. 259-269.

"Analfabetismo funcional" (2013), en *Wikipedia. La enciclopedia libre.* Disponible en: http://es.wikipedia.org/w/index.php?title=Analfabetismo_funcional&oldid=77706574 [Última revisión: 23 octubre 2014.]

Anguiano Peña, Gilberto (2006), *Modelo sociolingüístico del léxico del español usado en México,* México, El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Documento inédito.]

—— (2013a), *El léxico común del español de México,* México, El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Documento inédito.]

—— (2013b), *Palabras vacías del español de México.* México, El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Documento inédito.]

Barcala Rodríguez, Francisco Mario (2010), *Corpus lingüísticos estructurados de grandes dimensións: Metodoloxía e sistemas de recuperación de información,* tesis de doctorado, Universidade da Coruña, Departamento de Computación. Disponible en: http://ruc.udc.es/dspace/bitstream/2183/7171/1/tese_mario_barcala.pdf

Blázquez Ochando, Manuel (2013), *Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos,* Madrid, Universidad Complutense de Madrid. Disponible en: http://mblazquez.es/wp-content/uploads/ebook-mbo-tecnicas-avanzadas-recuperacion-informacion1.pdf

Bogomilova Lozanova Elena (2009), "Posibilidades y límites del análisis cuantitativo de corpus especializados", en Catalina Naumis Peña (coord.), *Memoria del I Simposio Internacional sobre Organización del Conocimiento: Bibliotecología y Terminología,* México, Universidad Nacional Autónoma de México, Centro Universitario de Investigaciones Bibliotecológicas, pp. 63-78.

Cabré, María Teresa (1999a), "La terminología hoy: concepciones, tendencias y aplicaciones", en *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos,* Barcelona, España, Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, pp. 17-37.

—— (1999b), "La terminología y documentación", en *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos,* Barcelona, España, Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, pp. 231-247.

—— (2002), "Textos especializados y unidades de conocimiento: metodología y tipologización", en Joaquín García Palacios y M. Teresa Fuentes (eds.), *Texto, terminología y traducción*, Salamanca, Ediciones Almar, pp. 15-36. Disponible en: http://www.upf.edu/pdi/dtf/teresa.cabre/docums/ca02te.pdf

—— y Estopà, Rosa (2002), "El conocimiento especializado y sus unidades de representación: diversidad cognitiva", en *Sendébar*, 13, pp. 141-153.

Cardero García, Ana María (1996), "La integración del corpus de la terminología de control de satélites en México", en *Actas del V Simposio Iberoamericano de Terminología*, México, Universidad Nacional Autónoma de México, pp. 106-111.

—— (1998), "Algunas observaciones de los conceptos, sus áreas temáticas. La sinonimia y la polisemia en tres vocabularios especializados en México", en *Actas del VI Simposio Iberoamericano de Terminología*, La Habana, Cuba, pp. 137-154.

—— (2003), "Unidad y variedad del español de América. Los vocabularios especializados", en Ignacio Guzmán Betancourt y María del Pilar Máynez Vidal (coords.), *Estudios de lingüística y filología hispánicas en honor de José G. Moreno de Alba,* México, Universidad Nacional Autónoma de México, pp. 299-322.

Cardero García, Ana María (2004), *Lingüística y terminología*, México, Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Acatlán.

—— (2005), "Algunas características lingüísticas de las denominaciones de una terminología", en *Lingüística Mexicana*, 2 (1), pp. 141-152.

—— (2009), "El descriptor y el término. Los conceptos y la lingüística", en Catalina Naumis Peña (coord.), *Memoria del I Simposio Internacional sobre Organización del Conocimiento*: *Bibliotecología y Terminología*, México, Universidad Nacional Autónoma de México, Centro Universitario de Investigaciones Bibliotecológicas, pp. 53-62.

CEMC (*Corpus del Español Mexicano Contemporáneo, 1921-1974*) (1975), María Isabel García Hidalgo, Luis Fernando Lara, Roberto Ham Chande *et al*., México, *Diccionario del español de México.* [Documento inédito.]

—— (*Corpus del Español Mexicano Contemporáneo, 1921-1974. Lematizado)* (2005), versión elaborada por Gilberto Anguiano Peña, Francisco Segovia y Erika Flores García, México, El Colegio de México, Universidad Nacional Autónoma de México, Instituto de Ingeniería. Disponible en: www.corpus.UNAM.mx/cemc/

Cunha, Iria de (2004), "Análisis discursivo, textos especializados y traducción", en Marisela Colín (ed.), *Manual de traducción de textos especializados. Nuevos enfoques, nuevas metodologías*, México, Universidad Nacional Autónoma de México, pp. 32-45

DEEM (Diccionario estadístico del español de México. Lematizado) (2005), Gilberto Anguiano Peña, Francisco Segovia y Erika Flores (eds.), México, El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Documento inédito.]

DEM (Diccionario del español de México) (2012), México, El Colegio de México, Centro de Estudios Lingüísticos y literarios. Disponible en: http://dem.colmex.mx/moduls/Default.aspx?id=8

DMLE (Diccionario Manual de la Lengua Española Vox) (2007), Larousse Editorial. Disponible en: http://es.thefreedictionary.com/tecnicismo

DTCE (*Diccionario de términos clave de ELE)* (2014), España, Biblioteca Virtual Cervantes. Disponible en: http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/sociolinguistica.htm

DRAE (Diccionario de la lengua española) (2001), 22a. ed., España, Real Academia Española, Espasa Calpe.

Estopà, Rosa (1998), "El léxico especializado en los diccionarios de lengua general: las marcas temáticas", en *Revista Española de Lingüística*, 28 (2), pp. 359-387.

Faber Benítez, P.; Moreno Ortiz, A. y Pérez Hernández, C. (1998), *Lexicografía Computacional y Lexicografía de Corpus.* Disponible en: http://www.ontoterm.com

Gómez González-Jover, Adelina (2005), *Terminografía, lenguajes profesionales y mediación interlingüística. Aplicación metodológica al léxico especializado del sector industrial del calzado y de las industrias afines,* tesis de doctorado, España, Universidad de Alicante, Facultad de Filosofía y Letras, Departamento de Filología Inglesa. Disponible en: http://rua.ua.es/dspace/bitstream/10045/760/1/tesis_doltoral_adelina_gomez.pdf

Halliday, M. A. K. (1979), *El lenguaje como semiótica social*, México, Fondo de Cultura Económica.

Jiménez del Castillo, Juan (2005), "Redefinición del analfabetismo: El analfabetismo funcional", en *Revista Educación*, 338, pp. 273-294. Disponible en: http://www.revistaeducacion.mec.es/re338/re338_17.pdf

Medina Urrea, Alfonso y Méndez Cruz, Carlos (2006), "Arquitectura del corpus histórico del español de México (CHEM)", en A. Hernández y José Luis Zechinelli Martini (eds.), *Avances en la ciencia de la computación,* México, Sociedad Mexicana de Ciencias de la Computación, pp. 248-253.

Lázaro Hernández, Jorge Adrián (2010), *Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico*, tesis de Licenciatura, México, Universidad Nacional Autónoma de México. [Documento inédito.]

Lara, Luis Fernando (1977), "Una base semántica para la lexicografía: la conceptualización del signo lingüístico", en *Nueva Revista de Filología Hispánica*, 26 (2), pp. 261-275.

—— (1984), "Una caracterización lingüística del discurso científico mexicano", en *Discurso: Cuadernos de Teoría y Análisis*, 2, pp. 33-42.

—— (1996), "Conocimiento y pragmática en los fundamentos de la semántica", en *Estudios de Lingüística Aplicada*, 23-24, pp. 236-243.

—— (1999), "Término y cultura: hacia una teoría del signo especializado", en María Teresa Cabré (ed.), *Terminología y modelos culturales*, Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, pp. 39-60.

—— (2001), *Ensayos de teoría semántica: lengua natural y lenguajes científicos,* México, El Colegio de México.

—— (2007), *Resultados numéricos del vocabulario fundamental del español de México,* México, El Colegio de México. Disponible en: http://dem.colmex.mx/moduls/Default.aspx?id=14

—— y Ham Chande, Roberto (1979), "Base estadística del Diccionario del Español de México", en Luis Fernando Lara, Roberto Ham Chande y María Isabel García Hidalgo, *Investigaciones lingüísticas en lexicografía* , México, El Colegio de México, pp. 7-39.

—— y Zahn, Jetta (1973), "El tecnicismo en el léxico del español mexicano. Posiciones posibles del DEM", en *Monografías generales del DEM,* México, El Diccionario del Español de México.

López-Barajas, Emilio (2009), "Alfabetización virtual y gestión del conocimiento", en *Revista Electrónica Teoría de la Educación. Educación y Cultura en la Sociedad de la Información*, 10 (2). Disponible en: http://www.usal.es/teoriaeducacion

López Morales, Humberto (2013), "¿Qué es la disponibilidad léxica?", en *DispoLex: investigación léxica*. Disponible en: http://www.dispolex.com/info/la-disponibilidad-lexica

Marinkovich, Juana (2008), "Palabra y término: ¿Diferenciación o complementación?", en *Revista Signos,* 41 (67). Disponible en: http://www.scielo.cl/scielo.php?pid=S0718-09342008000200002&script=sci_arttext

Moreno, María José (2011), *El lenguaje secreto de la ciencia.* Disponible en: http://ababol.laverdad.es/ciencia-y-salud/2985-el-lenguaje-secreto-de-la-ciencia

Naumis Peña, Catalina (1997), "Reconocimiento semi-automático de patrones temáticos y adaptación del lenguaje documental para mejorar la eficiencia en la recuperación del sistema INFOBILA", en *Primer Congreso Interno de la Comunidad Científica del CUIB: los investigadores y sus investigaciones* México, UNAM, CUIB, pp. 23-27.

—— (1999), *Tesauro latinoamericano en ciencia bibliotecológica y de la información. TELACIBIN,* México, UNAM, CUIB.

—— (2000), "Análisis de la confluencia entre término y descriptor en la elaboración de tesauros", en *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*, 14 (29), pp. 95-113.

—— (2003), "Indización y clasificación: un problema conceptual y terminológico" (Indexation and classification: a conceptual and terminologic problem), en *Documentación de las ciencias de la información*, 26, pp. 23-40. Disponible en: http://www.ucm.es/BUCM/revistas/inf/02104210/articulos/DCIN0303110023A.PDF

Rey-Debove, Josette (1971), *Étude linguistique et sémiotique des dictionnaires français contemporains,* París, Mouton the Hague.

Rubio Liniers, María Cruz (2004), "El análisis documental: indización y resumen en bases de datos especializadas", en *E-LIS: E-prints in Library and Information Science.* Disponible en: http://www.iberius.org/es/AisManager?Action=ViewDoc&Location=getdocs:///DocMapCSDOCS.dPortal/2519

Sager, Juan C. (1993), *Curso práctico sobre el procesamiento de la terminología,* Madrid, Fundación Germán Sánchez Ruipérez, Ediciones Pirámide.

Sánchez González, Aránzazu (2010), *El proceso de comunicación*. Disponible en: http://zazu897.blogspot.com/2010/10/el-proceso-de-comunicacion.html

Temmerman, Rita (2000), *Towards New Ways of Terminology Description: The Sociocognitive Approach,* Amsterdam/Philadelphia, John Benjamins.

Van Dijk, Teun A. (1992), *La ciencia del texto: un enfoque interdisciplinario*, 3a. ed., Barcelona, Paidós.

Wüster, Eugen (2003) [1998], *Introducción a la teoría general de la terminología y a la lexicografía terminológica,* María Teresa Cabré (ed.), España, Barcelona, Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Zipf, George Kingsley (1949), *Human behavior and the principle of least effort,* Oxford, Inglaterra, Addison-Wesley Press.

Para citar este artículo como revista electrónica:

Anguiano Peña, Gilberto y Catalina Naumis Peña. 2015. "Extracción de candidatos a términos de un corpus de la lengua general". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. 67: 19-45. Aquí se agrega la dirección electrónica (Consultado el día-mes-año)

Para citar este artículo tomado de un servicio de información:

Anguiano Peña, Gilberto y Catalina Naumis Peña. 2015. "Extracción de candidatos a términos de un corpus de la lengua general". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. 67: 19-45. En: Aquí se agrega el nombre del servicio de información y la dirección electrónica (Consultado el día-mes-año)