U N I V E R S I D A D N A C I O N A L
A U T Ó N O M A D E M É X I C O
POSGRADO EN BIBLIOTECOLOGÍA
Y ESTUDIOS DE LA INFORMACIÓN

INDIZACIÓN SEMIAUTOMÁTICA PARA ALMACENAR Y RECUPERAR INFORMACIÓN DEL LÉXICO DEL ESPAÑOL USADO EN MÉXICO

T E S I S

QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN BIBLIOTECOLOGÍA

P R E S E N T A

GILBERTO ANGUIANO PEÑA

DIRECTORA DE TESIS

DRA. CATALINA NAUMIS PEÑA

México, D. F., 2007

Con amor a mis padres:

Carmen Peña Báez y Jesús Anguiano Valdez

Con fraternidad y admiración a mis hermanos:

Jorge, Eduardo, Jesús, Griselda, Carina y Lilia

Con cariño y gratitud a:

Magdalena y Jorge

Con aprecio a:

La familia

AGRADECIMIENTOS

Doy gracias a Dios por todo lo que me ha dado en la vida y por que me ha permitido llegar a terminar esta investigación.

Mi agradecimiento a la Dra. Catalina Naumis Peña, Tutora de esta tesis, por su dirección académica y por la generosidad, comprensión y estímulo para mi persona, definitivos para lograr este trabajo de investigación.

Agradezco a cada uno de los integrantes del sínodo por su valiosa contribución académica al presente trabajo: Dr. Luis Fernando Lara, Dra. Ana María Cardero, Dra. Catalina Naumis Peña, Dr. Juan Voutssás Márquez y al Dr. Roberto Garduño Vera.

Con esta investigación agradezco a todas las personas que han apoyado de una u otra forma al Proyecto del Diccionario del Español de México (DEM) y particularmente a quienes han formado parte del equipo lexicográfico, ya que con su quehacer han enriquecido este invaluable tesoro cultural mexicano, y fue también gracias a sus puntuales aportaciones, las que indicaron el camino a seguir, que logré obtener los resultados presentados someramente en esta tesis.

Gracias a Erika Flores García por su apoyo personal y por su importante aportación al Diccionario del Español de México.

Finalmente mi agradecimiento infinito a mis seres queridos por darme el tiempo que les pertenecía y que usé para la elaboración de este trabajo, espero que me sepan perdonar por tanto abuso y las ausencias; ojalá les guste el resultado final de este esfuerzo compartido.

GILBERTO ANGUIANO PEÑA

México, 2007.

ÍNDICE

INTRODUCCIÓN

El problema a resolver en esta tesis es que el sistema perteneciente al Diccionario del Español de México (DEM) cuente con indización por asignación para agrupar palabras en los contenidos documentales y así disponga de registros bibliográficos y su liga a los documentos completos, en otras palabras, en la investigación que se presenta aquí se propone analizar el problema teórico que supone la indización y recuperación de información en el sistema de información del Diccionario del Español de México en función de la detección de necesidades de sus usuarios y la implementación de una solución práctica para satisfacerlas.

JUSTIFICACIÓN

La investigación sobre la indización del léxico del español usado en México surge de un aspecto cotidiano de trabajo documental, para responder a preguntas que se le hacían al sistema de información del Diccionario del Español de México (DEM). El sistema en su diseño original satisfizo las necesidades de información internas de este proyecto lexicográfico; pero al intentar abrir el acceso del sistema a los usuarios externos, resulta evidente que el mismo sistema en su estado inicial no responde a las necesidades de información manifestadas por los usuarios externos hoy en día, pues las respuestas a las solicitudes resultan parciales y para ser completas exigen un proceso manual tardado y difícil. Al revisar la base de usuarios del DEM con mirada crítica y confrontar las necesidades de estos usuarios con el tipo de respuestas que permitía dar el sistema automatizado, se identificó la existencia de una serie de problemas en aspectos de representación y control de la información. A partir de la búsqueda de una solución para controlar y recuperar la información del DEM se plantea en esta tesis un análisis del problema teórico que supone un sistema de información sobre un diccionario y la propuesta práctica de la indización semiautomática para almacenar y recuperar información del DEM.

Se pretende que con el resultado obtenido se abra el acceso a la riqueza de la información del corpus, para ser consultada por todo tipo de usuarios, tanto internos como externos.

Como consecuencia lógica de lo anterior surge la obligatoriedad de hacer los ajustes necesarios para mejorar la eficacia de este mismo recurso informativo en

su conjunto, es decir concluir las etapas de su proceso documental, pues en las condiciones que presenta el sistema desde 1990 si bien sí se recupera la información por medio de las grafías tal y como aparecen en los textos, mediante de la indización por extracción, con esto se logra un acceso parcial a los datos, lo que dificulta la recuperación automatizada del total de las variantes pertenecientes a una misma palabra, así como la identificación de su origen documental.

ANTECEDENTES

Si bien en el presente estudio se hace referencia al español usado en México, en realidad éste no es el tema central de la investigación en este trabajo, por lo cual se les recomienda a los interesados exclusivamente en este fascinante campo de conocimiento que consulten mejor las obras escritas *ex profeso* en la literatura correspondiente,[1] pues aquí sólo se hablará de este tema desde el punto de vista de su documentación.

Ahora bien, entrando en la materia de interés, se puede anotar que en la República Mexicana existen varios proyectos que tienen que ver con el manejo y gestión de información sobre palabras de la lengua natural, como son: de Concepción Company y Chantal Melis el corpus del *Léxico histórico del español de México* dedicado al periodo de la Nueva España; de Gerardo Sierra y Alfonso Medina el *Corpus lingüístico en Ingeniería*; de Raúl Ávila el corpus que ha servido a sus estudios sociolingüísticos; también son importantes los recursos del Centro de Lingüística Hispánica de la UNAM que han servido para elaborar *El habla de la Ciudad de México. Materiales para su estudio*; de Lope Blanch (director) el *Atlas*

[1] Entre otros destacados autores se pueden consultar: Beatriz Arias Álvarez, *El español de México en el siglo XVI: estudio filológico de quince documentos*, México, UNAM, Instituto de Investigaciones Filológicas, 1997, 521 p. (Publicaciones del Centro de Lingüística Hispánica; 44); Elisabeth Beniers Jacobs, "Bibliografía de trabajos descriptivos del español de México, publicados después de 1967", *Anuario de Letras*, México, 1996, núm. 34, pp. 293-349; Concepción María del Pilar Company Company y Chantal Melis, *Léxico histórico del español de México (régimen, clases funcionales, variación gráfica y frecuencias)*, México, UNAM, Instituto de Investigaciones Filológicas, 2002, 952 p. [123 páginas de índices]; Pedro Henríquez Ureña, "Observaciones sobre el español de México", en *Investigaciones Lingüísticas*, 1934, núm. 2, pp. 188-194; Luis Fernando Lara, "El español de México y de América Central", en *Lexikon der romanistischen Linguistik*, ed. por M. Metzeltin, C. Schmitt y G. Holtus, Tübingen, Max Niemeyer, 1989, t. 6, pp. 559-567; Luis Fernando Lara, "Pero... ¿qué es el español de México?", en *El Nacional*, Suplemento Cultural, 1983, (23 de enero); y Juan M. Lope Blanch, "Caracterización del español de México", en *Ensayos sobre el español de América*, UNAM, México, 1993, pp. 119-136.

lingüístico de México, y también de María del Rosario Heras Poncela (responsable) la *Investigación sobre el habla culta de la zona metropolitana de Guadalajara*, etcétera.

Sin embargo, en México solo existen dos instituciones que ofrecen un perfil de información documental similar a lo que se estudia en esta investigación; una de ellas es la Academia Mexicana de la Lengua, correspondiente de la española, con su *Índice de mexicanismos* y la otra institución es el Diccionario del Español de México perteneciente a El Colegio de México, con su *Corpus del español mexicano contemporáneo, 1921-1974*.

En el caso del *Índice de mexicanismos*, después de un breve análisis se puede notar que su carácter es esencialmente referencial, por lo que se entiende que con este tipo de instrumento no se pueden satisfacer necesidades de usuarios que busquen recuperar información de texto completo.

Respecto al Diccionario del Español de México (DEM)[2] su objetivo ha sido y es elaborar un diccionario que incluya el léxico hablado en las fronteras políticas mexicanas, razón por la cual los encargados de efectuar este proyecto, con la intención de conseguir sus objetivos y hacerlo de una manera académica tradicional,[3] tomaron la decisión de generar ellos mismos una fuente primaria de

---

[2] Véase al respecto Luis Fernando Lara, "Sobre la justificación de un diccionario de lengua española hablada en México", *La Gaceta del Fondo de Cultura Económica*, 1972, núm. 19, pp. 1-6.

[3] Como ha ocurrido en otros proyectos anteriores y posteriores al mexicano, tales como: Fernando Justicia Justicia, *El desarrollo del vocabulario: diccionario de frecuencias,* Granada, España, Universidad de Granada, 1995, 306 p.; Amparo Morales, *Léxico básico del español de Puerto Rico,* Puerto Rico, Academia Puertorriqueña de la Lengua Española, 1986, 349 p.; Manuel Ollero Toribio y Miguel Ángel Pineda Pérez, *Diccionario estadístico del léxico popular sevillano,* Sevilla, España, Publicaciones de la Universidad de Sevilla, 1992, 729 p.; José Ramón Alameda y Fernando Cuetos, *Diccionario de frecuencias de las unidades lingüísticas del castellano,* España, Universidad de Oviedo, 1995, 965 p. en 2 vols.; U. Bortolini, C. Tagliavini y A. Zampolli, *Lessico di frequenza della lingua italiana contemporanea,* Milán, 1972; Alphonse Juilland y Emilio Chang

información, éste resultó ser un corpus[4] de texto completo y marcado gramaticalmente, el *Corpus del español mexicano contemporáneo, 1921-1974* (CEMC),[5] el cual concluyeron en 1975. Con el mismo pretendían obtener una muestra representativa del léxico del español usado en México. Se procedió de esta manera porque se necesitaba información basada en datos reales y fidedignos en los que se pudiera tener certeza del uso de las palabras en los textos generados en México. Una vez obtenido el corpus y conformado como su sistema de información automatizado, iniciaron su trabajo lexicográfico, o sea, la elaboración de un diccionario mexicano.

En relación con los estudios sobre los *corpora*, en general se les atribuye a éstos gran importancia como fuentes de las que se pueden extraer datos de manera sistemática, y de su posible transformación en sistemas de

---

Rodríguez, *Frequency Dictionary of Spanish Words,* La Haya, Mouton & Co., 1964; P. Imbs, *Trésor de la langue francaise: dictionnaire de la langue du XIXe et du XXe siécle (1789-1960),* París, Centre National de la Recherche Scientifique, [1971-].

[4] Ejemplos de los estudios con los corpora son: Aquilino Sánchez *et al*., *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones,* Madrid, Sociedad General Española de Librería, 1995; E. Camarero García y M. F. Verdejo, "Un sistema pregunta-respuesta en castellano, sobre un corpus literario", *Boletín del Centro de Cálculo de la Universidad Complutense*, 1978, núm. 32, mayo, pp. 4-12; G. Corpas Pastor, "Localización de recursos y compilación de corpus vía Internet: aplicaciones para la didáctica de la traducción médica especializada", en García Yebra, V. Gonzalo García, C. (eds.), *Manual de documentación y terminología para la traducción especializada*, Madrid, Arco/Libros, pp. 223-506 (Colección Instrumenta Bibliologica)*;* M. Chantal Pérez Hernández, "Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento", *Estudios de Lingüística Española,* 2002, vol. 18. http://elies.rediris.es/elies18/index.html; Manuel Alvar Ezquerra y Juan Andrés Villena Ponsoda, *Estudios para un Corpus del Español*, Málaga, Universidad de Málaga, 1994, pp. 31-40.

[5] El Colegio de México. Diccionario del Español de México, *Corpus del español mexicano contemporáneo, 1921-1974* [cinta magnética], elaborado por García Hidalgo, María Isabel, Luis Fernando Lara, Roberto Ham Chande *et al*., México, Diccionario del Español de México, 1975.

almacenamiento y recuperación de información, como ocurre en este trabajo con el *corpus* del DEM, en realidad los corpora no son estudios nuevos.[6]

En esta tesis entre otras cosas se destaca la importancia de añadir valor al contenido del CEMC utilizando la indización por asignación para recuperar adecuadamente la información solicitada por los usuarios del DEM y para completar el proceso documental. Esto se hace, al mismo tiempo que se proponen los ajustes necesarios para que este recurso académico sea aprovechado, no sólo en la elaboración del diccionario mexicano, sino también para hacerlo accesible a los usuarios mexicanos e internacionales interesados en el español que se habla en México.

Por esta razón es pertinente explicar lo que es un *corpus lingüístico* y el significado de la *lingüística de corpus*, ante la muy viable transformación de un corpus hacia un sistema de almacenamiento y recuperación de información. Veamos entonces las siguientes definiciones del *Diccionario de lingüística moderna*:

> "**Corpus.** Los datos que utilizan en el trabajo de investigación lingüística constituyen el "corpus" o **inventario** del trabajo, los cuales son los enunciados de la comunicación, que están compuestos por formas lingüísticas, como oraciones, palabras y similares y también por categorías como sílabas, vocales, oclusivas, etc.

[6] Los hay de enfoques generales como los hechos por la Real Academia Española en su página de Internet con el *Corpus de Referencia del Español* (CREA) y con el CORDE; o bien en aspectos especializados como el de M. Chantal Pérez Hernández, "Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento", *Estudios de Lingüística Española*, 2002, vol. 18, http://elies.rediris.es/elies18/index.html; y en *Documentación y construcción de un corpus digital monolingüe sobre astronomía: criterios de clasificación para su aplicación a la investigación y a la docencia de la traducción especializada*, http://www.fti.uab.es/psanchez-gijon/Recerca/Treballs_20de_20recerca/Trecerca99.htm.

[…] por definición, es un **repertorio lingüístico** cerrado […] Los datos de este inventario o repertorio lingüístico deben ser: (a) válidos, esto es, deben ser representativos de la hipótesis que se quiere demostrar; (b) comprobables, (c) homogéneos […]".[7]

A lo anterior, conviene abundar con otra definición que dice: "Un *corpus lingüístico* es un conjunto, normalmente muy amplio, de ejemplos reales de uso de una lengua. Estos ejemplos pueden ser textos (típicamente), o muestras orales (normalmente transcritas)."[8]

Respecto a la definición de la Lingüística de corpus se afirma lo siguiente:

"**Lingüística de corpus.** Se llama 'lingüística de corpus' a la rama de la lingüística que estudia las lenguas basándose en los grandes repertorios lingüísticos llamados *corpora*. El nombre de CORPUS se aplica a toda colección de textos compilados según unos criterios explícitos que hacen que sea suficientemente representativo como para constituir, en sí mismo, un MODELO a escala de los aspectos lingüísticos que el investigador quiere examinar (Atkins, S. *et al*., 1992). Hay diferentes maneras de clasificar un corpus, ya sea por su carácter oral o escrito, por representar el uso general de la lengua o el de un dominio específico, o por su carácter monolingüe, bilingüe o multilingüe (*cf.* corpus). […] los orígenes, sin embargo, se remontan a la lingüística estructural posterior a Bloomfield, cuando los textos (escritos y hablados) se consideraban la fuente primaria y única para la investigación lingüística. Hay, sin

---

[7] Enrique Alcaraz Varó y María Antonia Martínez Linares, *Diccionario de lingüística moderna*, Barcelona, Editorial Ariel, 1997, p. 151. [Las negritas son mías].
[8] Defición que aparece en *Wikipedia. La enciclopedia libre*. http://es.wikipedia.org/wiki/Corpus_ling%C3%BC%C3%ADstico. [Las negritas son mías].

embargo, una diferencia entre la 'lingüística de corpus' de esta primera época y la concepción actual: se trata del uso de los ordenadores para almacenar y procesar grandes cantidades de información, lo cual supone facilitar la gestión de los datos. Tal como señala G. Leech (1991), esta innovación está relacionada con tres aspectos: (1) la importancia del corpus como una fuente de donde extraer datos de manera sistemática; (2) la importancia del corpus como banco de pruebas para el análisis lingüístico; y (3) la importancia del corpus como metodología del trabajo para la creación de sistemas para el procesamiento informático de la lengua."[9]

Bajo estas definiciones, principalmente la definición del *Diccionario de lingüística moderna*, donde señala "la importancia del corpus como una fuente de donde extraer datos de manera sistemática", es el criterio de partida para justificar el planteamiento formulado en esta tesis.

Pues bien, en el Área de documentación del Diccionario del Español de México (DEM) se ofrece una serie de servicios, entre los que destaca el proporcionado a partir de un sistema automatizado, al que le denominan CEMC.[10] De este servicio se conserva un registro de usuarios con las preguntas más frecuentes, esta información fue la que sirvió de base para diseñar un sistema de almacenamiento y recuperación de información retomando como recurso principal los datos que ya ofrecía ese mismo sistema automatizado del DEM, pero agregándole los elementos definidos en el análisis de las necesidades de los usuarios. Como se trató de responder a preguntas que tienen que ver con el significado, se buscó hacerlo en una forma semiautomática porque sólo así, puede

[9] Enrique Alcaraz Varó y María Antonia Martínez Linares, *Diccionario…, op. cit*., p. 337.
[10] Siglas del *Corpus del español mexicano contemporáneo*, 1921-1974.

lograrse la interpretación humana. En otras palabras, con esta acción se consideró oportuno reutilizar un sistema automatizado que ya existía, el cual en su momento sirvió de base para la elaboración interna del DEM, completando las etapas del proceso o cadena documental faltantes en el mismo, con el objeto de conseguir un sistema de recuperación de información (SRI) que responda a las actuales necesidades de información de los usuarios, principalmente externos, las cuales no fueron consideradas cuando este sistema se creó.

El sistema automatizado del CEMC, constaba de cuatro componentes originales: 1) una bibliografía (en papel), 2) una base de textos o ítems (los textos del corpus), 3) un índice estadístico de palabras obtenidas automáticamente por extracción, llamado *Diccionario estadístico del español de México*, y 4) un índice de concordancias, muy parecido al índice KWIC que se utilizaba para recuperar los textos o ítems del sistema. Estos fueron resultado de la indización automática y la recuperación de su información la basan en las palabras como aparecen en los textos obtenidas por medio de la indización por extracción.

Ahora bien, con el propósito de dar respuesta a las necesidades de información manifestadas por los usuarios del sistema, y teniendo en cuenta el estado que guardaba el mismo al inicio de esta investigación, se tomó la decisión de transformar a este sistema automatizado en un sistema de recuperación de información, basado en la indización semiautomática, con el que se resolvieran los problemas propios de la indización por extracción en cuanto a la recuperación de información se refiere. Esto se pensó hacer en dos pasos; el primer paso consistió en que cada término de indización obtenido por la indización por extracción fuera validado por un ser humano, siguiendo para ello ciertos lineamientos internos, con

lo cual se podría ofrecer a los usuarios internos y externos, un índice simplificado y controlado que sirviese de guía en las búsquedas de las palabras contenidas en el sistema.

El segundo paso, consistió en añadir valores al contenido de dos bases del CEMC, esto es, a la del índice estadístico y a la de los textos. Los contenidos que se eligieron ser incorporados a cada ítem fueron sus registros bibliográficos y su correspondiente clasificación temática o estratificación sociolingüística, pues con estas acciones se completaría en dichas bases su información bibliográfica y temática, lo que evitaría el que se tenga que consultar otras fuentes para interpretar los datos recuperados.

OBJETIVO

El objetivo de esta investigación es indizar en forma semiautomática puntos de acceso y relaciones en el corpus de tal manera que complementen el sistema del Diccionario del Español de México (DEM) con un subsistema de almacenamiento y recuperación de información para usuarios que responda a las necesidades de información detectadas. Esta información será extraída del *Corpus del español mexicano contemporáneo* (CEMC), el cual es una muestra aleatoria cerrada que contiene los datos del léxico del español de México y que es el sistema de información automatizado en el que se basa el DEM. El CEMC y como consecuencia el DEM están sustentados con un sistema adecuado a las necesidades que dieron origen a su construcción, pero no responden a las necesidades de consulta de todos los usuarios actuales del sistema.

Los objetivos específicos son:

1. Documentar e indizar un subsistema de información complementario del léxico del español usado en México con base en la detección de necesidades de sus usuarios.
2. Usar el CEMC como base de trabajo para documentar el léxico del español usado en México porque tiene la información en la que se basa el DEM.
3. Indizar en forma semiautomática puntos de acceso y relaciones para conformar un subsistema de almacenamiento y recuperación de información en el sistema del DEM.
4. Lematizar las palabras, es decir agrupar e indizar por asignación humana, las palabras obtenidas del análisis automático, junto a todas sus flexiones, validando de esta manera, cada una de ellas mediante el análisis de contenido en el corpus.
5. Completar los registros bibliográficos de los textos citados en el CEMC.
6. Agregar información sociolingüística a los textos mediante elementos clasificatorios que los definan en el sistema de información.
7. Obtener un sistema que responda a las preguntas frecuentes que no era posible responder con el CEMC y contar con una herramienta de búsqueda que guíe al usuario.

HIPÓTESIS

Se parte del supuesto que los usuarios tienen necesidades de información relacionadas con problemas de significado que no resuelve el sistema actual de información asociado al DEM.

Otro supuesto es que la indización por asignación o lematización (es decir agrupar por las flexiones correspondientes de una forma canónica o vocablo) aunada a la complementación de los registros bibliográficos en los textos usados para obtener las formas canónicas, y la incorporación de los elementos clasificatorios sociolingüísticos en los mismos textos es una necesidad frecuente de los usuarios del sistema automatizado del DEM o el CEMC, sistema que lo retroalimenta.

METODOLOGÍA

La metodología parte del estudio de caso de los registros de consulta realizados por los usuarios del DEM y especialmente del CEMC, para observar, definir y obtener los elementos complementarios que necesita el sistema actual. Es decir se analizan los resultados que ofrece el sistema anterior en uso al manipular la información para responder a preguntas concretas de los usuarios.

El proceso documental original es suficiente para apoyar con éxito las labores internas de documentación en el Diccionario del Español de México (DEM), sin embargo, para poder tener la información lexicográfica completamente controlada y automatizada, incluso para poder ofrecerla en consulta a usuarios internos y externos en un sistema de almacenamiento y recuperación de información, es necesario concluir el proceso documental, y para ello se procede a añadirle valor al contenido de los archivos que se encontraban estáticos desde 1990 en el sistema computacional del DEM o sea el CEMC.

Para el estudio de la complementación o añadirle valor al contenido del sistema automatizado no se consideró al índice de concordancias, por lo cual sólo se efectuó en tres de los elementos del sistema que sustentan los datos para conformar el DEM, estos son: a) *Bibliografía del CEMC*, con soporte en papel, convertida a base de datos para este estudio; b) La base de los textos del CEMC, con la idea de armonizar, complementar y relacionar los datos de tal manera que se pueda contar con un verdadero sistema de almacenamiento y recuperación de información y c) El *Diccionario estadístico del español de México*. Después de

analizar estos componentes y definir cómo abordar el problema se resuelve para cada una de ellas las acciones que se explican a continuación:

1) Al *Diccionario estadístico del español de México*, se le migra del programa INFORMIX a una base en Excel, en la que se pueden observar y trabajar las 64194 palabras gráficas, obtenidas por extracción. Posteriormente se procede a indizar por asignación o lematizar (es decir agrupar por las flexiones correspondientes de una forma canónica[1], por entrada o vocablo), a las palabras obtenidas por indización automática (por extracción) de la base de datos del CEMC. La indización por asignación manual permite identificar la forma canónica correspondiente a cada palabra gráfica. Esto sirve para completar los registros del Índice estadístico e identificar y controlar las palabras del lenguaje natural con que posteriormente se podrá recuperar la información buscada en la base de textos completos.

Una vez obtenidas las formas canónicas o vocablos se le añade valor al contenido del registro documental de cada una de ellas, mediante la identificación y disposición en la base de datos de columnas que las identifican en cuanto a su origen documental, con etiquetas como: tipo de lengua, nivel de lengua, géneros

[1] El significado de forma canónica, lo explica el *Diccionario de lingüística moderna*, de la manera siguiente: "LEMA, LEMARIO. En la teoría lexicográfica, el 'lema' es la forma canónica a la que se reduce todo un paradigma flexivo y que se forma como representante de todas las variantes morfológicas de la palabra. (Alvar: 1993a). El 'lema', también denominado entrada, palabra clave o voz guía, es el elemento que encabeza los artículos de los diccionarios de la lengua. El conjunto de las entradas o 'lemas' de que consta un diccionario se denomina macroestructura, nomenclatura y, más recientemente, lemario. El proceso de reducción morfológica que exige la lematización estricta presupone, en el usuario que consulta la obra lexicográfica, un cierto conocimiento de la gramática de su lengua (*cf* Azorín y Martínez: 1994: 95), a veces sobrevalorado por el lexicógrafo, lo que puede constituir un serio obstáculo de cara a la identificación de la entrada en la nomenclatura del diccionario por parte del lector, de ahí que, en ocasiones, la lematización se quiebre a favor de la facilidad del manejo del diccionario." Alcaraz Varó, Enrique y María Antonia Martínez Linares, *Diccionario de lingüística moderna*, Barcelona, Editorial Ariel, 1997, 643 p.

de lengua y tipo de texto; también se marcan aspectos como mayor frecuencia, mejor distribución y cuáles son las palabras significativas contenidas exclusivamente en textos especializados. Lo que permitirá además, clasificar por su uso a las palabras del léxico del español de México.

2) A la *Bibliografía del CEMC*, de un formato impreso, se le transforma en una base de datos. Esta bibliografía está basada en la indización precoordinada, guiada con un criterio de estratificación sociolingüístico o clasificación temática. En la base de datos a los 996 códigos de texto se les incorporaron 1932 registros bibliográficos correspondientes a los textos propiamente analizados y éstos a su vez se identificaron con 87 áreas del conocimiento. Posteriormente, a cada uno de los registros se le agregó de forma explicita la información sobre su estratificación sociolingüística o clasificación temática; consiguiéndose de esta manera una base de datos que relaciona los diferentes puntos de acceso de los contenidos documentales.

3) A la base de los textos completos del CEMC se le migra del programa INFORMIX al de Excel para manipularla y realizar los complementos de contenido en los aspectos bibliográficos y temáticos, tomados de la base de la bibliografía, para así establecer las relaciones que se requieran después de ser detectadas las necesidades de los usuarios.

Finalmente se entrelazan dichas acciones para tratar de obtener el subsistema de información integrado al sistema documental del léxico del español de México en el CEMC en cuanto a su estructura y el comportamiento de la información, y que sea complementado con un programa de ayudas de búsqueda.

Después de explicar aquí las acciones a desarrollar y la unidad que se guarda respecto a su viabilidad, estructura y consecuencia, se explica a continuación la secuencia de este trabajo.

En el primer capítulo se aborda la indización, el papel que juega en el proceso documental y se exponen los conceptos teóricos argumentados en la tesis; se explica cómo se entrelazan la documentación, el proceso o cadena documental, el análisis documental y la indización; además de justificar la asignación de términos de indización como el eje para guiar el trabajo.

En el segundo capítulo que trata la documentación del *Corpus del español mexicano contemporáneo*; se exponen los antecedentes de este servicio documental y se observan las necesidades de información de los usuarios, detectadas en las preguntas de consulta. El análisis de estos dos componentes muestra la necesidad de poner en práctica una indización orientada al usuario.

En el tercer capítulo y dentro de un enfoque integral del proceso documental se plantea una indización orientada al usuario y añadir valor al contenido del CEMC, para luego efectuar el diseño de las búsquedas correspondientes. Además de los complementos al proceso documental del CEMC y el diseño de búsquedas, se detalla la manera en que se desarrolla la indización manual, por medio de la indización asignada sobre las palabras gráficas producidas por la indización automática, recordando que desde un principio se pretendió hacer una indización exhaustiva por medio de la utilización de programas lingüísticos y estadísticos.

Respecto a las palabras gráficas obtenidas de la indización por extracción se constituyeron en la materia prima del trabajo documental y son las sometidas a la indización humana por asignación, con un enfoque lexicográfico de la información,

el de la lematización (asignación de la forma canónica de las palabras), ya que con esta forma económica de las palabras se representan todas las variantes gráficas documentadas de una palabra (de lengua hablada y escrita), tales como la raíz de palabra, las palabras truncadas o las formas flexivas y derivativas de los lexemas, tales como aumentativos, diminutivos, plurales, despectivos, femeninos, etc. La lematización se realiza con la finalidad de obtener como resultado una indización semiautomática.

Posteriormente, en un apartado, se discute lo ocurrido por el desarrollo de la misma y se presentan las consideraciones finales del caso, aquí se expone el estado que guarda ahora el sistema de información después de la aplicación de una indización orientada al usuario, que permitirá recuperar aproximadamente dos millones de palabras de la lengua natural que tienen documentación en el corpus lingüístico estudiado.

Finalmente se exponen los resultados más destacados conseguidos a lo largo de la investigación.

Como colofón se incluye la bibliografía utilizada en el desarrollo de este tema multidisciplinario.

CAPÍTULO I

LA INDIZACIÓN EN EL PROCESO DOCUMENTAL

Puesto que esta tesis se estructura a partir de la solución de un problema de indización es conveniente explicar el título "Indización semiautomática para almacenar y recuperar información del léxico del español usado en México", para entender qué es lo que se pretende exponer, pues de ello se deriva la investigación que sustenta el trabajo y su contenido.

Como primer elemento tenemos que Indización semiautomática se encuentra ubicada en los conocimientos concernientes a la Documentación y al proceso o cadena documental. De la Documentación, Ruiz Pérez define: "[…] es la ciencia general que tiene por objeto el estudio del proceso informativo-documental, en un plano universal y en un plano específico aplicado a una disciplina concreta".[1]

Esto se practica principalmente en centros y sistemas de información y de documentación, al igual que en bibliotecas especializadas, con el objetivo de satisfacer las necesidades de información de sus usuarios y facilitar el flujo de la información especializada, de tal forma que se fortalezca el proceso de comunicación en la comunidad científica donde comúnmente se anidan dichas instituciones informativas, y así también se coadyuva en la producción de nuevos conocimientos.

---

[1] Rafael Ruiz Pérez, *El análisis documental: bases terminológicas, conceptualización y estructura operativa*; presentación José Ramón Pérez Ávarez-Ossorio, Granada, España, Universidad de Granada. Grupo de Trabajo de Información y Documentación de la Comisión Española de Cooperación con la UNESCO, 1992, p. 19.

Pinto Molina añade al respecto "El proceso documental […] necesita de estos tres ejes básicos insustituibles para ser llevado a término: un emisor o *documentalista*, que como sujeto cualificado será el encargado de aplicar las técnicas pertinentes; un mensaje, el documento; y un destinatario o usuario, que será el beneficiario último de dicho proceso".[2]

El ejercicio de la investigación practicado en esta tesis se encuentra inmerso en el análisis documental, que el mismo Ruiz Pérez nos explica como: "conjunto de operaciones necesarias para extraer la información (o lo esencial de la misma), contenida en las fuentes primarias (documentos primarios) o (expresarla en elementos eficaces) para su posterior recuperación y utilización",[3] a lo que Pinto Molina detalla en cuanto a su función "la cantidad de documentos y la diversidad de las preguntas han obligado a introducir una etapa suplementaria, o intermediaria, que facilita la operación de interrogación. Esta fase intermediaria es precisamente la del Análisis Documental".[4]

Se reconocen en general dos niveles del análisis documental, uno que tiene que ver con su forma y otro que tiene que ver con su contenido. En cuanto tratamiento de la forma, se dedica a identificar todos los elementos aparentes y convencionales del soporte, que hacen posible la identificación formal de cada documento.

En cuanto al tratamiento del contenido para obtener el tema de un documento y extraer los elementos que representen los conceptos contenidos en

---

[2] María Pinto Molina, *Análisis documental: fundamentos y procedimientos*, Madrid, Eudema, 1993, p. 42.
[3] Rafael Ruiz Pérez, *El análisis…*, *op. cit*., p. 21.
[4] María Pinto Molina, *Análisis…, op. cit.,* p. 45.

este documento, tenemos que: la clasificación indica el contenido de un documento, incluyéndolo en cierta clase o rama del conocimiento; la indización extrae los conceptos del texto de un documento y a la vez los expresa por medio de la asignación de términos significativos con lo que indica de qué trata un documento; y el resumen que es la representación abreviada del contenido de un documento, donde se señala qué cosa dice el documento.

La indización es una técnica o forma de efectuar el análisis documental en cuanto a su nivel interno y específicamente en cuanto al análisis documental de contenido o mensaje; se realiza con la intención de identificar los contenidos documentales y así hacer posible su recuperación. Para lograr lo anterior, se efectúa la descripción de la información y su caracterización utilizando: palabras significativas, palabras clave, materias, temas, unitérminos, o descriptores representativos del contenido del documento que los contiene.

Si acaso el análisis de contenido de un documento no se articula en un lenguaje previamente establecido se entenderá que se usa la indización en lenguaje libre, denominado así porque la representación de los contenidos se efectúa con términos extraídos del lenguaje natural en que se encuentre el documento, en caso opuesto, se entenderá que se usa el leguaje documental o controlado.

En otras palabras, y por convenir a la justificación de esta tesis, se entenderá que si un sistema se fundamenta en el lenguaje natural, se encuentra ante una aplicación de la indización por lenguaje natural, la cual puede ser indización por

extracción o indización por asignación. Por último, es oportuno considerar que un producto del análisis documental puede ser un índice.[5]

En esta tesis se parte de un concepto amplio de indización el cual comprende la asignación en lenguaje natural de los textos.

Un aspecto que no hay que olvidar como factor que se incluirá en la indización semiautomática es el de la indización humana, la cual está basada en lo general en el concepto de la indización pero, además tiene otras características:

> "Van Slype señala como rasgo distintivo de la indización humana el ser una actividad fundamentada en la apreciación de un ser humano que se ejerce en dos planos: El de las unidades significativas reconocidas: distingue conceptos. El de la selectividad: se reconocen los elementos constitutivos.
>
> La indización humana, al fundamentarse en una apreciación de un ser humano, será proclive a las posibilidades de manipulación, ya que entra en juego la subjetividad del documentalista: los conceptos extraídos del documento pueden no ser los mismos de un documentalista a otro, ni ser compartidos por el usuario, que verá limitada su búsqueda a la conceptualización de otro ser humano."[6]

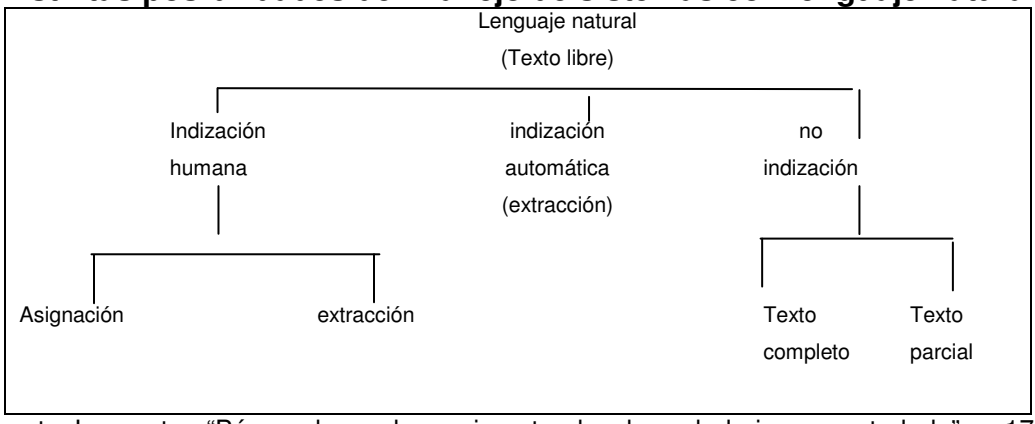
---

[5] El análisis interno según García Gutiérrez "…comprende dos fases: la Descripción Característica o Indización, tanto por métodos tradicionales como automatizados y cuyos productos son los índices y los indizados; y la Descripción Sustancial, esto es la operación de resumir, cuyo producto tradicional es el Resumen", A. L., García Gutiérrez, *El análisis documental*, *apud* Rafael Ruiz Pérez, *El análisis…*, *op. cit.*, p. 76.

[6] Georges van Slype, *Los lenguajes documentales de indización: concepción, construcción y utilización en los sistemas documentales*: traducción del francés Pedro Hípola [y] Félix de Moya. Madrid, Fundación Germán Sánchez Ruipérez, 1992, 198 p., *apud* Inmaculada Chacón Gutiérrez, *Efectos sociales del proceso documental de la fotografía de prensa*, http://www.ucm.es/info/multidoc/multidoc/revista/cuadern3/fotograf.htm.

Lancaster, aclara sobre la indización humana, que un sistema de lenguaje natural[7] puede estar basado en la indización humana, la indización automática o no existir indización.

Distintas posibilidades del manejo de sistemas con lenguaje natural



Fuente: Lancaster, "Búsqueda con lenguaje natural y el vocabulario poscontrolado", p. 178.

El mismo autor menciona que la indización automática o la humana por extracción se basa en extraer palabras o frases del texto de los documentos, pero que también en un sistema de lenguaje natural puede darse la asignación de términos adicionales que no aparezcan en el texto, es decir, que un indizador puede decidir que puede aplicar términos a un determinado texto, y en este caso se estaría hablando de la indización por asignación.

Esto queda sustentado por UNE 50113-5. (Equivale a la norma ISO 5127/3A-1981) donde se habla de la identificación y análisis de documentos y donde se aporta la siguiente definición:

---

[7] Véase en F. W. Lancaster, "Búsqueda con lenguaje natural y el vocabulario poscontrolado", en *El control del vocabulario en la recuperación de información*; tr. Alejandro de la Cueva Martín, València, España, Universitat de València, 1995, 286 pp. 177 y 188; y en F. W. Lancaster, "Indización de documentos científicos", en *Procesamiento de la información científica*, Madrid, Arco Libros, 2001, pp. 164-181.

"**Asignación de términos de indización**: elección y atribución de términos, aparezcan o no éstos en el texto, para representar documentos o datos de acuerdo con ciertas reglas." [8]

Lo anterior es muy necesario tomarlo en cuenta, pues lo que se va a desarrollar en esta tesis es a partir de lo que el humano, un documentalista, pueda asignar a la indización automática, por extracción, efectuada en un sistema automatizado, de tal manera que se complemente y así se logre consolidar lo que ahora se denomina la indización semiautomática,[9] en beneficio de los usuarios del propio sistema.

El siguiente párrafo ilustra cómo conseguir la indización semiautomática:

"La definición de la automatización de la indización se debe acometer desde una triple perspectiva: a) Programas informáticos que asisten en el proceso de almacenamiento de los términos de indización, una vez obtenidos de modo intelectual. (Indización Asistida por Ordenador Durante el Almacenamiento); b) Sistemas que analizan los documentos de modo automático, pero los términos de indización propuestos los valida y edita -si es necesario- un profesional (Indización Semiautomática); y c) Programas sin ningún tipo de validación, es decir, los

---

[8] Biblioteca Computense. Subdireción de Servicios Técnicos y Adquisiciones. Servicio de Proceso Técnico y Normalización, *Evaluación del uso de descriptores en los registros bibliográficos*, Definición en UNE 50113-5. (Equivale a la norma ISO 5127/3A-1981). http://www.ucm.es/bucm/intranet/doc6461.pdf. [Las negritas son mías].

[9] En trabajos como el de Dagobert Soergel, "Automatic and semi-automatic methods as an aid the construction of indexing lenguajes and theasauri", *International Classification*, vol. 1, núm. 1, may 1974, pp. 34-39, o en L. Alfonso Ureña López, *Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos*, España, Editorial Club Universitario, 2002, 156 p.

términos propuestos se almacenan directamente como descriptores de dicho documento. (Indización Automática)."[10]

Características específicas de la asignación de términos de indización para este estudio:

- El tipo de término o forma que se busca fijar en este estudio, es la llamada forma "canónica" de una palabra, con lo que se reducirían a una sola expresión todas las variantes de una palabra o término obtenido con anterioridad por extracción de la indización automática, y esto permitiría obtener los datos completos de las palabras buscadas en el sistema automatizado.
- Para efectuar este tipo de indización en el sistema del CEMC y mantener un nivel apropiado de control semántico se puede utilizar como apoyo las marcas gramaticales que ya existen asignadas en el anterior índice de palabras obtenidas por extracción o sea del *Diccionario estadístico del español de México*.
- En este tipo de indización, casi en su totalidad predomina la asignación con un solo término, sin embargo, cuando se presentan problemas de ambigüedad o de significado, se ha recurrido al uso de términos compuestos o bien de calificadores para diferenciar los distintos

[10] Isidoro Gil Leiva, "Sistema para la Indización Semiautomática (SISA) de artículos de revista de Biblioteconomía y Documentación", en *II Jornadas de Tratamiento y Recuperación de Información*, Madrid, Septiembre 2003, pp. 228-232.
http://personales.upv.es/isgil/SISA%20Demo%20Jotri%202003%20original.pdf.

significados de los homógrafos, y así que esto ayude a responder mejor las consultas de los usuarios.

- Se tomó como base de este tipo de indización la *Nomenclatura* del Diccionario del Español de México y como apoyo la misma las entradas del *Diccionario de la lengua española* de la Real Academia Española para fijar los términos establecidos por éstas, y se utilizaron otras fuentes si fuera necesario, aún siendo externas al sistema.
- Esta asignación produce una forma de control que permite ofrecer a los usuarios la información total correspondiente a una palabra, o todas las palabras solicitadas al sistema, de una manera expedita y sencilla.
- La asignación por razones económicas y de control de la información debe ser lo más exigente posible, pues no se desea incorporar nuevos términos y por esto se debe ceñir a la normatividad correspondiente, para este caso lexicográfica, elegida en el sistema que pretende ofrecer los servicios de información.

En este punto terminan los supuestos teóricos considerados para el desarrollo del proyecto en la tesis.

CAPÍTULO 2

LA DOCUMENTACIÓN Y EL *CORPUS DEL ESPAÑOL MEXICANO CONTEMPORÁNEO*

Respecto al Diccionario del Español de México (DEM) es un proyecto de El Colegio de México que se ubica en el Centro de Estudios Lingüísticos y Literarios (CELL), y en el que desde su origen, se han elaborado una serie de recursos documentales originales, con los que se respaldan sus labores lexicográficas, labores que corresponden a las efectuadas en los proyectos académicos modernos, con la metodología y las bases teóricas pertenecientes a la lingüística aplicada moderna. Esto corresponde al tipo de metodología en donde la información obtenida de fuentes primarias es considerada imprescindible en la elaboración y manipulación de un producto elaborado académicamente.

El proceso documental y el *corpus*[1] como fuente de información

En su tarea de hacer diccionarios, el lexicógrafo tiene necesidad constante de la mayor y mejor cantidad posible de información, que permita inventariar exhaustivamente la lengua, y aunque alguna información la recibe de forma natural en el habla cotidiana, por la radio, el periódico y la televisión u otros medios, no es sino la información documentada la que llega a ser verdaderamente

[1] Confróntense las definiciones en la Introducción de esta investigación, en las que para los estudios lexicográficos el *corpus* es considerado una herramienta para elaboración de diccionarios.

útil en las labores lexicográficas, ya que esta información debe servir como testimonio o prueba del conocimiento claro y seguro del uso de las palabras.

Desde un enfoque de la información, se reconoce como la unidad mínima de información a la "palabra", que es imprescindible para desarrollar las labores lexicográficas, misma que se encuentra tradicionalmente en las fuentes de información primaria, dichas fuentes son principalmente los diccionarios, enciclopedias, nomenclaturas, vocabularios, glosarios, etc., y en los estudios lingüísticos en los *corpora*.

El principal objetivo de la documentación en la lexicografía es asegurar el acceso a las fuentes primarias y secundarias que contienen los fragmentos de información o conocimiento necesarios en la elaboración de diccionarios.

Como consecuencia de esta necesidad documental, en el DEM se creó un recurso sustentado en la lingüística computacional, el análisis de textos, la indización automática y la estadística, este recurso al que en el ámbito de la lexicografía se le denomina *corpus*, se le nombró *Corpus del español mexicano contemporáneo* (1921-1974). Desde luego el objetivo del corpus es que sirviera para analizar y obtener información imparcial y confiable de una muestra representativa de cerca de dos millones de palabras contenidas en 1 932 textos analizados, con lo cual se esperaba obtener el estado del español hablado en México. Con este corpus se puede llegar a obtener el ciclo completo del proceso documental de la información y su recuperación real por los ítems que la contienen.

2.1 LAS NECESIDADES DE LOS USUARIOS DETECTADAS CON EL USO DEL SISTEMA DE INFORMACIÓN[2] DEL DICCIONARIO DEL ESPAÑOL DE MÉXICO

Los servicios ofrecidos a los usuarios del CEMC

Antes de hablar propiamente del sistema automatizado del CEMC, del que abreva el DEM, es importante hablar de la *Bibliografía del CEMC*, pues hay que hacer notar que la bibliografía sirvió primeramente en la selección y adquisición de material, y después en ella se realizó la clasificación sociolingüística del sistema. Finalmente ésta sigue siendo la clave para tener el acceso al corpus, como se describe en el "Acceso a la información" de la "Introducción de la bibliografía":

> "Esta muestra del español usual en México es la que constituye el *Corpus del español mexicano contemporáneo*, cuya bibliografía se ofrece a continuación. […] El objeto de esta bibliografía es ofrecer a los interesados en el uso del CEMC, los datos necesarios para pedir resultados al equipo lexicográfico del DEM."[3]

Ejemplos de la bibliografía de los textos que se encuentran en el CEMC

000. Azuela, Mariano, *Los de abajo*, Botas, México, 1944.
326. Garza Mercado, Ario, *Las bibliotecas de la Universidad de Nuevo León*, Universidad de Nuevo León, Monterrey, México, 1966, Volumen VII.
526. García Trejo, Antonio, *Plaguicidas en el medio ambiente*, IMIQ. Nuevos Procesos, México, 1974. Volumen 15; núm. 2; páginas 84 al la 93.
609. Reyes Heroles, Jesús, *Seis discursos*, Imprenta Madero, México, 1973.
719. Murúa Beltrán, Dámaso, *El Güilo Mentiras*, Impresora Técnica Moderna México, 1971.

[2] El sistema computacional está basado en el siguiente documento El Colegio de México. Diccionario del Español de México, *Corpus del español mexicano contemporáneo, 1921-1974* [cinta magnética], elaborado por García Hidalgo, María Isabel, Luis Fernando Lara, Roberto Ham Chande *et al*., México, Diccionario del Español de México, 1975.
[3] Aquí queda clara la orientación de ofrecer servicios al público desde sus inicios, véase El Colegio de México. Diccionario del Español de México, *Bibliografía del CEMC*, [h. 5].

767. *El llorar*, *(versión 59* Chalahuite, Hgo., 1967. Cintas Colegio.
958. Agustín, José, *Círculo vicioso*, Mortiz, México, 1974.

A esta aclaración se le puede añadir otro párrafo de la misma "Introducción":

> "En la bibliografía, el número que se encuentra a la izquierda de la ficha representa el código del texto dentro del corpus. Toda petición de resultados debe hacer referencia a él. Cada línea de texto está precedida por nueve dígitos: los tres primeros representan el número del texto, los tres siguientes la página en la obra original y los tres últimos la línea en la cinta magnética."[4]

Ejemplo de textos con información bibliográfica para su identificación

[000090**021**]	[TODOS RÍEN ESTREPITOSAMENTE. SÓLO EL MECO, CON MUCHA GRAVEDAD E]
[000090**022**]	[INDIFERENCIA, CANTABA EN HORRIBLE FALSETE:]
[000090**022**]	[YO LE DABA UN CENTAVO Y ELLA ME DIJO QUE NO...]
[000090**023**]	[YO LE DABA MEDIO Y NO LO QUISO AGARRAR.]
[000090**024**]	[TANTO ME ESTUVO ROGANDO HASTA QUE ME SACÓ UN RIAL.]
[000090**025**]	[!AY, QUÉ MUJERES INGRATAS NO SABEN CONSIDERAR!]
[000094**016**]	[- NO, CURRO - RESPONDIÓ DEMETRIO SONRIENDO Y CON GESTO DESDEÑOSO - ;]
[000094**017**]	[NOSOTROS CAEMOS CUANDO ELLOS MENOS SE LO ESPEREN, Y YA. ASÍ LO HEMOS]
[000094**018**]	[HECHO MUCHAS VECES. ?HA VISTO CÓMO SACAN LA CABEZA LAS ARDILLAS POR LA]

Los recursos automatizados

Después de explicada la relación del CEMC con la bibliografía, se retoma la descripción general del sistema automatizado del CEMC. Una vez desarrollado el corpus, se consolidaron dos importantes puntos de acceso para la recuperación de la información dignos de destacarse, pues con distintas funciones consolidan el sistema computacional: el archivo de concordancias y el archivo de datos

[4] *Ibidem*.

estadísticos. Estos pueden consultarse desde 1991 por medio del Sistema Computacional del Diccionario del Español de México (SCDEM), mediante el programa INFORMIX. Este sistema que está basado en el CEMC, contiene tanto lengua hablada como escrita, ambas compiladas de fuentes originales.

Ahora bien, en esta parte de la investigación se detallan las funciones del *Sistema Computacional del Diccionario del Español de México* (SCDEM), que en 1990 estructuró Isabel García Hidalgo,[5] el mismo que se encuentra disponible en una base de datos, en la plataforma de INFORMIX, con las cuatro funciones de búsqueda que se diseñaron para su funcionamiento,[6] las cuales son: TIPOS, CONCOR, RECONCOR y TEXTOS, que se describen a continuación:

La función **TIPOS.** Permite recuperar los datos principalmente de frecuencias y de relaciones porcentuales estadísticas respecto a las palabras que se encuentran en el *Diccionario estadístico del español de México.*

Después de excluir información previamente determinada (principalmente nombres propios y números) y del análisis computacional, la computadora puede recuperar información de 1.891.058 palabras gráficas[7] por medio de este índice estadístico compuesto de 64194 palabras indizadas por extracción.

La información estadística puede ser solicitada por: número de orden, tipos (formas gráficas: palabra completas, raíz de palabra o palabra truncada),

---

[5] María Isabel García Hidalgo, *Versión para microcomputadoras IBM del sistema computacional del DEM* [Programa de computadora]; con la colaboración de María Luisa Pérez Valdespino, México, Diccionario del Español de México, 1990.
[6] Diseño implementado por el Dr. Boris Fridman Mintz, en el documento *Estructura de la base de datos del DEM*, México, Diccionario del Español de México, 1993, 13 h.
[7] En la indización automática se entiende como palabra gráfica a la grafía o palabra que resulta del análisis de textos, y se le identifica por estar separada por blancos de otras palabras en los textos.

categorías gramaticales, lemas (si se les asignó manualmente), frecuencia absoluta del tipo y porcentaje respecto al total de la muestra.

Ya recuperada la información, aparecen en pantalla, junto con los datos anteriores, los tres índices estadísticos *KF* (frecuencia corregida), *S* (índice *S* de corrección) y *C* (índice *C* de dispersión); y 42 celdas que indican la distribución de apariciones y frecuencias de la palabra por extracción o tipo consultado. Esto último puede ser consultado respecto a la clasificación interna del corpus; una de estas presentaciones muestra la información por los tres niveles de lengua (culta, sub-culta o no estándar), segundo nivel de clasificación. Y la otra, muestra los catorce géneros de lengua, tercer nivel de clasificación en que están divididos los textos del corpus. La recuperación en este caso es automatizada, aunque también hay índices impresos de estos mismos datos.

Ejemplo de recuperación de una palabra gráfica en el *Diccionario estadístico del español de México*

---

número [28941] tipo [GATO] categoría [8]
lema [GATO] f. tot. [45] e. tot. [.00238]

---14 GENEROS-

kf [34.2734975] s [.7616333] c [.7589208]

Frec. x género		Entre los géneros	Dentro del género
g1	[16]	[35.55556]	[.00593]
g2	[1]	[2.22222]	[.00033]
g3	[8]	[17.77778]	[.00231]
g4	[4]	[8.88889]	[.00197]
g5	[0]	[.00000]	[.00000]
g6	[0]	[.00000]	[.00000]
g7	[1]	[2.22222]	[.00144]
g8	[1]	[2.22222]	[.00078]
g9	[0]	[.00000]	[.00000]
g10	[1]	[2.22222]	[.00221]
g11	[12]	[26.66667]	[.00462]
g12	[1]	[2.22222]	[.00146]
g13	[0]	[.00000]	[.00000]
g14	[0]	[.00000]	[.00000]

Interpretación del ejemplo anterior, con el apoyo de la versión mecanografiada de la clasificación sociolingüística del CEMC:

Estos datos recuperados indican que el "tipo" o palabra gráfica [GATO] tiene el número de orden [28941]; es de categoría gramatical nominal [8]; está agrupado en el lema [GATO]; tiene como frecuencia absoluta [45]; que respecto al total de la muestra tiene [0.00238] de porcentaje; y que además de las medidas estadísticas *kf*, *s* y *c*, tiene información en los niveles de lengua (g1) **Lengua culta**: g2 periodismo, g3 ciencias, g4 técnicas, g7 habla culta; (g2) **Lengua sub-culta**, g8 literatura popular; g10 lírica popular y (g3) **Lengua no-estándar**:, g11 textos dialectales, y en g12 documentos antropológicos. Se concluye que tiene buena distribución de su información.

Esta función tiene su salida de información en archivos de sólo texto con tabuladores, que se pueden grabar en disco duro o en red.

Ejemplo abreviado de la información contenida en el *Diccionario estadístico del español de México*

núm.	*tipo o palabra gráfica*[1]	*categoría*	*vocablo*[2]	*frecuencia*	%	*KF*	*S*	*C*
43	A-	9	a2	16	0.00085	7.5494225	0.4718389	0.4658287
44	A	4	a2	44809	2.36953	44569.7766793	0.9946613	0.9946005
45	A	6	a2	1	0.00005	0.0361578	0.0361578	0.0251898
46	A/	4	a2	8	0.00042	3.5373627	0.4421703	0.4358225
47	ADIOS	8	a2	1	0.00005	0.1426661	0.1426661	0.1329101
48	AA	A	¡ah!	9	0.00048	1.2368447	0.1374272	0.1276116
49	AAA	A	¡ah!	1	0.00005	0.1374272	0.1374272	0.1276116
50	AAAAAAAAAYYYYY	A	¡ay!	1	0.00005	0.0674014	0.0674014	0.0567889
51	AAAAAAAAYY	A	¡ay!	1	0.00005	0.0674014	0.0674014	0.0567889
52	AAAAAAAY	A	¡ay!	1	0.00005	0.1374272	0.1374272	0.1276116
53	AAAAAAHHG	A	¡ahg!	1	0.00005	0.0674014	0.0674014	0.0567889
54	AAAAAYYYYY	A	¡ay!	1	0.00005	0.0674014	0.0674014	0.0567889
55	AAAH	A	¡ah!	2	0.00011	0.2022928	0.1011464	0.0909179
56	AAAY	A	¡ay!	4	0.00021	0.5066917	0.1266729	0.1167349
57	AAHHH	A	¡ah!	1	0.00005	0.0674014	0.0674014	0.0567889
58	AAPPROB-	9	aprobado	1	0.00005	0.0169065	0.0169065	0.0057194
59	AAY	A	¡ay!	4	0.00021	0.7129570	0.1782392	0.168888

60	AAYYYY	A	¡ay!	1	0.00005	0.0674014	0.0674014	0.0567889
61	AB-	9	[ambigua]	19	0.001	6.9536448	0.3659813	0.3587665
62	AB	4	ab	1	0.00005	0.1831331	0.1831331	0.1738376
63	ABA-	9	basar	1	0.00005	0.0287994	0.0287994	0.0177477
64	ABACA/	8	abacá	1	0.00005	0.0169065	0.0169065	0.0057194
65	A/BACO	8	ábaco	1	0.00005	0.1831331	0.1831331	0.1738376
66	ABADEJO	8	abadejo	1	0.00005	0.1831331	0.1831331	0.1738376

[1]Un vocablo puede tener varias palabras gráficas obtenidas por extracción, de alta o baja frecuencia, por esto mismo aquí se debe hablar de datos con frecuencias parciales.

[2]La columna de los vocablos no existe en los resultados originales automatizados, se tiene que incorporar posteriormente por medio de la indización humana por asignación, y se incluyó en este ejemplo, para aclarar la información presentada; por otro lado, los números en los vocablos se refieren a un arreglo lexicográfico interno y funciona como un calificador o símbolo, utilizado para diferenciar los distintos significados de los homógrafos.

La función **CONCOR.** Con esta función del INFORMIX se puede generar y recuperar las concordancias en su contexto del CEMC. La computadora permite recuperar 1891058 palabras gráficas analizadas en los 1932 textos completos. La recuperación se logra a partir de la previa identificación y la correspondiente solicitud al sistema de los “tipos” o palabras gráficas (palabras completas, raíz de palabra o palabra truncada), grafías específicas, palabras y cadenas de palabras. Se puede hacer esta solicitud simple, o acompañadas de sus categorías gramaticales.

Al contar con la indización por extracción que presenta el *Diccionario estadístico del español de México*, se pueden identificar las posibles búsquedas parciales, abiertas o delimitadas de concordancias o ítems en el sistema del corpus, y se pueden especificar solicitando las palabras con su categoría:

Cuadro de categorías gramaticales para solicitar información con características lingüísticas

Categorías	*Marcas en el Diccionario estadístico*	*Códigos de búsqueda en la base de textos*	*Palabras gráficas analizadas*
ambigua	0	a-n	51
adverbio	1	a	1918
adjetivo	2	b	837
conjunción	3	c	102
preposición	4	d	141
pronombre	5	e	670
artículo	6	f	204
contracción	7	g	2
nominal	8	h	47136
verbo	9	i	12450
(interjección)	A	a-n	488
(sin categoría)	[…]	a-n	195
Totales			64194

Se pueden hacer búsquedas específicas de las palabras gráficas conocidas, al incluir su categoría gramatical, o de sus posibles variantes al truncar la palabra o poner la raíz de la palabra, como en el siguiente ejemplo:

hHOMBREh u *hHOMBRE*h* u *hHOMBR*h*

También se pueden hacer búsquedas de cualquier palabra añadiendo "cualquier marca" gramatical, por ejemplo: pedir la palabra SER, sin importar si es nominal o si es verbo:

[a-n]SER[a-n]

O expresar búsquedas sin importar las palabras sino la combinación de categorías; utilizando las marcas del cuadro anterior se pueden hacer búsquedas por marcas gramaticales, por ejemplo: en el que se pida información que contenga: conjunción + artículo + nominal, en el que puede recuperarse algo como esto:

chYch *fELf* *hFERROCARRIL*

Los resultados de la generación de concordancias se almacenan automáticamente en archivos del directorio f:\usr\DEM\concor.txt, en formato de sólo texto, con lo que luego se pueden manipular en programas de procesamiento de palabras o en bases de datos, para ser entregados a los usuarios internos y externos que los solicitan.

Como resultado de este tipo de solicitud se obtienen las "concordancias", las cuales son registros muy parecidos a lo que conocemos como índice de palabras clave en su contexto (KWIC), donde se presentan tres líneas de texto; en la línea de en medio está la información solicitada y van acompañadas con una línea anterior y una posterior, lo que equivale a una recuperación de tipo real.

Esta información se presenta además con un código numérico que conlleva cada concordancia, este código se compone de 9 dígitos, los tres primeros son el código de texto, los otros tres son el número de página correspondiente al documento original de donde se tomó la información y los últimos tres corresponden a un número progresivo de control interno de captura. En esta etapa era necesario acudir a la lista bibliográfica de las obras que componen el *corpus*, que está ilustrado líneas arriba en el archivo bibliográfico.

Ejemplo de la información con concordancias[8] del CEMC
(3 concordancias de ayate)

AYATE

All (SIC), = AH/I) CONOC/I AL CAIM/AN
761001314 CON UN PESCADO EN **L'AYATE**
AYER QUE ME FUI A BA+AR

[8] En este ejemplo se nota el uso controlado de caracteres que equivalen a los signos diacríticos del texto original, esto se hizo para su almacenamiento por medio de tarjetas perforadas a mediados de la década de 1970. Una vez obtenida la información, estos caracteres se pueden convertir por un procesador de palabras en los signos diacríticos correspondientes.

834318037 ACOSTUMBRADA LA GENTE DE QUE CADA DOMINGO DAN LA DOMINICA PARA QUE SE SOSTENGA EL PADRE DE LA IGLESIA. SALEN VARIOS HOMBRES CON UN **AYATE** Y UNA ALCANCIA QUE VA RETRATADA LA PATRONA DE NUESTRA SE+ORA SANTA ANA.

AYATES

576026022 QUE VA COSIDA UNA BOLSA DE FORMA PIRAMIDAL DE MALLA MUY CERRADA (GENERALMENTE EMPLEAN **AYATES).** POR ESTA RAZ/ON DEL DI/AMETRO DE LA MALLA NO PASA DEL CENT/IMETRO. SU MANIOBRA

Interpretación del ejemplo anterior, con el apoyo de la versión mecanografiada de la *Bibliografía del CEMC*

La información recuperada corresponde a tres palabras gráficas; dos de AYATE y una de AYATES con sus respectivas concordancias.

Lo que significa que la palabra AYATE tiene tres concordancias como frecuencia total. Y que de estos datos se puede saber lo siguiente:

Palabra gráfica o "Tipo" **AYATE**

Concordancia 1 de 2.

Referencia bibliográfica: núm. **761**. *El Caimán I* [versión 21]. Cintas Colegio, Tamazunchale, S.L.P., 1963; Página 001 (Es la primera página del documento de *El Caimán I*).

Datos del levantamiento:

Estrato sociolingüístico: Es de un uso de lengua estándar, nivel lengua sub-culta; género 10; lírica popular; habla media.

Palabra gráfica o "Tipo" **AYATE**

Concordancia 2 de 2.

Referencia bibliográfica: núm. **834.** Parsons, Elsie Clews. "Folklore from Santa Ana Xalmimiluico, Puebla, México". *The Journal of American Foklore,* 45 (1932), 318-362.

Datos del levantamiento: [México, D.F., 1929. Un informante, 105: 22 a., m., medio.], la página 318.

Estrato sociolingüístico: Es de un uso de lengua no estándar, nivel lengua no estándar; género 1l; textos dialectales; lengua hablada.

Palabra gráfica o "Tipo" **AYATES**

Concordancia 1 de 1.

Referencia bibliográfica: núm. **576.** Mercado Sánchez, Pedro. *Breve reseña sobre las principales artes de pesca usadas en México.* Secretaría de Industria y Comercio, México, 1959. 79 pp., la página 26.

Datos del levantamiento:

Estrato sociolingüístico: Es de un uso de lengua estándar; nivel lengua culta; Género 4; técnicas; caza y pesca.

La función **RECONCOR.** Esta función sirve para reconstruir concordancias previamente generadas en CONCOR y que se hayan borrado del directorio CONCOR.TXT. Para esto hay que ver el directorio f:\usr\DEM\concor.lst.

La función **TEXTOS.** Con esta función se pueden consultar las líneas de texto junto con sus marcas gramaticales. Para contar con los textos del corpus, se capturaron 1932 textos distintos, que resultaron ser 219122 líneas de registros documentales, lo que los lexicógrafos llaman concordancias. La búsqueda es a partir de la solicitud de los nueve números de la propia concordancia, si se conocen, o también se pueden recuperar textos completos indicando de qué texto a qué texto se busca, Por ejemplo:

clave
[**762001279**]

texto

[e**LE**e i**DICES**i ch**QUE**ch i**ES**i f**UN**f h**GATITO**h]

(En esta línea de texto se puede apreciar que: *le* = pronombre; *dices* = verbo; *que* = conjunción; *es* = verbo; *un* = artículo, y *gatito* = nominal.)

Con el desarrollo de estos ejemplos de uso queda manifiesta la necesidad de agrupar la información por medio de la lematización o indización por asignación de

estas palabras gráficas en la computadora, y con ello poder consolidar así la unidad de los datos que todavía se mantenían electrónicamente separados. Se aclara esto, porque físicamente en los archiveros del DEM, ya casi todos los vocablos se encontraban agrupados de manera manual, es decir, los registros de cada vocablo y sus datos estadísticos ya se encontraban juntos en su "monografía documental", y fue con la ayuda de estos registros manuales que también se pudieron resolver una buena cantidad de problemas en los registros hechos para la computadora, con la intención de obtener por completo un sistema de almacenamiento y recuperación de información. Ahora se podrá comprender mejor qué es lo que sucedía al querer responder a las consultas de los usuarios, como se explica en el siguiente apartado.

2.2 LAS NECESIDADES DE LOS USUARIOS DETECTADAS POR SUS CONSULTAS

Una de las líneas que más se ha cuidado en el Diccionario del Español de México y la sección de documentación del mismo, y en general por el equipo lexicográfico, es la de prestar servicio de información a los usuarios que lo requieran, siempre desde un punto de vista práctico y solidario con la comunidad científica de México, y la de El Colegio de México, institución en que se desarrolla este proyecto académico. La respuesta a las necesidades de información se practica de manera general con usuarios externos nacionales e internacionales, que vienen a solicitar información sobre el léxico del español que se habla en México, aunque siempre

ha sido la prioridad de los servicios informativos del DEM[9] el proporcionar la mejor y mayor información posible a los investigadores internos para que realicen la redacción del Diccionario.

Por años en la sección de documentación, junto con los otros miembros del equipo del Diccionario, se ha ejercido la tarea de responder a las solicitudes formuladas por usuarios internos y externos, por medio del correo, correo electrónico, teléfono, oficios, o de forma directa en las mismas instalaciones del DEM. De este servicio se ha conservado un registro de usuarios, que ahora tiene forma de una base de datos,[10] la cual contiene información como: fecha de solicitud, nombre del solicitante, tipo de usuario, el perfil de la información solicitada, institución a la que pertenece el usuario, nombre del asesor (si tiene), título o tipo de investigación que realiza, tipo de respuesta ofrecida y tipo de información entregada.

Con la revisión de esta base de datos se ha obtenido una caracterización de las necesidades de información manifestadas por los distintos usuarios especialistas: internos y externos, personas o instituciones, estudiantes o investigadores, nacionales o internacionales, etc., a los que se ha dado respuesta, con los recursos que cuenta el DEM.

---

[9] Como se puede constatar al consultar documentos como: El Colegio de México. Diccionario del Español de México, *Proyecto de reglamento para la utilización de los resultados de DEM*, México, Diccionario del Español de México, 1978, [4 h]; Javier Becerra, "Los servicios del Diccionario del Español de México", en ponencia en 4° Simposio de la Asociación Mexicana de Lingüística Aplicada, *Presente y perspectivas de la lingüística computacional en México*, México: 24 al 26 nov. 1987, México, UNAM, 1988 y en Gilberto Anguiano Peña, *Documentación de palabras. Sistema de información de los recursos documentales del DEM. Hoja tipo WEB* [consulta por Intranet], México, El Colegio de México. Diccionario del Español de México, [2000-]. F:\palabras\PALABRAS-BASE DE DATOS.htm.

[10] Gilberto Anguiano Peña, *Usuarios del DEM* [base de datos], México, El Colegio de México. Diccionario del Español de México, [2003-]. Documento de consulta interna.

Los principales usuarios de información del Corpus del DEM son los mismos lexicógrafos del Diccionario, pero los usuarios externos han manifestado sus necesidades de información efectuando consultas al equipo lexicográfico, y al ser analizadas estas consultas se han podido identificar que los requerimientos más destacados son respecto a: uso de las palabras; información lingüística especializada; información testimonial e información estadística, que es obtenida del corpus lingüístico. Como es lógico, la interpretación de las consultas recibidas es la fuente para decidir las características de un sistema que responda a las necesidades internas y externas de los usuarios del DEM.

Por esto mismo a continuación se procede a describir las necesidades que se han visto manifestadas con cierta consistencia en las consulta de información de los usuarios sobre los materiales del corpus. En definitiva, el estudio de las necesidades ayuda a definir la estructura de la indización por asignación para implementar las acciones documentales pertinentes con el objetivo de su satisfacción y la complementación del sistema automatizado del DEM, en la parte de consulta. El análisis de la base de datos se realizó con el enfoque de la solución que se podía ofrecer, al estar limitados por un sistema previo donde se encuentra alojada la información.

Características de la información solicitada en la Base de usuarios del DEM, hasta 2006

Tipo de consulta	*Total*
Sobre el corpus (especializado)	293
General (varios**)**	225
Total	518

Las respuestas a la información solicitada del corpus

Tipo de respuesta	*Total*
Sobre el corpus	
Contextos, concordancias o ítems	203
Datos estadísticos de las palabras	60
Información sobre el Corpus	30

Algunas necesidades de información detectadas en la consulta del *Diccionario estadístico del español de México*

Tipo de información solicitada (datos estadísticos)
adjetivos (200) con alta frecuencia
adjetivos (300) con mayor frecuencia
Diccionario de la primaria, nomenclatura del
Diccionario del Español de México: listas de los tipos con sus respectivas frecuencias por género, frecuencia total su categoría gramatical
Diccionario estadístico de tipos. Orden decreciente de tipos 3 tomos (del 12501 a 25000).
Diccionario estadístico de tipos. Orden decreciente.
Diccionario fundamental del español de México, lista del
entradas (848) más utilizadas con su frecuencia, de la estadística fundamental del corpus.
español fundamental, Lista de frecuencias del
frecuencia más alta de aparición, frecuencia total y por géneros
frecuencias típicas de letras, diagramas, trigramas y palabras en el idioma Castellano
habla culta (etiquetado del género)
habla media (etiquetado del género)
hampa, textos del “CEMC”, la lista de palabras con su frecuencia de
humanidades, datos de textos del Corpus
índice C, la información (menor a 0.6) del
lemas
letras (Frecuencias de las) en la escritura del español a partir del CEMC
letras frecuentes en el español de México, un grupo de
léxico básico del mexicano
léxico con la mayor y mejor dispersión. Lista decreciente de tipos 2do. tomo (del 12501 a 25000).
léxico del español mexicano
morfemas no excesivas, listas de
Nomenclatura del DEM
nominales del CEMC (base de datos)
palabras claves, sus raíces y lemas
palabras con alta frecuencia, 200 sustantivos y 200 adjetivos (abstractas, concretas, objetos vivos o animados e inanimados, de seis a ocho). Lista en orden descendente de
palabras de mayor frecuencia en el CEMC para hacer un análisis de legibilidad de materiales didácticos, Lista de
palabras de mayor relevancia (1301) en el español de México con base en sus medidas estadísticas.
palabras lematizadas del CEMC
palabras más frecuentes en el CEMC
palabras más frecuentes en el CEMC para la elaboración de apoyos didácticos, lista de
palabras más frecuentemente usadas en español, Estoy buscando un listado de las
palabras que confrontan el Diccionario de la RAE con el Diccionario básico, Listado de
palabras, frecuencia de uso de
palabras: frecuencias absolutas e índices C

para la elaboración de una prueba de evaluación audiológica.
Periodismo (etiquetado del género)
sílaba en español de México, patrón canónico de la (tablas de resultado de la investigación sobre el)
sílabas, frecuencia de
sustantivos (150) de dos sílabas en singular, terminados en vocal (de entre 5 y 7 letras)
sustantivos (200) con alta frecuencia,
sustantivos (300) con mayor frecuencia. Los 300 adjetivos con mayor frecuencia
sustantivos incluidos en el corpus, frecuencias absolutas y relativas de los
Textos jurídicos incluidos en el Corpus del DEM. Las entradas marcadas con “Derecho”.
tipos de las letras A-C-P con índice de dispersión mayor que 0.5.
verbos, (150) de dos sílabas en infinitivo con terminación *-ar*, *-er* e *-ir* (el número de letra puede variar entre 5 y 7)
verbos, con sus frecuencias
verbos, de baja frecuencia
verbos, de movimiento, datos estadísticos
verbos, del CEMC (base de datos)
verbos, más comunes, bases de datos de frecuencias verbales
verbos, más frecuentes (los setecientos) y con mayor grado de dispersión
verbos, más frecuentes con su frecuencia absoluta
Verbos, que aparecen en el género 6 y 7 y copia del listado de la Nomenclatura del DEM. Versión para computadora personal del Analizador gramatical.
vocabulario básico del español mexicano, lista del
vocabulario fundamental y lista de palabras que muestra su longitud en sílabas
vocabulario fundamental, orden creciente, orden alfabético, tipos estadísticos

Con esta tabla se nota también la imperiosa necesidad de contar con la información completa de cada palabra, y el sistema automatizado como se vio en la descripción de los servicios la presenta de manera parcial y por separado.

Necesidades de información detectadas en la consulta de los textos por medio de índices de concordancias del CEMC

Tipo de información solicitada (ítems)
acronimia (para una investigación de)
adjetivos
adjetivos denominales
adverbios
artículos
artículos (concordancias sintácticas de los)
Balún Canán: lista de palabras
clíticos
colectivos léxicos
complementos preposicionales
concordancias
conjugaciones
conjunciones

construcciones
corpus, acceso al CEMC para un estudio sobre el léxico,
cuerpo (partes del)
derivación (para el estudio de la)
determinantes indefinidos
estudio de tú, usted y vos, en la deixis social
frases preposicionales
genética y la biología molecular (términos de la)
grafías (diferentes grafías: cs, cc, x de la A a la Z)
lengua culta (datos de la)
lengua no estándar (datos de la)
locuciones
medio ambiente (términos relacionados con el)
muerte (términos relacionados con la)
nexo: aunque
oraciones completivas de nombre
oraciones con estructuras
oraciones condicionales
oraciones irreales o contrafactuales
oraciones pasivas perifrásticas
palabras (científicas)
palabras (concordancias de)
palabras (que inician con)
palabras (que terrminan con)
palabras (terminadas con el sufijo)
partículas
posesivos en construcciones en las que canónicamente se usa artículo
prefijos
preposiciones
procedimientos derivacionales (para el estudio de)
pronombre reflexivo más verbo
pronombres
pronombres personales átonos de tercera persona
secuencias
sufijos
sustantivos
terminaciones
términos
variación de sustantivo a su uso como adverbio
variantes
verbos, con preposición relacionada con el verbo por no más de dos niveles sintácticos de análisis
verbos, con significados modales
verbos, con un complemento en plural
verbos, cuando hay dos verbos, uno finito y otro infinito
verbos, de movimiento
verbos, de movimiento que presenten alternancia de preposiciones a /
verbos, de opinión y sus completivas

verbos, en tercera persona en singular y plural
verbos, indicando los registros de lengua a los que pertenecen dichos contextos
verbos, más participio
verbos, modales del español (comportamiento sintáctico de los)
verbos, prepositivos
verbos, que aparecen con su complemento introducido por la construcción de que
verbos, seguidos de preposición de, más sustantivo
verbos, transitivos
verbos, y su alternancia con otro verbo

Esta tabla destaca la necesidad de responder con información completa sobre palabras de una determinada categoría gramatical y con suficiente información sintáctica, pero también se nota la búsqueda de información temática y por clasificación sociolingüística. En todo caso para cualquier resultado que se ofrezca se hace indispensable que cada ítem cuente con su correspondiente información bibliográfica y no sólo con un código referencial que obliga al usuario a consultar otras fuentes. En el caso de este sistema, la información bibliográfica y la clasificación sociolingüística fue un problema resuelto de manera precoordinada y manual para cada caso en particular, por lo que sólo se requeriría incorporar estos registros de manera automática a cada uno de los correspondientes ítems.

OBSERVACIONES

De la interpretación de las necesidades de información sobre el sistema computacional del DEM, es lógico concluir que lo más pertinente es agregar en forma permanente al sistema automatizado los puntos de acceso solicitados y aumentar el nivel de rapidez y calidad en las respuestas ofrecidas por el sistema. Recordemos uno de los argumentos que alientan este tipo de enfoques:

"Las investigaciones bibliotecológicas de los años ochenta resaltan la importancia de contemplar las necesidades de los usuarios como un elemento básico, así como el propio sistema de información… "[11]

Los datos sobre los usuarios y sus necesidades de información resultaron cruciales para tomar la decisión de agregar información en forma permanente para hacer más eficiente el control, manejo y acceso a la información lexicográfica de este corpus y convertirlo en un sistema de almacenamiento y recuperación de información. Para aumentar la eficiencia del sistema es imprescindible añadir valor agregado al contenido almacenado en la plataforma del programa INFORMIX de los componentes del *Corpus del español mexicano contemporáneo* (CEMC), esto es, además relacionando la bibliografía, la base de los textos, y el Diccionario estadístico.

En este mismo punto que se puede argumentar que si al *CEMC*, se le añadiera valor al contenido (en la recuperación de información), la consulta especializada ya no sería el único medio para tener acceso a la información de este recurso; es decir, resultaría práctico y enriquecedor para propios y extraños el abrir su consulta a todo tipo de usuarios, y que ellos mismos pudieran hacer sus propias búsquedas en el mismo corpus, reservando las consultas especializadas para que las responda el documentalista del DEM. Durante el tiempo que se desarrolló la investigación y en el futuro las consultas fueron y seguirán siendo controladas para retroalimentar el sistema.

---

[11] Catalina Naumis Peña, *Modelo de construcción de tesauros documentales multimedia: aplicaciones a los contenidos educativos en televisión*. Memoria para optar al grado de doctor, p. 381. http://www.ucm.es/BUCM/tesis/inf/ucm-t25976.pdf.

CAPÍTULO 3

COMPLEMENTOS AL PROCESO DOCUMENTAL DEL *CEMC*

En su tarea de hacer diccionarios, el lexicógrafo tiene necesidad constante de la mayor y mejor cantidad posible de información, que permita inventariar exhaustivamente la lengua, y aunque alguna información la recibe de forma natural en el habla cotidiana, por la radio, el periódico y la televisión u otros medios, no es sino la información documentada la que llega a ser verdaderamente útil en las labores lexicográficas, ya que esta información debe servir como testimonio o prueba del conocimiento claro y seguro del uso de las palabras.

Por lo previamente expuesto, se puede comprender que el trabajo de documentación es muy importante para el Diccionario del Español de México (DEM)**,** ya que de él depende la calidad que pueda adquirir la obra en cuanto al valor y la utilización de fuentes informativas.

Un aspecto destacado de la documentación se circunscribe a las actividades que se realizan para documentar o testificar la información utilizada en la producción de un diccionario.

Para establecer un contexto adecuado que permita valorar la importancia de estas actividades, es necesario aclarar que a la documentación en la disciplina lexicográfica, se le considera como parte del conjunto de *técnicas* y *criterios*[1] aplicados para la elaboración de léxicos y diccionarios, por consiguiente los procesos documentales tales como la selección, la adquisición, la catalogación, la

[1] Como quedó estipulado en dos documentos: El Colegio de México. Diccionario del Español de México, "Organización del equipo lexicográfico"*,* en *Manual de redacción del DEM*, México, Diccionario del Español de México, 1976, [11 h] (Manual de redacción: 4). (Documentos de trabajo del DEM) y en El Colegio de México. Diccionario del Español de México, *Manual DEM*, México, Diccionario del Español de México, 1973, [61 h] (Documentos de trabajo del DEM).

clasificación, el control, el análisis de contenido, la indización, la difusión y la recuperación de documentos, están estrechamente vinculados a la producción de diccionarios y en las labores lexicográficas, como se observa en la siguiente lista de pasos que se siguieron en la elaboración del sistema automatizado basado en el CEMC. En esta misma lista se indican con un asterisco [*] los pasos que no fueron previstos en el sistema automatizado del CEMC y que faltaron para completar la cadena documental:

1. Elección de una clasificación para este sistema.
2. Selección del material.
3. Adquisición del material.
4. Elaboración de la bibliografía del CEMC (en base de datos) y asignación de temas.(*)
5. Indización precoordinada.

5.1 Resultados de asignación de descriptores temáticos a la bibliografía.

6. Captura y almacenamiento de los textos.
7. Análisis de contenido

7.1. Análisis gramatical.

7.2. Análisis estadístico.

8. Indización automática.
9. Indización por lenguaje natural

9.1. Indización por extracción.

10. Indización por asignación.(*)
11. Indización semiautomática.(*)
12. Búsquedas.(*)

Con esta lista se puede identificar que sólo faltan los puntos 4, 10 11 y 12 para tener el sistema completo.

Complementación de la cadena documental

La documentación de la información, la que es indispensable para conseguir los intereses del DEM está basada en el análisis de contenido de los textos, éste por cierto tiene que hacerse de forma exhaustiva, pues para efectos de la investigación lexicográfica se necesita identificar el uso y origen de cada una de las palabras contenidas en los textos que componen esta muestra del español que se habla en México, el CEMC.

Con la meta de controlar totalmente la información generada por el CEMC y hacerlo de forma totalmente automatizada, se comenzó en el año 2002 a agregar valor a los contenidos de las bases que conforman el sistema automatizado del CEMC, con el firme propósito de poder abrir, en algún momento próximo, el acceso a la riqueza de la información del corpus, para que pueda ser consultada por todo tipo de usuarios, tanto internos como externos.

Además, esto se hizo también con la intención de poder transformar los rígidos resultados obtenidos de las consultas al sistema automático en Informix, de tal manera que pudieran ser fácilmente manipulados por cualquier tipo de usuario, es decir, la intención principal fue lograr que la recuperación de la información se pueda hacer de manera individual y fácilmente por parte de usuarios internos y usuarios externos, que hacen consultas sobre la información contenida en el CEMC.

Para aclarar conceptos, reflexionemos aquí lo que significa la recuperación de información a partir de lo que dice García Ejarque: "Acción y efecto de recuperar, previa búsqueda y localización los datos o información concreta que se desea de entre la almacenada en un fondo documental o en una memoria de ordenador".[2] Este enfoque es muy parecido a lo que se buscaba desarrollar en esta tesis, y es por esto mismo que lo tomaré como propio en esta tesis.

Pues bien, ya con esta meta de completar este proceso documental y con la idea de conseguirlo de la mejor manera posible, se observó la necesidad de que esos valores y puntos de acceso que había que añadir se tendrían que hacer tanto en el contenido como en la forma.

Se observó entonces que la validación de la indización automática por indización asignada y el llenado por completo de los registros de los textos o ítems con la información formal de la bibliografía y su clasificación sociolingüística, permitiría recuperar para cada palabra sus datos completos en las futuras búsquedas en el sistema de recuperación, con lo que ya no sería necesario recurrir a otras fuentes informativas.

Esto resultaba necesario hacerlo en estos tiempos, pues aunque estas etapas habían sido vislumbradas para cumplirse desde la misma elaboración del corpus, por razones de prioridades respecto a la redacción del diccionario y por carencias de equipamiento tecnológico en la sección de documentación en el DEM, no se habían concluido antes.

---

[2] Luis García Ejarque, *Diccionario del archivero bibliotecario: terminología de la elaboración, tratamiento y utilización de los materiales propios de los centros documentales*, Gijón, Asturias, TREA, 2000, xiv, 442, [4] p.

Por estas consideraciones quedó claro que el valor añadido al contenido, para conseguir la indización semiautomática, se tenía que incorporar básicamente en tres bases de datos: 1) La Bibliografía del CEMC, que había que generar como base; 2) Los textos del CEMC, que ya existía, y 3) El Diccionario estadístico, que también ya existía, para conseguir el sistema de almacenamiento y recuperación de información buscada.

3.1 *La indización por asignación: el valor añadido al contenido del* Diccionario estadístico

Como primer paso, en el año de 2002 en la sección de documentación del (DEM), se procedió a migrar a una PC personal, la información de la base de datos del *Diccionario estadístico* al programa Excel con la idea de que se pudiera tener acceso a sus datos duros desde cualquier máquina del DEM[3] y no únicamente a través de la licencia única de usuario, otorgada por el programa INFORMIX, pues esto mismo centralizaba todas las funciones del sistema en una computadora y a un operador de la misma. Lo más importante fue que con esta migración de datos, se pusieron las bases para que posteriormente en más adelante, ya en Excel, se manipulara la información y se pudiera poner en práctica la indización por asignación para resolver los problemas de agrupamiento, control y servicios de la información.

[3] Este diccionario estadístico se encuentra desde 1993 almacenado y dispuesto para su consulta como base de datos en una poderosa plataforma informática denominada Informix-SQL.

Hay que recordar que en el análisis automatizado original, la prioridad era efectuar un tipo de indización exhaustiva, con la que se identificaran de manera completa, sin excepciones, restricciones o límites, las palabras significativas con las que se puede representar el contenido de los textos completos que integran el corpus analizado.

Con el *Diccionario estadístico del español de México* en forma de base de datos se pudieron indizar por asignación las 64194 palabras gráficas (indizadas por extracción) y agrupar los datos que sobre una misma palabra se pudieran encontrar dispersos por medio de sus variantes. Los datos que se agruparían de estas 64194 palabras gráficas fueron los siguientes: número progresivo en el CEMC; palabra gráfica; categoría gramatical; frecuencia parcial; porcentaje; *KF*; *S*; *C*; y frecuencia por género, porcentaje entre géneros, porcentaje dentro del género, de los 14 géneros de lengua.

La lematización o indización por asignación

En los diccionarios el registro de entradas se sujeta a las convenciones lexicográficas tradicionales, con un afán económico y sistemático, pues con estas convenciones las variantes de una palabra se reducen a su forma canónica, es decir, se elige la forma más breve y "económica" para representar un conjunto de palabras, efectuándose con esto en la práctica un ciclo de condensación, mediante un proceso que los lexicógrafos llaman lematización y el cual es el que se siguió en la tarea de la indización por asignación en esta tesis.

A partir de 2004 y hasta mediados del 2006 se añadió el valor agregado al contenido del *Diccionario estadístico del español de México*, es decir, a los 64194 tipos o palabras gráficas obtenidas por indización automática se procedió a asignarles, a cada una de éstas, su correspondiente forma canónica o lema, por medio de la indización humana manual o lo que los lexicógrafos llaman lematización,[4] es decir, validar y editar los términos de indización propuestos por el análisis automático del lenguaje natural.

Para esto se tomó como punto de partida los resultados de la indización automática, o sea las palabras gráficas, las cuales se podían encontrar como raíz de palabra, palabra truncada y/o en palabras completas, y se procedió a asignar a cada una de éstas, su forma canónica correspondiente, o sea, agruparlas en torno a su lema.

De esta manera se obtuvo que todas las palabras gráficas (las variantes de una palabra de lengua natural) pudieran ser representadas bajo su forma representativa o canónica, y así poder sumarlas luego, con lo que se pudo obtener los totales de algunos datos estadísticos, y así también de esta manera se podría identificar y controlar las distintas formas o variantes que tiene una palabra contenida en el corpus, con el objetivo de poder recuperarlas o bien por su forma canónica o por la grafía de la palabra obtenida por extracción, utilizando para ello búsquedas prediseñadas por la sección de documentación del DEM en beneficio de los usuarios.

---

[4] Véase la definición de lematización del *Diccionario de lingüística*, México, Red Editorial Iberoamericana, 1991, p. 174, que dice: "*Lex* Reducción, a menudo automatizada, de las formas flexivas de los lexemas que aparecen en un determinado texto a su respectivo lema o forma de cita convencional; p. ej. las formas *tengo*, *tienen*, *tuvo*, *tenido*, etc. Con respecto del lema *tener*."

Ejemplo de lematización de las palabras gráficas

número en el CEMC	*palabra gráfica indizada*	*categoría gramatical*	*lema*	*frecuencia*	*porcentaje*	*KF*	*S*	*C*
136	ABEJA	nom	abeja	3	0.00016	0.5064400	0.1688133	0.1593548
137	ABEJAS	nom	abeja	8	0.00042	1.6328317	0.204104	0.1950471
6710	AVEJA	nom	abeja	1	0.00005	0.1831331	0.1831331	0.1738376
				12	0.00063			
982	ACTITUD	nom	actitud	204	0.01079	142.6448229	0.6992393	0.6958168
983	ACTITUDES	nom	actitud	55	0.00291	34.3496307	0.6245387	0.6202662
				259	0.0137			
1221	ADEMÁN-	v	ademán	1	0.00005	0.1426661	0.1426661	0.1329101
1222	ADEMAN	nom	ademán	1	0.00005	0.0674014	0.0674014	0.0567889
1223	ADEMÁN	nom	ademán	11	0.00058	2.2508696	0.2046245	0.1955735
1225	ADEMANES	nom	ademán	9	0.00048	3.5608571	0.3956508	0.3887736
				22	0.00116			
3056	ALUSION	nom	alusión	1	0.00005	0.1426661	0.1426661	0.1329101
3057	ALUSIÓN	nom	alusión	11	0.00058	3.8502445	0.3500222	0.3426258
3058	ALUSIONES	nom	alusión	7	0.00037	2.1018861	0.3002694	0.2923069
				19	0.001			
3253	ÁMBITO	nom	ámbito	59	0.00312	37.7448242	0.6397428	0.6356432
3254	ÁMBITOS	nom	ámbito	3	0.00016	0.9946542	0.3315514	0.3239448
				62	0.00328			
3838	ANIMA	nom	ánima	2	0.00011	0.2748544	0.1374272	0.1276116
3839	ÁNIMA	nom	ánima	1	0.00005	0.1426661	0.1426661	0.1329101
3862	ÁNIMAS	nom	ánima	5	0.00026	2.3202519	0.4640504	0.4579516
				8	0.00042			
4611	APOGEO	nom	apogeo	12	0.00063	5.9992595	0.4999383	0.4942478
4827	APUGEO	nom	apogeo	1	0.00005	0.0287994	0.0287994	0.0177477
				13	0.00068			

Fundamentos de la agrupación

En cuanto a los antecedentes de esta indización por asignación o lematización en el CEMC, es adecuado traer y transcribir la parte correspondiente al *Manual de*

redacción[5] donde se menciona la manera en que debe agruparse la información correspondiente a un vocablo[6] para obtener su correspondiente lema,[7] es decir, donde se explica como se debe proceder en la lematización de la información:

> "Dentro de las ocurrencias del CEMC, nos encontramos que un mismo vocablo puede aparecer con distintos tipos. En un ejemplo hipotético, se podría tener el vocablo aceptar con los tipos siguientes y número de ocurrencias:

N° de orden	*Palabras*	*Total de ocurrencias*
164	Aceptar	433
167	Aceptable	24
169	Aceptadas	45
170	Aceptado	36
172	Acepte	17
173	Acepto	18

> En este ejemplo, el vocablo aceptar en realidad ocurre con un total de 573 ocurrencias.
>
> La computadora produciría todas las frecuencias, absolutas y relativas, conjuntamente con las medidas estadísticas de modo semejante al cuadro que se

[5] El Colegio de México. Diccionario del Español de México, *Manual de redacción del DEM*, México, Diccionario del Español de México, 1976, *ca*. 250 h.
[6] "Vocablo. […] || *Lex* En estadística léxica, cualquier realización de un mismo lema o lexema; p. ej. *quepo* y *cupo* son formas de un mismo vocablo *caber*." Definición del *Diccionario de lingüística*, *op. cit.*, p. 301.
[7] "Lema. *Lex* En lexicografía, entrada léxica del diccionario en que se suministra diversa información y es a menudo representativa de diversas formas flexionadas; p. ej. *ir* es lema de *voy*, *vas*, *íbamos*, *fueron* y el resto de sus formas conjugadas." Definición del *Diccionario de lingüística*, *Ibid.*, p. 174.

presenta. Estas frecuencias y parámetros de todos los tipos deben reunirse en la del vocablo genérico.

En el ejemplo, el vocablo aceptar tiene las frecuencias y medidas estadísticas señaladas en el renglón en rojo. Para el caso de las frecuencias absolutas tendríamos por ejemplo:

Para el total: 573 = 433 + 24 + 45 + 36 + 17 + 18

Para G_1: 79 = 65 + 3 + 4 + 4 + 2 + 1

Para G_7: 19 = 13 + 1 + 3 + 2 etc."[8]

Fuentes para la indización

Es importante aclarar respecto a la indización por asignación o lematización que se efectuaron para completar los registros en la computadora, que se hizo casi en su totalidad apoyado en el propio índice automatizado del *Diccionario estadístico del Español de México*, o sea que la guía fueron los propios registros de palabras gráficas de manera directa,[9] pero cuando hubo dudas importantes, el apoyo se obtuvo de dos fuentes importantes para la lematización, una fue la nomenclatura[10] o lista de vocablos que existen en el mismo Diccionario del Español de México (DEM), (de donde se obtuvieron además las marcas que distinguen vocablos

[8] Explicación presentada en El Colegio de México. Diccionario del Español de México, *Manual de redacción del DEM*, *op. cit*.

[9] Aquí es justo aclarar que desde 1993 ya se encontraban asignados en la computadora 15929 lemas, de un total de 64194, los cuales eran correspondientes en su mayoría al *Vocabulario fundamental*, asignados por Boris Fridman y Luis Fernando Lara con base en los datos del documento Luis Fernando Lara, *Vocabulario fundamental,* México, El Colegio de México, 1979, mismos que en este 2007 el DEM publicó en un cuadernillo bajo el título *Resultados numéricos del vocabulario fundamental del español de México.*

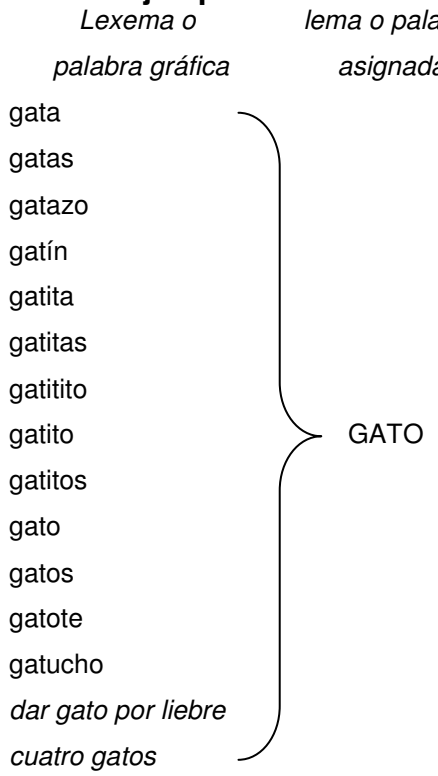
[10] Según el *Diccionario de la lengua española*, la entrada nomenclatura dice: ".||3. *Ling*. Serie de las voces lematizadas en un diccionario."

homógrafos), y la otra fue el *Diccionario de la lengua española* de la Real Academia Española. Sin embargo, cuando seguían las dudas sobre cómo lematizar una palabra gráfica, la orientación se obtuvo de los archivos de la documentación que se encuentran en el propio DEM, los cuales han sido desarrollados y resguardados por los lexicógrafos desde el inicio de este proyecto.

Es oportuno anotar que en la lematización, de manera general se hace la elección de la forma canónica para una entrada de diccionario en la forma masculina y singular. Esta forma elegida no niega la existencia de las otras formas de la misma palabra sino que es la manera económica de representarlas. Es decir, bajo una entrada en masculino y singular, el usuario o lector deberá imaginar o suponer que se encuentran todas sus otras formas: plural, femenino, aumentativo, diminutivo, superlativo, despectivo, apócope, etc., a menos que por alguna razón existan excepciones de alguna de estas formas, y estas excepciones en general sólo existen cuando han adquirido por consenso social un valor autónomo suficiente y llegan a ser entradas independientes.

Veamos en la siguiente página, como ejemplo de esto, lo que ocurre al lematizar la palabra GATO, con un ejemplo que no es real pero sí posible, en el que por ser sustantivo esta palabra, se da preferencia para su representación a la forma masculina y singular:

Posible ejemplo de lematización



Este tipo de ejemplo corresponde principalmente a los nominales o sustantivos, pero hay otros casos como los siguientes:

Ejemplo de información que es representada por una entrada

*Grafía**	*Razón para no ser lema*	*Forma canónica que la representa*
aigre	lengua hablada	**aire**
gata	femenino	**gato**
afueras	plural	**afuera**
puertota	aumentativo	**puerta**
ratititito	diminutivo	**rato**
fui	conjugación del verbo	**ir**
pus	lengua hablada	**pues**
sr	abreviatura	**señor**
wc	iniciales	**water**
le cayó el veinte	frase hecha	**veinte**

También es oportuno apuntar que una forma de gran importancia en la lematización es representar todas las variantes de un verbo bajo su forma de infinitivo o en su forma pronominal. Es decir, que todas las conjugaciones debemos de pensarlas representadas por su forma en infinitivo, como se explicará en este mismo apartado con el ejemplo del verbo ir, que incluye formas tan aparentemente lejanas como: fui, iba, ir, juites, va, vaya, vete, voy, yendo.

La indización por asignación en el CEMC

A continuación, se presentan algunos asuntos puntuales que se observaron en la indización semiautomática de este estudio, que por cierto, no abarcan la totalidad de los aspectos confrontados en la indización que se efectuó, pero se espera sirvan para sensibilizar al lector de este estudio sobre la indización humana agregada a las 64194 palabras gráficas obtenidas de la indización automática. Además, estos asuntos concretos pueden llegar a ser útiles a los documentalistas o a los mismos usuarios en la recuperación de la información, para comprender mejor la información que está representada en las fuentes de información de tipo léxico cuando las consulten:

Ejemplos de la Indización por asignación

Formas que no aparecen como entradas, sino que son representadas por otra forma canónica. La forma femenina se integra a la forma masculina en singular:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
77	ABANDERADA	nom	**abanderado**	2

De cómo una forma que aparece en plural y se lematiza en singular:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
75	ABALORIOS	nom	**abalorio**	1

Los diminutivos se incluyen en su forma canónica:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
71	ABAJITO	adv	**abajo**	4
47191	POCARITO	nom	**pókar**	1
58619	TAQUITOS	nom	**taco**	3

Respecto a los aumentativos hay que ubicarlos bajo su forma canónica:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
73	ABAJOTE	nom	**abajo**	1

Los superlativos por igual:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
54710	SANTÍSIMA	nom	**santo**	2
25945	EXCELENTÍSIMOS	nom	**excelente**	1

Otras se consideran abreviaturas que se agrupan también bajo la forma canónica:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
11172	CFR	nom	**confrontar**	1

21223	DR	nom	**doctor**	23
21713	EJ	nom	**ejemplo**	1
29756	GRM	nom	**gramo**	1
36584	LIC	nom	**licenciado**	12
35745	KM	nom	**kilómetro**	14

Los símbolos también se agrupan en la forma canónica:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
449	ÁC	nom	**ácido**	36
11642	CL	nom	**cloro**	3

Algunas interjecciones y expresiones (categoría A) que son claramente derivadas de otra palabra, se mantienen bajo la forma canónica:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
41484	NARANJAS	A	**naranja**	1
63566	VIVA	A	**vivir**	19

De cómo la forma canónica, principalmente para los adjetivos y sustantivos, es el masculino singular:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
46047	PERRA	nom	**perro**	28
46053	PERRAS	nom	**perro**	2
46054	PERRAZO	nom	**perro**	1
46059	PERRILLO	nom	**perro**	1
46060	PERRILLOS	nom	**perro**	2
46061	PERRITA	nom	**perro**	4
46062	PERRITO	nom	**perro**	14
46063	PERRITOS	nom	**perro**	3
46064	PERRO-	v	**perro**	1

46065	PERRO	nom	**perro**	131
46066	PERROS	nom	**perro**	95
46067	PERROTE	nom	**perro**	2

Aparentes diminutivos. Cuando una palabra que no es diminutivo mantiene su forma como entrada:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
13356	FINITO	nom	**finito**	6
22234	PERITO	nom	**perito**	4

Aparentes aumentativos. Hay que recordar que algunas palabras parecen ser aumentativos pero no lo son:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
138	ABEJÓN	nom	**abejón**	1

Aparentes plurales. Hay palabras del habla común y científica que son singulares y plurales a la vez:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
110	ABARROTES	nom	**abarrotes**	4
3544	ANÁLISIS	nom	**análisis**	216
7890	BINOCULARES	nom	**binoculares**	3
19760	DIABETES	nom	**diabetes**	5

Los apócopes se agrupan bajo su forma canónica. Hay información que abrevia otra palabra y que se le considera su apócope:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
48407	PREPA	nom	**preparatoria**	26
58869	TELE	nom	**televisión**	7

Algunas expresiones que no tienen vínculo con otras palabras se mantienen aparte:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
1157	ACHIS	A	**achis!**	1
16736	CHIN	nom	**chin!**	4
5275	ARRE	A	**arre!**	1

Diferencias en cuanto al registro de información. En algunos sistemas lexicográficos los adverbios no son tomados en cuenta y en otros sí:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
80	ABANDERAMIENTO	nom	**abanderamiento**	1

Cuando se detecta información de una locución latina. De cuando los datos se componen de dos o más palabras gráficas se ponen bajo su forma canónica:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
47896	POSTERIORI	adj	**a posteriori**	1
28102	FRAGANTI	nom	**in fraganti**	1
56426	SITU	nom	**in situ**	3

Cuando el sistema identifica formas verbales, la forma recuperada es la raíz de palabra:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
74	ABAL-	v	**abalar**	1

De cómo la forma canónica para el verbo en general, es el infinitivo y bajo éste debe agruparse la información, como aquí es el caso del verbo (ir):

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
7410	BAS	nom	**ir**	1
28372	FU-	v	**ir**	1355
28378	FUE	v	**ir**	2822
28390	FUERA	v	**ir**	97
28396	FUERAS	v	**ir**	1
28427	FUÍ-	v	**ir**	13
28428	FUI	v	**ir**	424
28430	FUÍMOS-	v	**ir**	4
28432	FUISM-	v	**ir**	1
28433	FUISTE	nom	**ir**	2
28434	FUISTE	v	**ir**	1
28435	FUISTES	nom	**ir**	1
31754	I-	v	**ir**	2
31756	IBA-	v	**ir**	249
31757	IBA	nom	**ir**	1
31758	IBA	v	**ir**	1
31759	IBAMOS-	v	**ir**	3
31760	ÍBAMOS-	v	**ir**	23
31761	ÍBAMOS	v	**ir**	1
31762	IBAN-	v	**ir**	85
31763	IBAN	nom	**ir**	1
31764	IBAN	v	**ir**	2
31765	IBAS-	v	**ir**	4
31788	ID	v	**ir**	6
31874	IDO-	v	**ir**	5
31875	IDO	v	**ir**	2
34842	IR-	v	**ir**	3422
34843	IR -	v	**ir**	2
34844	IR	v	**ir**	857
34845	IRA-	v	**ir**	1
34846	IRÁ-	v	**ir**	20
34848	IRA	A	**ir**	2
34849	IRÁ	v	**ir**	1
34856	IRAN-	v	**ir**	1
34857	IRÁN-	v	**ir**	6
34859	IRÁS-	v	**ir**	5

34860	IRAS	nom	**ir**	2
34864	IRÉ-	v	**ir**	16
34865	IREMOS-	v	**ir**	8
34866	IREMOS	v	**ir**	1
34868	IRÍA-	v	**ir**	7
34869	IRÍAN-	v	**ir**	3
34870	IRÍAS-	v	**ir**	1
34880	IRLE-	v	**ir**	2
34881	IRLE	nom	**ir**	1
34882	IRME-	v	**ir**	1
34986	IRSE-	v	**ir**	1
34987	IRSE	nom	**ir**	6
34988	IR=	v	**ir**	2
34989	IS-	v	**ir**	1
35536	JUI-	v	**ir**	1
35545	JUITES-	v	**ir**	1
62111	V-	v	**ir**	5333
62115	VA-	v	**ir**	704
62116	VA	nom	**ir**	8
62117	VA	A	**ir**	9
62118	VA	v	**ir**	6
62212	VÁIS-	v	**ir**	1
62313	VÁMONOS-	v	**ir**	1
62314	VÁMONOS	A	**ir**	3
62315	VÁMONOS	nom	**ir**	1
62316	VAMOS	nom	**ir**	14
62317	VAMOS	A	**ir**	13
62318	VAMOS	v	**ir**	2
62322	VAN-	v	**ir**	2
62323	VAN	nom	**ir**	1
62339	VANIR-	v	**ir**	1
62431	VAS-	v	**ir**	30
62432	VAS	nom	**ir**	2
62475	VAY-	v	**ir**	140
62476	VÁY-	v	**ir**	1
62477	VAYA-	v	**ir**	101
62480	VAYA	v	**ir**	1
62481	VAYAS-	v	**ir**	3
62482	VÁYASE	A	**ir**	1
62520	VEE	v	**ir**	1
63088	VETE-	v	**ir**	2
63089	VETE	nom	**ir**	1
63774	VOR-	v	**ir**	1
63797	VOY	v	**ir**	1262
63941	YENDO	v	**ir**	53
63979	YOY-	v	**ir**	1
Total				17177

Cuando hay palabras que se escriben igual pero gramaticalmente son distintas. De cómo existen homógrafos que únicamente se pueden diferenciar por su categoría gramatical de tal manera que tendrán distinta entrada en el diccionario, aquí por ejemplo la entrada ser[1] es verbo a diferencia de ser[2] que es sustantivo:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Entrada del diccionario*	*Frecuencia*
55625	SER	v	**ser**	**ser1**	22
55648	SERES	nom	**ser**	**ser2**	118

Problemas de indización

-En distintas categorías gramaticales. Hay problemas de lematización cuando algún lema está integrado por varias palabras gráficas, que a su vez tienen distintas categorías gramaticales. Por lo que le toca al redactor verificar si existe un error de análisis y resolver su ubicación, primero con la lematización, y luego con la asignación de entrada en el diccionario. Aunque hay que agregar que en la práctica lexicográfica, una misma entrada de diccionario puede tener distintas categorías gramaticales, dependiendo esto de sus usos, por ejemplo a la expresión ¡viva!, se le ha agrupado en general en la entrada del verbo *vivir*, o en otros casos, hay palabras que se les identifica como usadas como sustantivo y también como adjetivo.

-Con información confusa. De cuando una información recuperada tuvo datos de distintas palabras, por lo que no se pudo lematizar, entonces se marcó así [ambigua]:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
6956	B-	verbo	[ambigua]	9

-Sin recuperación información. Cuando un dato no pudo ser recuperado por medio del sistema se le puso la marca [no recuperada]:

Número de orden	*Ind. por extracción*	*Categoría gramatical*	*Ind. por asignación*	*Frecuencia*
42526	Ñ	nom	[no recuperada]	1

Recuperación con indización por asignación

Una vez que se le dio el valor agregado "lema" al contenido del *Diccionario estadístico* y que con este procedimiento se completó el registro automatizado de cada palabra gráfica, también se consiguió que todas las variantes de una palabra, pudieran ser representadas bajo su forma canónica y así pudieran ser sumadas, obtenerse los totales de sus datos estadísticos, y establecer sus relaciones estadísticas en toda la muestra. También con esto mismo el poder reconocer las distintas formas en que se encuentran estas variantes en el corpus, para poderlas recuperar en búsquedas prediseñadas en beneficio de los usuarios.

Todo esto permitió agrupar las 64194 palabras gráficas distintas que se pueden utilizar en la recuperación de información en los textos y contextos en El *Diccionario estadístico del español de México*, y ahora ya se pueden identificar los datos de 30899 vocablos o palabras indizadas por lenguaje natural con datos que indican sus aspectos estadísticos y de uso de la lengua: lemas; categoría gramatical; frecuencia; porcentaje; datos en los 14 géneros de lengua g [1], g [2], g

[3], g [4], g [5], g [6], g [7], g [8], g [9], g [10], g [11], g [12], g [13], g [14] y uso del español, nivel de lengua, Textos, mayor frecuencia, mejor distribución, palabras clave**,** clave de texto, temas.

Ejemplo del índice de palabras por asignación con la identificación documental de su uso

Palabra asignada	*cat. gram*	*frec. total*	*% total*	*uso del español*	*nivel de lengua*	*Textos*	*mayor frecuencia*	*mejor distribución*	*palabra clave*	*CLAVE*	*tema 1*
a2	prep	44836	2.37095	estándar			vocabulario fundamental	léxico común			
abacá	s	1	0.00005	estándar	lengua culta	discursos políticos			abacá	615	Discurso político
ábaco	s	1	0.00005	estándar	lengua culta	ciencias			ábaco	493	Artes gráficas
abadejo	s	1	0.00005	estándar	lengua culta	ciencias			abadejo	408	Biología
abadesa	s	5	0.00026	estándar							
abajeño	adj	4	0.00021	estándar	lengua sub-culta	lírica popular			abajeño	782	Habla regional
abajo	adv; interj	246	0.01300	estándar			vocabulario fundamental	léxico común			
abalar	v	1	0.00005	estándar	lengua culta	literatura			abalar	18	Obras de literatura
abalorio	s	1	0.00005	estándar	lengua culta	literatura			abalorio	101	Cuentos y ensayos aparecidos en revistas y suplementos
abanderado	adj; s; pp	14	0.00074	estándar	lengua culta						
abanderamiento	s	1	0.00005	estándar	lengua culta	técnicas			abanderamiento	514	Transporte
abanderar	v	2	0.00011	estándar	lengua culta	técnicas			abanderar	515	Transporte

3.2 COMPLEMENTACIÓN FORMAL DE CADA REGISTRO: EL VALOR AÑADIDO A LA *BIBLIOGRAFÍA DEL CEMC*

Una vez que estuvo indizado el CEMC, y se hicieron las primeras pruebas en la recuperación de ítems con ese sistema, se incrementó la vieja necesidad de contar con una base de datos de la bibliografía del CEMC, y de esta manera poder interpretar en forma rápida los resultados de las concordancias, pues dicha bibliografía se encontraba en papel y estaba como otro instrumento de referencia y ayuda a la interpretación de la información.

Por estas razones se procedió a la elaboración de la base de datos de la *Bibliografía del CEMC*, añadiendo a cada registro la correspondiente clasificación sociolingüística, del tipo de nivel de lengua y texto al que correspondía cada registro bibliográfico, aprovechando que estaban estos conceptos indicados claramente en la bibliografía en papel.

Los datos significativos en esta bibliografía bajo un criterio de clasificación precoordinada de tipo sociolingüístico, fueron integrados en la base de datos de la siguiente forma: hay *996* códigos de texto del CEMC con 1932 registros bibliográficos de igual cantidad de documentos de texto libre, éstos se refieren a 87 áreas temáticas distintas del conocimiento. A cada registro se le añadió sus marcas sociolingüísticas de “uso de lengua”; “nivel de lengua”; “género de lengua” y “texto” al que pertenece.

Ejemplo de base *Bibliografía del CEMC* con valor añadido al contenido

código *del CEMC*	*autor o entrevistador*	*título o núm. de cinta*	*editorial, periódico, etc.*	*lugar de edición o grabación*	*fecha de edición o grabación*	*Descripción*	*Uso del idioma*	*Nivel de lengua*	*Género*	*Textos*
000	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	español estándar	lengua culta	Literatura	Obras de literatura
001	Rulfo, Juan	*Pedro Páramo*	FCE	México	1964	0	español estándar	lengua culta	Literatura	Obras de literatura
002	Fuentes, Carlos	*La región más transparente*	FCE	México	1968	0	español estándar	lengua culta	Literatura	Obras de literatura
003	Guzmán, Martín Luis	*La sombra del caudillo*	Compañía General de Ediciones	México	1967	0	español estándar	lengua culta	Literatura	Obras de literatura
004	Vasconcelos, José	*El viento de Bagdad*	Letras de México	México	1945	pp. 151 a la 187	español estándar	lengua culta	Literatura	Obras de literatura
005	Rojas González, Francisco	*El diosero*	FCE	México	1960	0	español estándar	lengua culta	Literatura	Obras de literatura
006	Abreu Gómez, Ermilio	*Canek*	Robredo	México	1969	0	español estándar	lengua culta	Literatura	Obras de literatura

3.3 LOS REGISTROS BIBLIOGRÁFICOS Y LOS DATOS DE ESTRATIFICACIÓN:

EL VALOR AÑADIDO A LA BASE DE LOS TEXTOS DEL CEMC

El siguiente paso correspondió a migrar a una PC personal la información original de la Base de datos de los textos, la cual es de donde se generaran propiamente las “concordancias”, y se hizo al programa Excel, con la idea de que los redactores pudieran tener acceso a ellos desde cualquier computadora del DEM.

De esta manera a cada una de las 219122 líneas de texto capturado y sus códigos de concordancias, se les añadieron sus datos correspondientes, y ahora se pueden recuperar los siguientes datos en cada ítem: código del texto en el

CEMC; número de concordancia; texto etiquetado en el corpus; autor o entrevistador; título o núm. de cinta; editorial, periódico, etc.; lugar de edición o grabación; Fecha de edición o grabación; datos de publicación; uso del idioma, nivel de lengua, género, textos y número progresivo de captura. Con lo que se obtuvo el nuevo valor de contenido a la base de textos en cuanto a "registro bibliográfico" y a "niveles de lengua" y con ello tenemos 219122 registros documentales completos, cada uno con sus datos bibliográficos y de uso de lengua.

Respecto a la base de "textos" ahora ya se pueden identificar los datos de 219122 líneas de texto con datos que indican: como se usó en un tipo de lengua, en un nivel de lengua, en un léxico especializado, por un autor determinado, el título de la obra en que se dijo, los datos del pie de imprenta de la obra y el número de línea consecutivo dentro del corpus.

Después de efectuado todo lo anterior se obtuvieron los datos duros que existen hasta el momento por separado en cada base de datos, y se ha llegado a estimar pertinente que es el momento de desarrollar un sistema de consulta que pueda ofrecer la información contenida en las dos bases de datos en una vista o consulta, efectuada por algún usuario no especializado.

Para conseguir esto último, y lograrlo de la mejor manera posible, se pensó en solicitar la colaboración de la Unidad de Cómputo de El Colegio de México,[11] para que se consolide un sistema de consulta para Internet que facilite la

[11] Gilberto Anguiano Peña y Martha Elva Gómez Malagón, *Proyecto de un sistema de consulta vía Internet de la documentación del Diccionario del Español de México,* México, El Colegio de México, Diccionario del Español de México/Unidad de Cómputo, [2007-].

interacción de estas dos bases, con la idea de que los usuarios puedan tener el acceso al corpus por medio del índice por asignación (lematizado) de las 30899 palabras distintas de la lengua natural, acompañada de sus datos estadísticos totales o bien por medio del índice por extracción de las 64194 palabras gráficas (grafías tal cual se encuentran usadas las palabras en el corpus). Se trata pues, de utilizar uno, los dos índices, o su combinación para tener acceso a las concordancias. De las cuales se darían una cantidad limitada de ejemplos.

Recuperación de textos con marca gramatical y con el valor agregado a su contenido

Número de concordancia	*texto con marcas en el corpus*	*Autor o entrevistador*	*Título o Núm. de cinta*	*Editorial, periódico, etc.*	*Lugar de edición o grabación*	*Fecha de edición o grabación*	*Datos de publicación*	*Uso del idioma*	*Nivel de lengua*	*Género*	*Textos*
[000007003]	[dDEd aPRONTOa eSEe iOYÓi fUNf hDISPAROh, fELf hPERROh iLANZÓi fUNf hGEMIDOh hSORDOh cYc aNOa]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000007004]	[iLADRÓi aMÁSa.]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000013014]	[lDEMETRIOl@ iDESPERTÓi aSOBRESALTADOa, iVADEÓi fELf hRÍOh cYc iTOMÓi fLAf hVERTIENTEh]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000013015]	[hOPUESTAh gDELg hCAÑÓNh. aCOMOa hHORMIGAh hARRIERAh iASCENDIÓi fLAf hCRESTERÍAh,]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura

[000013016]	[hCRISPADASh fLASf hMANOSh dENd fLASf hPEÑASh cYc hRAMAZONESh, hCRISPADASh fLASf hPLANTASh]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000013017]	[dSOBREd fLASf hGUIJASh dDEd fLAf hVEREDAh.]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000020007]	[fLOSf hFEDERALESh iGRITABANi dAd fLOSf hENEMIGOSh, jQUEj, hOCULTOSh, hQUIETOSh cYc]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000020008]	[hCALLADOSh, eSEe iCONTENTABANi dCONd iSEGUIRi iHACIENDOi hGALAh dDEd fUNAf hPUNTERÍAh jQUEj]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura
[000020009]	[aYAa eLOSe iHABÍAi iHECHOi hFAMOSOSh.]	Azuela, Mariano	*Los de abajo*	Botas	México	1944	0	Español estándar	Lengua culta	Literatura	Obras de literatura

Recuperación de textos sin marca gramatical de la palabra GATO

Concordancia	Texto sin marcas en el corpus	Autor o entrevistador	Título o N° de Cinta	Editorial, periódico, etc.	Lugar de edición o grabación	Fecha de edición o grabación	Uso del idioma	Nivel de lengua	Género	Textos
[00802011]	[GATOS Y LOS PERROS DEL VECINDARIO.]	Benítez, Fernando	*El agua envenenada*	FCE	México	1973	Español estándar	Lengua culta	Literatura	Obras de literatura
[01205509]	[RELAMIÓ COMO UN GATO Y BISBISEÓ:	Magdaleno, Mauricio	*El ardiente verano*	FCE	México	1970	Español estándar	Lengua culta	Literatura	Obras de literatura

	'!CARAY, CARAY!' BARRIENTOS LO]									
[014018030]	[RUMBO DE MI CASA, UNAS VECES MAULLANDO COMO GATO, Y OTRAS, LADRANDO]	Romero, José Rubén	*La vida inútil de Pito Pérez*	Robredo	México	1944	Español estándar	Lengua culta	Literatura	Obras de literatura
[025117006]	[DE GATOS NI DE PERROS. EL TELÉFONO FUE ARRANCADO DE CUAJO, LAS]	Pacheco, José Emilio	*El viento distante*	Era	México	1969	Español estándar	Lengua culta	Literatura	Obras de literatura
[025125025]	[NO DE UN GATO SINO DE UN PERRO-LOBO.]	Pacheco, José Emilio	*El viento distante*	Era	México	1969	Español estándar	Lengua culta	Literatura	Obras de literatura

3.4 PROPUESTAS DE APOYO PARA LAS BÚSQUEDAS EN EL CEMC

Búsquedas internas

Una vez que se completaron los registros de las palabras gráficas del *Diccionario estadístico* se puede identificar qué palabras gráficas existen y cómo están marcadas en los textos del corpus, es entonces que se pueden recuperar estas mismas palabras en el sistema de recuperación (actualmente en el programa Excel), por medio de búsquedas de tipo lingüístico, utilizando las marcas gramaticales y teniendo en cuenta la tabla siguiente:

Tabla de códigos y etiquetas para efectuar búsquedas lingüísticas en el sistema automatizado de documentación del DEM

Categorías gramaticales	*Otra información*	*Etiquetas en los textos en el corpus*	*¿A qué responde?*
	ambigua	*xPALABRAx*	Busca una palabra ambigua
Adverbio (adv)		*aPALABRAa*	Busca un adverbio
Adjetivo (adj)		*bPALABRAb*	Busca un adjetivo
Conjunción (conj)		*cPALABRAc*	Busca una conjunción
Preposición (prep)		*dPALABRAd*	Busca una preposición
Pronombre (pron)		*ePALABRAe*	Busca un pronombre
Artículo (art)		*fPALABRAf*	Busca un artículo
Contracción (contrac)		*gPALABRAg*	Busca una contracción
Nominal (nom)		*hPALABRAh*	Busca una palabra nominal
nombre propio de objetos		*hPALABRAh^*	Busca el nombre de algo
nombre propio de lugares		*hPALABRAh@*	Busca el nombre de algún lugar
Verbo (v)		*iPALABRAi*	Busca un verbo
(conjunción?)		*jPALABRAj*	
exclamación e interjección		*kPALABRAk*	Busca una exclamación o una interjección
	Nombre propio	*lPALABRAl@*	Busca el nombre de alguien
	números	mNÚMEROm	Busca números, fechas, etc.
	Todas las categorías	*[a-n]PALABRA[a-n]*	Busca todas las categorías de una palabra

NOTA: en general el símbolo (^) después de una marca gramatical, corresponderá a su relación con algún nombre de persona, animal o cosa.

Para búsquedas delimitadas por su grafía

¿Qué buscar?	*Equivalencias*	*Ejemplo de búsqueda en el corpus*	*resultados*
Busca cualquier palabra en la base	Las palabras como se solicitan, sin restricción alguna.	palabra o PALABRA (Es decir, es igual poner los datos en minúsculas o en mayúsculas)	Recupera la información de una grafía que se encuentra en la base de datos, incluso, aunque no haya sido analizada por el CEMC (si se quiere tener toda la información de un vocablo hay que hacer una búsqueda para todas las variantes de una palabra)
Busca una grafía específica en el CEMC	Palabras tal y como aparecen en los textos	hPALABRAh (En mayúsculas más su categoría gramatical)	Recupera únicamente la grafía que tiene la categoría gramatical buscada.
Busca una grafía del CEMC con cualquier categoría	palabras gráficas o por extracción del CEMC	*PALABRA*	Recupera información de una palabra con cualquier categoría gramatical.
Busca las palabras que empiecen con	raíz de palabra o palabra truncadas	hPALABR*	Recupera información de las palabras que inician o tienen el prefijo buscado
Busca las grafías que terminen con	terminaciones o sufijos	*ALABRAh	Recupera información de las palabras que terminan o tienen el sufijo buscado

Organización para búsquedas de usuarios externos

Respecto a la forma de buscar y recuperar la información contenida en el corpus y pensando principalmente en los usuarios externos, fue que se consideró pertinente ordenar la información de una manera que el usuario tenga a la vista únicamente los vocablos controlados que fueron obtenidos de la indización semiautomática y que conforman el *Índice de frecuencias del Léxico del español*

usado en México. Esto para que a partir de la solicitud de búsqueda de un único vocablo se efectúe la recuperación de las distintas palabras gráficas que estén representadas por éste, mediante la aplicación de una fórmula de búsqueda interna.

En otras palabras, a partir de la solicitud que haga un usuario de un vocablo que está identificado con un número progresivo de entrada (el cual corresponde a su orden alfabético) se procederá, de manera interna e imperceptible al usuario, a recuperar en los textos del corpus los ejemplos limitados y aleatorios de todas las palabras gráficas que componen al vocablo buscado y que necesariamente tienen marcado el mismo número de entrada.

Para que este tipo de búsqueda tuviese éxito y no se recuperaran datos ajenos a los del vocablo solicitado, resultó un requisito indispensable el que las palabras gráficas en la búsqueda llevasen su marca gramatical original, tal y como se encuentran marcadas y almacenadas en los textos del corpus.

Ejemplo de la estructura interna (semi-oculta al usuario) para búsquedas de información controlada

Búsqueda con palabras asignadas (dato único a la vista del usuario)	*Núm. de entrada*	*Palabras gráficas extraídas con categoría buscadas en los textos*	*Frecuencia parcial*	*Registro progresivo*
a2	1	iA-i	16	1
	1	dAd	44809	2
	1	fAf	1	3
	1	dÁd	8	4
	1	hADIOSh	1	5
	1	aAVERa	1	6
abacá	2	hABACÁh	1	7
ábaco	3	hÁBACOh	1	8
abadejo	4	hABADEJOh	1	9
abadesa	5	hABADESAh	4	10

	5	fABADESAf	1	11
abajeño	6	hABAJEÑASh	4	12
abajo	7	aABAJITOa	4	13
	7	aABAJOa	241	14
	7	hABAJOTEh	1	15
abalar	8	iABAL-i	1	16
abalorio	9	hABALORIOSh	1	17
abanderado	10	hABANDERADAh	2	18
	10	hABANDERADOh	8	19
	10	hABANDERADOSh	4	20
abanderamiento	11	hABANDERAMIENTOh	1	21
abanderar	12	iABANDER-i	2	22
abandonado	13	hABANDONADAh	5	23
	13	hABANDONADASh	4	24
	13	fABANDONADASf	1	25
	13	hABANDONADOh	8	26
	13	iABANDONADOi	1	27
	13	hABANDONADOSh	6	28
abandonar	14	iABANDON-i	157	29
	14	iABANDONi	1	30
	14	iABANDONA-i	1	31
	14	iABANDONADO-i	1	32
	14	hABANDONANDOh	1	33
	14	hABANDONARh	1	34
	14	iABANDONAS-i	1	35
	14	hABANDONEh	1	36
	14	iHABANDON-i	1	37
abandono	15	hABANDONOh	24	38
abanicar	16	iABANIC-i	3	39
	16	iABANICA-i	2	40
abanico	17	hABANICOh	10	41
	17	hABANICOSh	2	42
abaratar	18	iABARA-i	1	43
	18	iABARAT-i	4	44
abarcar	19	iABARC-i	73	45
	19	iABARCAi	1	46

Búsquedas almacenadas

Una vez que ha sido organizada la información como en el cuadro anterior, se puede facilitar un tipo de búsqueda controlada en beneficio de los usuarios, si se desarrolla una interfaz que resulte cómoda en alguna base de datos y la cual guíe al usuario para que éste escriba en una casilla en blanco la palabra que quiere solicitar. Además, junto a dicha casilla se deberá informar al usuario que la búsqueda será por acercamiento, es decir, que en cuanto vaya escribiendo su palabra, la lista de vocablos controlado (que sí están contenidos en el sistema) se irá moviendo de tal manera que se muestre el vocablo buscado, o el más parecido al de la solicitud. De esta manera el usuario se dará cuenta por sí mismo si el sistema tiene o no tiene la palabra que le interesaba consultar y/o si existe otra que le pueda servir a sus propósitos. En caso de que el usuario encuentre una palabra de su interés, entonces podrá presionar el botón de “buscar textos”, para que el sistema recupere una muestra limitada y aleatoria de concordancias o de todos los ítems que componen dicho vocablo. La entrega de textos se realizará sin etiquetas gramaticales para no confundir al usuario. Los textos con marcas sólo se proporcionarán a solicitud por escrito de los usuarios, vía correo electrónico de preferencia, y después de ser autorizada su solicitud.

Ejemplo de búsqueda en el Corpus por Internet
(Propuesta para usuarios externos)

DICCIONARIO DEL ESPAÑOL DE MÉXICO

(Primera ventana)

Corpus del español mexicano contemporáneo, 1921-1974
*Antecedentes
*Explicación
* Bibliografía
*Búsquedas en textos

(Segunda ventana)

Búsquedas en textos
*Búsqueda por vocablos recuperables en el corpus
*Búsqueda abierta

(Tercera ventana)

Búsqueda por vocablos recuperables en el corpus	
Lista de palabras recuperables en el sistema	Búsqueda por acercamiento
ABANDERADO ABANDERAMIENTO ABANDERAR ABANDONADO ABANDONAR ABANDONO ABANICAR	Escriba la palabra: ABANDO (Búsqueda de vocablos por acercamiento)
Buscar textos (presionar)	

Búsquedas ilimitadas

Es importante explicar que el sistema de búsquedas contará en esta misma tercera ventana, con una opción ilimitada en la búsqueda, y podrá manejarse como cualquier procesador de palabras en la que el usuario podrá solicitar libremente una palabra en cualquiera de sus variantes y si así lo prefiere también combinada con otras palabras, de tal manera que en el caso de que hubiera información en el corpus se recuperarán los contextos de la información solicitada. Con esta búsqueda se podrán recuperar diminutivos, aumentativos, despectivos, plurales, femeninos, etc., y también se podrán encontrar frases hechas, locuciones, clichés y más, pero el inconveniente de esta búsqueda, es que el

usuario para localizar cada una de estas variantes, deberá intuirlas y tendrá que hacer una búsqueda por separado para cada una de ellas.

Búsquedas temáticas[12] por direcciones desglosadas

Como parte de la actualización y mejoramiento de los servicios de información se podrá ofrecer la búsqueda por temas asignados. Lo que permitirá delimitar las búsquedas en sectores del corpus. Llegando a obtenerse la información de las palabras pertenecientes a distintos sectores del CEMC, tales como:

Uso de lengua estándar.

Uso de lengua no estándar.

Nivel de lengua culto.

Nivel de lengua sub-culto.

Nivel de lengua no estándar.

Por 14 géneros de lengua.

Por 87 temas de los textos.

Para hacer búsquedas por los 87 temas de los textos, el usuario se podrá apoyar en los temas precoordinados explícitos en la Bibliografía del CEMC, tal y como lo dejaron establecidos originalmente los lexicógrafos desde un principio, según se puede constatar en el siguiente desglose de la estructura del CEMC:

[12] La información contenida en los siguientes cuadros la elaboré como parte de la investigación para esta tesis y para la versión preparada para Internet, de este mismo Corpus.

PRIMERA CLASIFICACIÓN:
usos del idioma en el CEMC

Uso	*Género*	*Códigos de texto*	*Concordancia inicial y final*
Español estándar	G1-G10	000-784	[000007003] a [784001200]
Español no estándar	G11-G14	785-999	[785038027] a [999001207]

SEGUNDA CLASIFICACIÓN:
niveles de lengua en el CEMC

Nivel	*Género*	*Códigos de texto*	*Concordancia inicial y final*
Lengua culta	G1-G7	000-667	[000007003] a [667418419]
Lengua sub-culta	G8-G10	668-784	[668032055] a [784001200]
Lengua no estándar	G11-G14	785-999	[785038027] a [999001207]

TERCERA CLASIFICACIÓN:
géneros dentro de los niveles de lengua en el CEMC

Géneros	*Género*	*Códigos de texto*	*Concordancias*
G1 Literatura	G1	000-149	[000007003] a [149001041]
G2 Periodismo	G2	150-325	[150011034] a [325101157]
G3 Ciencias	G3	326-505	[326001019] a [505071054]
G4 Técnicas	G4	606-607	[506010014] a [607110055]
G5 Discursos políticos	G5	608-625	[608011036] a [625089019]
G6 Religión	G6	626-637	[626241001] a [637216008]
G7 Habla culta	G7	638-667	[638011011] a [667418419]
G8 Literatura popular	G8	668-730	[668032055] a [730143004]
G9 Habla media	G9	731-760	[731003001] a [760033193]
G10 Lírica popular	G10	761-784	[761001001] a [784001200]
G11 Textos dialectales	G11	785-914	[785038027] a [914008215]
G12 Documentos antropológicos	G12	915-947	[915038043] a [947395049]
G13 Textos jergales	G13	948-959	[948002001] a [959038204]
G14 Textos del hampa y conversación popular	G14	960-999	[960000001] a [999001207]

Listado de temas precoordinados en la bibliografía.
Orden alfabético de los textos clasificados del CEMC

textos y temas	*Género*	*nivel*	*uso*	*Clave de género*	*código de texto*	*Identificación bibliográfica: por texto, página y línea de captura*
Administración	ciencias	culta	estándar	G3	429-435	[**429**023**040**] a [**435**060**126**]
Agropecuarias	técnicas	culta	estándar	G4	557-575	[**557**564**010**] a [**575**020**026**]
Albañilería	técnicas	culta	estándar	G4	552-553	[**552**005**019**] a [**553**001**236**]
Antropología	ciencias	culta	estándar	G3	351-354	[**351**008**022**] a [**354**223**008**]
Arqueología	ciencias	culta	estándar	G3	355-356	[**355**020**033**] a [**356**982**008**]
Arquitectura	ciencias	culta	estándar	G3	479-483	[**479**301**027**] a [**483**248**027**]
Arte dramático	ciencias	culta	estándar	G3	495-496	[**495**010**007**] a [**496**256**014**]
Artes coreográficas	ciencias	culta	estándar	G3	484-486	[**484**006**003**] a [**486**343**009**]
Artes gráficas	ciencias	culta	estándar	G3	493-494	[**493**022**004**] a [**494**128**034**]
Artes plásticas	ciencias	culta	estándar	G3	487-492	[**487**015**034**] a [**492**052**096**]
Astronomía	ciencias	culta	estándar	G3	379-383	[**379**014**024**] a [**383**119**202**]
Bibliotecología	ciencias	culta	estándar	G3	326-327	[**326**001**019**] a [**327**376**014**]
Biología	ciencias	culta	estándar	G3	404-415	[**404**043**018**] a [**415**107**091**]
Carpintería	técnicas	culta	estándar	G4	537-540	[**537**013**007**] a [**540**053**178**]
Caza y pesca	técnicas	culta	estándar	G4	576-579	[**576**009**003**] a [**579**020**040**]
Charrería	técnicas	culta	estándar	G4	582	[**582**014**060**] a [**582**039**063**]
Cine y fotografía	ciencias	culta	estándar	G3	502-505	[**502**007**002**] a [**505**071**054**]
Comercio	ciencias	culta	estándar	G3	449-453	[**449**014**006**] a [**453**001**178**]
Computación	ciencias	culta	estándar	G3	400-403	[**400**001**002**] a [**403**061**025**]
Contabilidad	ciencias	culta	estándar	G3	436-448	[**436**003**017**] a [**448**018**245**]
Conversación popular	Textos del hampa y conversación popular	no estándar	no estándar	G14	970-999	[**970**001**001**] a [**999**001**207**]
Correos y filatelia	técnicas	culta	estándar	G4	506-507	[**506**010**014**] a [**507**00**3273**]
Corte y confección	técnicas	culta	estándar	G4	551	[**551**011**005**] a [**551**191**018**]

Cuentos y ensayos aparecidos en revistas y suplementos culturales	Literatura	culta	estándar	G1	095-149	[**095**020**003**] a [**149**001**041**]
Culturas indígenas	ciencias	culta	estándar	G3	339-342	[**339**007**014**] a [**342**293**035**]
Derecho	ciencias	culta	estándar	G3	357-362	[**357**012**005**] a [**362**695**030**]
Dibujo técnico	técnicas	culta	estándar	G4	548	[**548**013**004**] a [**548**013**260**]
Discurso político	discursos políticos	culta	estándar	G5	608-625	[**608**011**036**] a [**625**089**019**]
documentos antropológicos	documentos antropológicos	no estándar	no estándar	G12	915-947	[**915**038**043**] a [**947**395**049**]
Economía	ciencias	culta	estándar	G3	363-367	[**363**056**012**] a [**367**103**021**]
Editoriales	periodismo	culta	estándar	G2	172-206	[**172**043**108**] a [**206**057**076**]
Educación y pedagogía	ciencias	culta	estándar	G3	343-344	[**343**026**042**] a [**344**248**017**]
Ejército	técnicas	culta	estándar	G4	580-581	[**580**073**002**] a [**581**031**181**]
Electricidad	técnicas	culta	estándar	G4	541-544	[**541**188**019**] a [**544**093**032**]
Electrónica y electricidad	ciencias	culta	estándar	G3	389-391	[**389**021**002**] a [**391**027**144**]
Enfermería	técnicas	culta	estándar	G4	549-550	[**549**013**003**] a [**550**139**055**]
Filosofía	ciencias	culta	estándar	G3	328-332	[**328**014**019**] a [**332**107**047**]
Física	ciencias	culta	estándar	G3	392-395	[**392**017**014**] a [**395**067**188**]
Fotonovela	Literatura popular	sub-culta	estándar	G8	694-708	[**694**009**001**] a [**708**001**498**]
Geofísica	ciencias	culta	estándar	G3	396-399	[**396**014**007**] a [**399**068**003**]
Geografía	ciencias	culta	estándar	G3	368-370	[**368**010**031**] a [**370**351**040**]
Habla de la Ciudad de México	habla culta	culta	estándar	G7	638-667	[**638**011**011**] a [**667**418**419**]
Habla media	lírica popular	sub-culta	estándar	G10	761-764	[**761**001**001**] a [**764**001**430**]
Habla media de la Ciudad de México	habla media	sub-culta	estándar	G9	731-760	[**731**003**001**] a [**760**033**193**]
Habla regional	lírica popular	sub-culta	estándar	G10	765-784	[**765**001**001**] a [**784**001**200**]
Herrería	técnicas	culta	estándar	G4	556	[**556**122**034**] a [**556**181**020**]
Historia	ciencias	culta	estándar	G3	333-338	[**333**019**016**] a [**338**054**036**]
Historieta	literatura popular	sub-culta	estándar	G8	709-718	[**709**001**001**] a [**718**001**254**]

Ingeniería aérea	técnicas	culto	estándar	G4	531-532	[**531**009**033**] a [**532**099**013**]
Ingeniería automotriz	técnicas	culta	estándar	G4	528-530	[**528**002**017**] a [**530**036**046**]
Ingeniería civil	técnicas	culta	estándar	G4	518-522	[**518**011**014**] a [**522**029**177**]
Ingeniería de ferrocarriles	técnicas	culta	estándar	G4	533	[**533**026**012**] a [**533**419**004**]
Ingeniería de minas	técnicas	culta	estándar	G4	535-536	[**535**001**002**] a [**536**136**016**]
Ingeniería industrial	técnicas	culta	estándar	G4	523-524	[**523**011**001**] a [**524**026**179**]
Ingeniería naval	técnicas	culta	estándar	G4	534	[**534**017**004**] a [**534**376**030**]
Ingeniería química	técnicas	culta	estándar	G4	525-527	[**525**013**032**] a [**527**031**254**]
Matemáticas	ciencias	culta	estándar	G3	384-388	[**384**013**003**] a [**388**011**092**]
Mecánica	técnicas	culta	estándar	G4	545-547	[**545**001**012**] a [**547**163**013**]
Medicina humana	ciencias	culta	estándar	G3	458-478	[**458**075**035**] a [**478**328**009**]
Medicina y veterinaria	ciencias	culta	estándar	G3	454-457	[**454**019**033**] a [**457**259**173**]
Mercadotecnia	técnicas	culta	estándar	G4	516-517	[**516**009**021**] a [**517**107**028**]
Música	ciencias	culta	estándar	G3	497-501	[**497**007**007**] a [**501**005**233**]
Novela popular	literatura popular	sub-culta	estándar	G8	719-730	[**719**017**029**] a [**730**143**004**]
Novela rosa	literatura popular	sub-culta	estándar	G8	668-685	[**668**032**055**] a [**685**127**023**]
Obras de literatura	literatura	culta	estándar	G1	000-094	[**000**013**015**] a [**094**043**077**]
Periodismo	técnicas	culta	estándar	G4	508-509	[**508**021**003**] a [**509**104**012**]
Plomería	técnicas	culta	estándar	G4	554-555	[**554**005**003**] a [**555**141**016**]
Política	ciencias	culta	estándar	G3	371-374	[**371**021**027**] a [**374**150**033**]
Psicología	ciencias	culta	estándar	G3	345-350	[**345**025**012**] a [**350**417**020**]
Publicidad	técnicas	culta	estándar	G4	510-511	[**510**029**021**] a [**511**131**012**]
Química	ciencias	culta	estándar	G3	416-428	[**416**052**045**] a [**428**163**012**]
Radio y televisión	técnicas	culta	estándar	G4	512-513	[**512**017**017**] a [**513**310**008**]
Religión		culta	estándar	G6	626-637	[**626**241**001**] a [**637**216**008**]
Reportajes de autores mexicanos	periodismo	culta	estándar	G2	150-171	[**150**011**034**] a [**171**009**262**]
Reseñas culturales	periodismo	culta	estándar	G2	250-284	[**250**081**004**] a [**284**066**199**]

Reseñas deportivas	periodismo	culta	estándar	G2	285-309	[**285**024**016**] a [**309**216**251**]
Reseñas policíacas	periodismo	culta	estándar	G2	310-317	[**310**013**78**] a [**317**706**181**]
Reseñas políticas	periodismo	culta	estándar	G2	207-241	[**207**015**111**] a [**241**056**180**]
Reseñas sociales	periodismo	culta	estándar	G2	242-249	[**242**293**071**] a [**249**387**032**]
Reseñas taurinas	periodismo	culta	estándar	G2	318-325	[**318**161**001**] a [**325**101**157**]
Sociología	ciencias	culta	estándar	G3	375-378	[**375**005**001**] a [**378**176**037**]
Telenovela	literatura popular	Sub-culta	estándar	G8	686-693	[**686**004**001**] a [**693**077**077**]
Textos del hampa	textos del hampa y conversación popular	no estándar	no estándar	G14	960-966	[**960**000**001**] a [**966**000**163**]
Textos del hogar	técnicas	culta	estándar	G4	583-607	[**583**020**020**] a [**607**110**055**]
Textos dialectales	textos dialectales	no estándar	no estándar	G11	785-914	[**785**038**027**] a [**914**008**215**]
Textos jergales	textos jergales	no estándar	no estándar	G13	948-959	[**948**002**001**] a [**959**038**204**]
Transporte	técnicas	culta	estándar	G4	514-515	[**514**017**027**] a [**515**324**026**]

Esta tabla con forma de índice es de alguna manera, la descripción mínima para identificar y recuperar la información por tema contenida en el corpus.

DISCUSIÓN Y CONSIDERACIONES FINALES

Los resultados fueron estructurados a partir de la conformación y complementación de nuevos datos y relaciones en las diferentes bases de datos que alimentan el sistema que permite desarrollar y mantener el DEM. Las bases de datos y las tareas que se realizan en torno al DEM tienen el objetivo primordial de obtener el testimonio de uso de las palabras, y a la vez, documentar con información lexicográfica imparcial y fidedigna el uso y significado de las palabras del español que se habla en México.

La conformación y complementación de las bases de datos se realizó a partir del análisis de las necesidades de información de los usuarios externos y la información sobre las palabras requerida por los redactores, correctores y revisores de este mismo proyecto, para proceder a hacer su análisis, síntesis, redacción y edición para cada una de las palabras incluidas, y de esta manera producir un diccionario hecho por mexicanos para mexicanos.

El enfoque en esta investigación fue orientado por la documentación y la indización, pues esto permitió concluir el proceso documental que se detectó como una necesidad de los usuarios del sistema del DEM.

El proceso documental original llegó a un estado de desarrollo lo suficientemente consolidado para poder apoyar con éxito en las labores internas de documentación en el Diccionario del Español de México (DEM), sin embargo, para poder tener la información lexicográfica completamente controlada y automatizada, incluso para poder ofrecerla en consulta a usuarios internos y externos en un sistema de almacenamiento y recuperación de información, fue

necesario concluir el proceso documental, y para ello se procedió al estudio que respaldó las acciones que lo completan.

Como se ha explicado a lo largo de las páginas precedentes, los puntos de acceso y elementos de recuperación añadidos se incorporaron básicamente a tres bases de datos distintas:

-El Diccionario estadístico.

-La Bibliografía del CEMC.

-Los textos del CEMC.

Al indizar por asignación y/o agrupamiento las palabras detectadas por indización automática en el Diccionario estadístico se pudo identificar y relacionar la forma canónica que correspondía a cada una de las palabras gráficas presentes en el sistema. Esto permite identificar las palabras del lenguaje natural con que se puede recuperar la información de la base de textos completos.

Después de identificar y agrupar las 64194 palabras gráficas en sus 30899 formas canónicas, se pudo hacer una suma simple para obtener la frecuencia total y también el porcentaje total de cada palabra del lenguaje natural. Una vez que se tuvieron las palabras de lengua natural o vocablos se incluyeron los datos que las identifican con el uso de lengua, nivel de lengua, géneros de lengua, mayor frecuencia, mejor distribución y en los casos que correspondiera se identificaron las palabras significativas o palabras clave contenidas en textos especializados.

La manipulación de la bibliografía como base de datos *ex profeso* permitió definir varios puntos de acceso para los 1932 registros bibliográficos y obtener una

recuperación acorde con las necesidades de los usuarios. La estratificación y ubicación física en el corpus crea nuevas relaciones y usos para la información que ofrecen los registros bibliográficos.

En la base de textos del CEMC se generaron dos versiones de ítems, uno con marcas gramaticales y otra sin marcas, para poder hacer distintas clases de búsquedas internas y externas, de manera tal que la manipulación sea más legible para usuarios externos, sin perder de vista las condiciones que necesita mantener para la consulta interna del equipo que realiza el DEM. También se relacionó con los registros bibliográficos y la estratificación de las palabras, completando los registros de los ítems, con información formal y clasificación sociolingüística.

Finalmente se obtuvo el estado de cómo se encuentra el sistema documental del léxico español de México en el CEMC en cuanto a su estructura y el comportamiento de la información.

RESULTADOS ANTERIORES DE LA INDIZACIÓN AUTOMÁTICA (EXTRACCIÓN)

2086050 Total palabras gráficas capturadas

- 194992 Palabras gráficas rechazadas para el análisis, principalmente por ser nombres y números.

1891058 *Total* de palabras de lengua natural aceptadas para analizar

Estado general del sistema

Uso de lengua	2
Nivel de lengua	3
Géneros de lengua	14
Núm de claves de textos asignadas	996
Temas en los textos analizados	87
Documentos adquiridos de texto completo	1932
Líneas capturadas	219122
palabras de lengua natural capturadas	2086050
Palabras aprobadas de lengua natural	1891058
Palabras gráficas obtenidas de la indización automática (indización por extracción)	64194
palabras indizadas por lenguaje natural (indización por asignación)	30899

Recuperación real

palabras de la lengua natural en el CEMC	*Palabras indizadas por extracción*	*Palabras indizadas por asignación*
1891058	64133	30899

Resultados de la indización por asignación

Con la indización por lenguaje natural manual por asignación se redujeron las distintas palabras pertenecientes a la lengua natural. Lo que ayudó a efectuar la operación de sumar únicamente los datos del *Diccionario estadístico*, porque no se efectuaron las operaciones estadísticas sobre el KF, S y C, que requieren desarrollar fórmulas complejas y que deberán realizarse con posterioridad.[1]

[1] Junto con el lexicógrafo y poeta, Francisco Segovia, inicié esta idea de transformar al corpus en algo accesible para todo mundo, aunque por mi necesidad de terminar esta tesis y por la suya de hacer las bases del diccionario electrónico no hemos podido seguir en el mismo camino hasta aquí recorrido, pero es seguro que volveremos a unirnos con este mismo fin más adelante. Además Francisco me ha dado ideas sobre el manejo de la información y me ha brindado consejos muy oportunos durante el desarrollo de esta investigación.

Con dicha operación de sumar se descubrieron varios aspectos importantes en la identificación de las palabras que se pondrán a disposición de los usuarios del DEM:

-Con estas acciones se ha podido identificar de manera categórica un total de 30899 formas canónicas o de lengua natural que por asignación se encuentran documentadas en el sistema.

-En cada palabra se pudo agrupar las distintas categorías gramaticales con que fueron marcadas automáticamente.

-Se obtuvo la frecuencia total de cada palabra que permite su ponderación en el español de México a partir de los datos absolutos de frecuencia, porcentaje y distribución (pues antes estaban parcialmente automatizados y no sumados).

-Se obtuvieron las palabras de mayor frecuencia.

-Se obtuvo el porcentaje total de cada palabra respecto al total de la muestra de cerca de dos millones de palabras.

-Se identificaron las palabras con la mejor dispersión en la muestra.

-Se identificaron el total de apariciones que una palabra tuvo por su uso en la lengua estándar o en la no estándar.

-Se identificó el total de apariciones de una palabra en los niveles de lengua: lengua culta, lengua sub-culta o lengua no estándar.

-Se identificó el total de apariciones de una palabra en catorce géneros de lengua.

-La delimitación de las palabras gramaticales o palabras vacías (las que sirven de enlace entre las palabras significativas, pero que ellas mismas no son consideradas útiles para la recuperación de la información).

-Se identificó si hay palabras exclusivas de las clasificaciones sociolingüísticas o de un léxico individual o a un texto especializado.

-La identificación documental de las palabras significativas conforme a su origen de uso en la lengua estándar o no estándar.

-La identificación documental de las palabras significativas respecto al nivel de lengua en el que se usaron en los textos: lengua culta, lengua sub-culta, lengua no estándar.

-La identificación documental de las palabras significativas respecto al género de lengua en que aparecieron: literatura, periodismo, ciencias, técnicas, discursos políticos, religión, habla culta, literatura popular, habla media, lírica popular, textos dialectales, documentos antropológicos, textos jergales, textos del hampa y conversación popular.

-La identificación documental de las palabras significativas de uso compartido en dos géneros de lengua especializados (por ejemplo: ciencias y técnicas).

-La identificación documental de las palabras significativas de uso exclusivo en un léxico especializado, considerando para esto el uso en los textos de un campo de conocimiento: filosofía, administración, agropecuarias, albañilería, antropología, arqueología, arquitectura, arte dramático, artes coreográficas, artes gráficas, artes plásticas, astronomía, bibliotecología, biología, etc., etc.

-La estructuración de un *Índice de frecuencias del léxico del español usado en México.*

Resumen de indización de palabras por asignación en el sistema

	Indización automática		*Indización humana*	
Palabras Iniciadas con la letra:	*Palabras por extracción (tipos)*	*% palabras por extracción (tipos)*	*palabras por asignación*	*% palabras por asignación*
a	6916	10.7735925	3376	10.92592
b	1913	2.98002929	1071	3.46613
c	7501	11.6848927	3568	11.54730
ch	633	0.98607346	359	1.16185
d	4420	6.8853787	2049	6.63128
e	5198	8.09732997	2469	7.99055
f	2020	3.14671153	1037	3.35610
g	1428	2.22450696	777	2.51464
h	1697	2.64354924	814	2.63439
i	3341	5.20453625	1590	5.14580
j	579	0.90195345	272	0.88029
k	85	0.13241113	46	0.14887
l	1589	2.47530922	778	2.51788
ll	205	0.31934449	45	0.14564
m	3786	5.89774745	1859	6.01638
n	1187	1.84908247	534	1.72821
ñ	14	0.02180889	5	0.01618
o	1436	2.23696919	623	2.01625
p	6256	9.74545908	2930	9.48251
q	410	0.63868897	144	0.46603
r	3595	5.60021186	1765	5.71216
s	4033	6.28251862	1902	6.15554
t	3381	5.26684737	1678	5.43060
u	474	0.73838677	196	0.63432
v	1749	2.7245537	776	2.51141
w	23	0.03582889	20	0.06473
x	31	0.04829112	23	0.07444
y	90	0.14020002	56	0.18124
z	194	0.30220893	137	0.44338
OTROS	10	0.01557778	0	0
totales	*64194*	*100*	*30899*	*100*

Con todas estas aportaciones y operaciones sobre el *Diccionario estadístico del español de México* se pudo obtener un producto esperado desde el principio de la investigación con el corpus, el *Léxico del español usado en México*, y ya se puede tener como una opción pertinente "la recuperación de información lexicográfica" en los textos del corpus, pues este recurso proporciona información primaria con datos fidedignos, imparciales y suficientes que permiten identificar a las palabras de la lengua natural.

Seguimiento de la indización por asignación en el corpus

Tabla del procesamiento de la información sobre las palabras del corpus

Letra inicial	*Palabras analizadas*	*% Palabras analizada*	*Indización automática*	*% Indización automática*	*Indización humana*	*% Indización humana*
a	152716	8.06994	6916	10.77359	3376	10.92592
b	22599	1.19335	1913	2.98003	1071	3.46613
c	144842	7.65341	7501	11.68489	3568	11.54730
ch	3235	0.17047	633	0.98607	359	1.16185
d	158453	8.37508	4420	6.88538	2049	6.63128
e	228373	12.07125	5198	8.09733	2469	7.99055
f	32079	1.69468	2020	3.14671	1037	3.35610
g	18999	1.00379	1428	2.22451	777	2.51464
h	48798	2.57882	1697	2.64355	814	2.63439
i	34493	1.82150	3341	5.20454	1590	5.14580
j	6328	0.33426	579	0.90195	272	0.88029
k	688	0.03624	85	0.13241	46	0.14887
l	169018	8.93669	1589	2.47531	778	2.51788
ll	8235	0.43540	205	0.31934	45	0.14564
m	96634	5.11601	3786	5.89775	1859	6.01638
n	65306	3.45244	1187	1.84908	534	1.72821
ñ	23	0.00119	14	0.02181	5	0.01618
O	34615	1.82946	1436	2.23697	623	2.01625

p	158199	8.36030	6256	9.74546	2930	9.48251
q	81861	4.32840	410	0.63869	144	0.46603
r	35830	1.89176	3595	5.60021	1765	5.71216
s	147434	7.79303	4033	6.28252	1902	6.15554
t	75018	3.96405	3381	5.26685	1678	5.43060
u	50389	2.66414	474	0.73839	196	0.63432
v	38845	2.05301	1749	2.72455	776	2.51141
w	62	0.00324	23	0.03583	20	0.06473
x	51	0.00264	31	0.04829	23	0.07444
y	76762	4.05914	90	0.14020	56	0.18124
z	1116	0.05882	194	0.30221	137	0.44338
OTROS	57	0.00300	10	0.01558	0	0.00000
	1891058	99.95551	64194	100.00000	30899	100.00002

Esta fue la forma en que se comportó la información en su aspecto general

Procesos	*Logros obtenidos*	*Unidades*
Textos completos	Palabras almacenadas	2086050
Indización automática	Palabras analizadas	1891058
Indización automática	Palabras por extracción	64184
Indización humana	Palabras Vocabulario fundamental	1418294
Indización humana	Palabras léxico común	1277673
Indización humana	Palabras por asignación	30899
Indización humana	Palabras de léxico especializado asignadas	14720

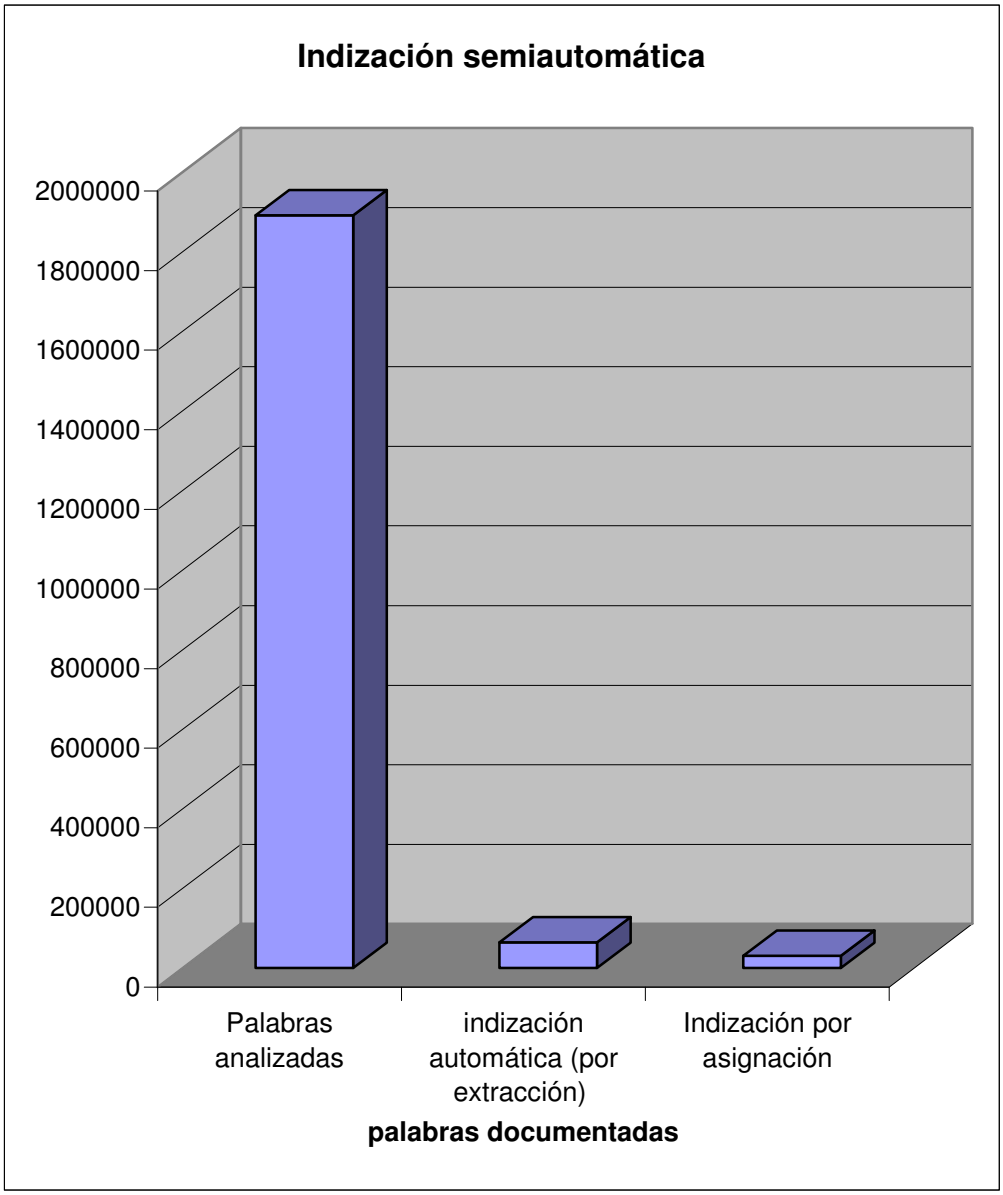
Resultado de las palabras asignadas de que por uso pertenecen exclusivamente* a un género de lengua

Género	*Claves*	*Palabras asignadas*
Literatura	G1	2706
Periodismo	G2	1880
Ciencias	G3	4876
Técnicas	G4	1574
CyT	G3 Y G4	664
Discursos políticos	G5	137
Religión	G6	122
Habla culta	G7	190
Literatura popular	G8	763
Habla media	G9	186
Lírica popular	G10	206
Textos dialectales	G11	819
Documentos antropológicos	G12	154
Textos jergales	G13	312
Textos del hampa y conversación popular	G14	131
Totales		14720

* La forma en que se identificó estos datos fue por medio de la operación de exclusión, es decir, que no aparecen en los otros géneros de lengua.

Como resultado de lo expuesto aquí, en esta tesis se muestra cómo se han logrado reducir las búsquedas a efectuar en el sistema a través de la forma canónica y con esto resultan las consultas mucho más rápidas y la información buscada sobre las palabras es completa, como se muestra en la siguiente gráfica.

Reducción y agrupación de palabras significativas para recuperar información en el CEMC



Palabras analizadas	*indización automática (por extracción)*	*Indización humana (por asignación)*
1891058	64194	30899

Respecto a la integración de los servicios de información del DEM

Con el objeto de apoyar con información a los redactores en su toma de decisiones y para agilizar la recuperación de la información lexicográfica, se ha creado una página tipo WEB para el sistema Intranet. Esta página denominada "*Documentación de palabras. Sistema de información de los recursos documentales del DEM*"[2] está constituido por varios tipos de información lexicográfica, lexicológica, documental y bibliográfica, algunos de los cuales son generados y actualizados sistemáticamente en el propio DEM:

-Información primaria (documentos originales):

-Información testimonial (bases de datos de documentos completos);

-Información estadística con datos numéricos sobre las palabras;

-Información referencial en la que hay registros sobre las palabras y se remite a otras fuentes).

Como resultado de la investigación la documentación de las palabras ya se puede consultar por los especialistas en las instalaciones del Diccionario del Español de México por medio del sistema de información en Intranet, y a fines del 2007 todo usuario interesado podrá tener acceso por medio del Internet.

---

[2] Gilberto Anguiano Peña, *Documentación de palabras. Sistema de información de los recursos documentales del DEM. Hoja tipo WEB* [consulta por Intranet], México, El Colegio de México. Diccionario del Español de México, [2000-]. F:\palabras\PALABRAS-BASE DE DATOS.htm.

Esta investigación podrá ser consultada en principio por los investigadores del diccionario, por medio de la red interna de El Colegio de México-DEM, junto con otros recursos informativos en F:\palabras\PALABRAS-BASE DE DATOS.htm.

DICCIONARIO DEL ESPAÑOL DE MÉXICO

DOCUMENTACIÓN DE PALABRAS

SISTEMA DE INFORMACIÓN DE

LOS RECURSOS DOCUMENTALES DEL DEM

(Estos datos son preliminares. Los datos sobre los vocablos básicamente son correctos y definitivos, pero si se llega a identificarse algo por corregir o si hay que actualizar información, se procede a efectuar los ajustes correspondientes de manera periódica en las distintas bases de datos):

(Elija una letra inicial)

♠ PALABRAS DOCUMENTADAS EN EL DEM

(INFORMACIÓN GENERAL Y REFERENCIAL):

A B C CH D E F G H I J K
L LL M N Ñ O P Q R S T U
V W X Y Z

♠ ÍNDICE SEMIAUTOMÁTICO DEL ESPAÑOL USADO EN MÉXICO CON: ESTRATIFICACIÓN, FRECUENCIA Y PORCENTAJE (INFORMACIÓN ESTADÍSTICA):

TODOS (A-Z)

♠ TEXTOS DEL CORPUS DEL ESPAÑOL MEXICANO CONTEMPORÁNEO (1921-1974) (INFORMACIÓN TESTIMONIAL):

♠ ESTRUCTURA DEL CEMC POR ESTRATIFICACIÓN

(ORGANIZACIÓN DE LA INFORMACIÓN PARA SU ANÁLISIS SOCIOLINGÜÍSTICO)

♠ TABLA DEL CÓDIGO DE BÚSQUEDAS EN EL CEMC (INSTRUCTIVO)

♠ PALABRAS GRÁFICAS PARA LA BÚSQUEDA Y

RECUPERACIÓN DE INFORMACIÓN PARA EL CEMC (INFORMACIÓN REFERENCIAL)

♠ TEXTOS DEL CORPUS SIN MARCA GRAMATICAL (INFORMACIÓN TESTIMONIAL):

G1 G2 G3 G4 G5 G6 G7

G8 G9 G10 G11 G12 G13 G14

TODOS (G1-G14)

♠ TEXTOS DEL CORPUS CON MARCA GRAMATICAL (INFORMACIÓN TESTIMONIAL):

G1 G2 G3 G4 G5 G6 G7

G8 G9 G10 G11 G12 G13 G14

TODOS (G1-G14)

♠ BIBLIOGRAFÍA DEL CEMC

♠ DICCIONARIO ESTADÍSTICO DEL ESPAÑOL DE MÉXICO

POR LETRA INICIAL (INFORMACIÓN ESTADÍSTICA):

TODOS (A-Z)

OTROS RECURSOS DE IMPORTANCIA:

♠ ÍNDICE DE MEXICANISMOS (INFORMACIÓN REFERENCIAL)

♠ BIBLIOTECA DEL COLMEX (INFORMACIÓN REFERENCIAL)

♠ BIBLIOGRAFÍA DEL DEM (INFORMACIÓN REFERENCIAL)

♠ BIBLIOGRAFÍA DE FUENTES PARA DOCUMENTAR
EL ESPAÑOL DE MÉXICO (INFORMACIÓN REFERENCIAL)

DICCIONARIO DEL ESPAÑOL USUAL EN MÉXICO
(BASE DE DATOS EN INTERNET)

♠ Base de datos "USUARIOS"

RESULTADOS OBTENIDOS DE LA INVESTIGACIÓN

Respecto al problema a resolver en esta tesis, considero que se cumplió debido a que se logró obtener un subsistema de información, el cual se integró al sistema documental del léxico del español de México perteneciente al DEM, mismo que ahora cuenta con la indización por asignación para agrupar palabras en los contenidos documentales, con lo que se resuelven la mayoría de los problemas propios de la indización por extracción. Asimismo para potenciar la recuperación de la información, se logró incorporar los registros bibliográficos y su liga a los documentos en texto completo.

Con lo anterior se completó el proceso documental del DEM y se logró su transformación en un sistema de almacenamiento y recuperación de información del que se podrán extraer datos de manera sistemática y así responder a las necesidades de información manifestadas por los usuarios en cuanto a significado y a registros documentales completos. A partir de esto mismo las respuestas a las solicitudes podrán ser completas, controladas, ilimitadas y fáciles de efectuar por el mismo usuario al utilizar una computadora personal.

En cuanto a la metodología utilizada en el estudio de caso pienso que fue la idónea ya que se logra mostrar la viabilidad para transformar los sistemas basados en los *corpora* a sistemas de recuperación de información que beneficien ampliamente a sus usuarios, completando el proceso documental, que en este estudio se consolidó con tres aspectos: a) Como resultado del análisis léxico del diccionario se obtuvo para el sistema del DEM una indización por asignación, para

agrupar palabras de los contenidos documentales; b) Como resultado de la agrupación de las palabras se obtuvieron registros estadísticos completos. Es decir, las diferentes variantes de una palabra al estar agrupadas reflejan la frecuencia absoluta, el porcentaje respecto al total de la muestra y otros índices estadísticos que ahora se pueden consultar en el sistema, y c) Como resultado del análisis documental se elaboraron los registros bibliográficos y las ligas a los documentos originales.

Respecto a los distintos objetivos que orientaron esta investigación se pude afirmar que fueron cumplidos debido a los siguientes aspectos. En cuanto al objetivo principal buscado con la indización semiautomática se estructuraron puntos de acceso y relaciones que complementan el sistema original. En consecuencia se obtuvo un subsistema de información complementario al original, en el que se tomó en cuenta las necesidades de los usuarios, a partir de lineamentos propios al centro de información en que se encuentra dicho sistema.

Se propone y fundamenta la consolidación del uso del CEMC y su base de datos en la documentación del léxico del español de México; se validaron o rechazaron las palabras obtenidas por la indización por extracción en el CEMC; se completó la información de los ítems con sus registros bibliográficos; se completó la información de los ítems con información sociolingüística; y se obtuvo un sistema que responde a las necesidades de los usuarios y los guiará en sus búsquedas.

Asimismo el índice de palabras obtenidas por extracción de manera automática llamado *Diccionario estadístico del español de México* al ser objeto de

una indización por asignación por parte de un documentalista, cuenta ahora con información lematizada o agrupada, lo cual permitió se completaran los datos sobre las palabras contenidas en el sistema. Esta agrupación hizo posible que se efectuaran las operaciones matemáticas correspondientes y que se obtuvieran dos tipos de resultados, unos fueron estadísticos y otros documentales.

Los resultados estadísticos. Con la agrupación de las palabras obtenidas por la indización por asignación, se obtuvieron los siguientes datos: frecuencia total, porcentaje total, índice estadístico *KF* de corrección e índice *C* de distribución; además se obtuvieron tres datos para los 14 géneros del corpus: frecuencia por género, porcentaje por género y porcentaje dentro del género.

Los resultados documentales. Éstos fueron obtenidos también de la agrupación pero se le dio preponderancia a la frecuencia por género, pues con ésta se pudo detectar el origen u orígenes documentales del uso de cada palabra respecto a los 14 géneros en que está subdividido el corpus, lo que hace posible ubicar una palabra con respecto a su clasificación sociolingüística por su uso, nivel y género de la lengua, o incluso puede servir para establecer su ubicación en relación con el área temática del texto analizado.

Como resultado de las operaciones estadísticas y documentales se generó un “índice semiautomático del español usado en México”, el cual cumple con una función práctica para los documentalistas pues les sirve para identificar la información con que cuenta el sistema, controlarla y ofrecerla a los usuarios.

Se elaboraron búsquedas controladas a partir del mismo índice, lo cual exigió adaptaciones de carácter interno, que simplifican la utilización del sistema, lo que

servirá a los usuarios internos y externos para orientar sus búsquedas y visualizar la información con que cuenta el *Corpus del español mexicano contemporáneo* (CEMC).

Este índice servirá a lexicógrafos, terminógrafos y documentalistas en la selección y elaboración de nomenclaturas basadas en uso, frecuencia, dispersión u origen de las palabras. Asimismo, el índice será de mucha utilidad a documentalistas en la indización asistida por la computadora, ya que con éste se podría diferenciar el lenguaje general que muestra el CEMC del especializado analizado en nuevos textos.

Por lo anterior se asume que se cumplieron los supuestos y objetivos propuestos en esta investigación respecto a consolidar este sistema de recuperación de información. El sistema logrado es más eficiente debido a que simplifica, controla y ayuda en las búsquedas de los usuarios en forma más eficiente debido a que se recupera la información documental completa y se visualiza en forma ágil la información recuperada por los usuarios.

Se piensa que esta tesis fomentará el conocimiento y aprovechamiento de la información obtenida mediante la aplicación de la indización por asignación sobre documentos de texto completo que muestran el uso del español de México.

BIBLIOGRAFÍA

Academia Mexicana. Correspondiente de la Española, *Índice de mexicanismos: registrados en 138 listas publicadas desde 1761,* México, Academia Mexicana, CNCA, FCE, 2000, 696 p.

AGUILAR-AMAT, Anna y María José RECODER, *Documentación y construcción de un corpus digital monolingüe sobre astronomía: criterios de clasificación para su aplicación a la investigación y a la docencia de la traducción especializada*, http://www.fti.uab.es/psanchez-gijon/Recerca/Treballs_20de_20recerca/Trecerca99.htm

ALCARAZ VARÓ, Enrique y María Antonia MARTÍNEZ LINARES, *Diccionario de Lingüística moderna*, Barcelona, Editorial Ariel, 1997, viii, 643 p.

ALVAR EZQUERRA, Manuel y Juan Andrés VILLENA PONSODA, *Estudios para un Corpus del Español*, Málaga, Universidad de Málaga, 1994, pp. 31-40.

ANGUIANO PEÑA, Gilberto, *La relevancia de la información bibliográfica en la documentación de un diccionario*, México, El autor, 1991, 196 p Tesis (Licenciado en Bibliotecología) Universidad Nacional Autónoma de México.

_____, *Documentación de palabras. Sistema de información de los recursos documentales del DEM. Hoja tipo WEB* [consulta por Intranet], México, El Colegio de México, Diccionario del Español de México, [2000-]. F:\palabras\PALABRAS-BASE DE DATOS.htm.

_____, *Usuarios del DEM* [base de datos], México, El Colegio de México. Diccionario del Español de México, [2003-]. Documento de consulta interna.

ARANO, Silvia, *La ontología: una zona de interacción entre la Lingüística y la Documentación* [on line]. *Hipertext.net*, núm. 2, 2003. http://www.hipertext.net/web/pag220.htm#La%20lingüística%20y%20la%20documentación.

BECERRA, Javier, “Los servicios del Diccionario del Español de México”, en 4° Simposio de la Asociación Mexicana de Lingüística Aplicada, *Presente y perspectivas de la lingüística computacional en México*, México: 24 al 26 nov. 1987, México, UNAM, 1988.

BRONSOILER FRID, Charlotte, “Indización automatizada”, *Ciencia bibliotecaria,* Vol. 5, núm. 2 (abr. 1982), pp. 85-93.

CORPAS PASTOR, G., “Localización de recursos y compilación de corpus vía Internet: aplicaciones para la didáctica de la traducción médica especializada”, en GARCÍA YEBRA, V. y C. GONZALO GARCÍA, (eds.), *Manual de documentación y terminología para la traducción especializada*, Madrid, Arco/Libros, 2004, pp. 223-506 (Colección Instrumenta Bibliológica).

Diccionario de lingüística, México, Red Editorial Iberoamericana, 1991, xvi, 308 p.

Diccionario de Organización y Representación del Conocimiento. Clasificación, Indización, Terminología. http://eubca1.eubca.edu.uy/diccionario/letra_i.htm.

Diccionario del Español de México, *Corpus del español mexicano contemporáneo, 1921-1974* [cinta magnética], elaborado por García Hidalgo, María Isabel, Luis Fernando Lara, Roberto Ham Chande *et al.*, México, Diccionario del Español de México, 1975. 1 Cinta magnética.

_____, *Corpus del español mexicano contemporáneo, 1921-1974. Lista bibliográfica del corpus,* México, Diccionario del Español de México, 1975, *ca.*, 100 h.

_____, *Proyecto de reglamento para la utilización de los resultados de DEM*, México, Diccionario del Español de México, 1978, 4 h.

Diccionario del español usual en México, Luis Fernando Lara (dir.), México, El Colegio de México, Centro de Estudios Lingüísticos y Literarios, Diccionario del Español de México, 1996, 937 p.; en Biblioteca Virtual Miguel de Cervantes, [1999-],

http://www.cervantesvirtual.com/servlet/SirveObras/04927038278217213276310 3/ y en El Colegio de México http://wodka/Scripts/Dem/principal.htm.

Documentación y construcción de un corpus digital monolingüe sobre astronomía: criterios de clasificación para su aplicación a la investigación y a la docencia de la traducción especializada. http://www.fti.uab.es/psanchez-gijon/Recerca/Treballs_20de_20recerca/Trecerca99.htm.

FOX, Virginia, *Análisis documental de contenido: principios y prácticas*, Buenos Aires, Alfagrama, 2005, 253 p.

FRIDMAN MINTZ, Boris, *Estructura de la base de datos del DEM*, México, Diccionario del Español de México, 1993, 13 h.

GARCÍA EJARQUE, Luis, *Diccionario del archivero bibliotecario: terminología de la elaboración, tratamiento y utilización de los materiales propios de los centros documentales*, Gijón, Asturias, TREA, 2000, xiv, 442, [4] p.

GARCÍA HIDALGO, María Isabel, Jorge SERRANO LIMÓN y Manuel ORONA, *Diccionario del Español de México: programas de análisis automático*, México, Diccionario del Español en México, 1974, 92 h.

García Hidalgo, María Isabel y María Luisa PÉREZ VALDESPINO, *Versión para microcomputadoras IBM del sistema computacional del DEM* [Programa de computadora] México, Diccionario del Español de México, 1990.

GIL LEIVA, Isidoro, "Automatización de la indización", en su *La automatización de la indización de documentos*, Gijón, Ediciones Trea, 1999, pp. 57-85 (Biblioteconomía y administración cultural; 25). http://www.primeravistalibros.com/fichaLibro.jsp?codigo=56.

_____, "Sistema para la Indización Semiautomática (SISA) de artículos de revista de Biblioteconomía y Documentación", en *II Jornadas de Tratamiento y Recuperación de Información*, Madrid, Septiembre 2003, pp. 228-232,

http://personales.upv.es/isgil/SISA%20Demo%20Jotri%202003%20original.pdf.

_____ y José Vicente RODRÍGUEZ MUÑOZ, "De la indización humana a la indización automática", en García Marco, Francisco Javier (ed.), *Organización del conocimiento en sistemas de información y documentación,* Zaragoza, Facultad de Filosofía y Letras, Universidad de Zaragoza, 1997, pp. 201-215.

Glosario ALA de Bibliotecología y Ciencias de la Información, Heartsill YOUNG, con la colaboración de Terry BELANGER [y otros]. Traducción Blanca de MENDIZABAL Allende, Madrid, Díaz de Santos, 1988, xvi, 473 p.

GRANDA, J. Simón, E. de LEMA GARZÓN, "Primeras experiencias sobre el análisis de textos en castellano aplicado a la indexación automática de información", en *Terceras Jornadas Españolas de Documentación Automatizada*, 1990, pp. 1255-1270.

HAM CHANDE, Roberto, *Sistema estadístico* [Fórmulas estadísticas], México, Diccionario del Español de México, 1974.

LANCASTER, Frederick Wilfrid, "El lenguaje natural en la recuperación de la información", en su *Indización y resúmenes: teoría y práctica*, Buenos Aires, EB Publicaciones, 1996, pp. 200-228.

_____, *El control del vocabulario en la recuperación de información*, tr. Alejandro de la Cueva Martín, Valencia, España, Universitat de València, 1995, 286 p.

_____, *Indización y resúmenes: teoría y práctica*, Buenos Aires, EB Publicaciones, 1996, xii, 337 p.

_____ y María PINTO (coords.), *Procesamiento de la información científica*, Madrid, Arco Libros, 2001, 270 p.

LARA, Luis Fernando, "La cuantificación en el Diccionario del español de México", en su *Dimensiones de la lexicografía. A propósito del Diccionario del español de México*, México, El Colegio de México, 1990, pp. 51-84 (Jornadas 116).

_____, *Vocabulario fundamental*, México, El Colegio de México, 1979.

_____, *Resultados numéricos del vocabulario fundamental del español de México*, México, El Colegio de México, Centro de Estudio Lingüísticos y Literarios, Diccionario del Español de México, 2007, 43 p.

_____ y Roberto HAM CHANDE, "Base estadística del Diccionario del Español de México", en su *Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, Centro de Estudio Lingüísticos y Literarios, 1980, p. 15 (Jornadas; 89).

_____, Roberto HAM CHANDE y María Isabel GARCÍA HIDALGO, *Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, Centro de Estudio Lingüísticos y Literarios, 1980, vii, 266 p. (Jornadas; 89).

MOREIRO GONZÁLEZ, José Antonio, *Introducción al estudio de la información y la documentación*, Colombia, Editorial Universidad de Antioquia, 1998, 188 p. (Medios y mensajes).

NAUMIS PEÑA, Catalina, "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, 1997, pp. 23-26.

_____, "Indización y clasificación: un problema conceptual y terminológico", *Documentación de las ciencias de la información* [Indexation and classification: a conceptual and terminologic problem], Vol. 26 (2003) p. 23-40.

_____, *Modelo de construcción de tesauros documentales multimedia: aplicaciones a los contenidos educativos en televisión*, p. 69. Memoria para optar al grado de doctor, http://www.ucm.es/BUCM/tesis/inf/ucm-t25976.pdf.

_____, *Los tesaurus documentales y su aplicación en la información impresa, digital y multimedia*, México, UNAM, CUIB, 2007, 288 p. (Sistematización de la Información Documental / CUIB).

PÉREZ, C., A. Moreno y P. FABER, "Lexicografía computacional y lexicografía de corpus", *Revista de la Asociación Española de Lingüística Aplicada*, Panorama de la Investigación en Lingüística Computacional, volumen Monográfico, pp. 175-214.

PÉREZ HERNÁNDEZ, M. Chantal, "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento", *Estudios de Lingüística Española,* volumen 18, 2002. http://elies.rediris.es/elies18/index.html.

PINTO MOLINA, María, *Análisis documental de contenido: procesamiento de información*, prólogo de T. A. van DIJK, Madrid, Editorial Síntesis, [1996], 158 p. (Ciencias información. Biblioteconomía y documentación; 10).

Real Academia Española, *Corpus de Referencia del Español Actual* (CREA); HTTP://BUSCON.RAE.ES/DRAEI/.

_____, *Corpus Diacrónico del Español* (CORDE), http://www.rae.es/

_____, *Diccionario de la Lengua Española*, vigésima segunda edición, España, Gredos, 2001, en CD-ROM. Versión 1.0.

RUIZ PÉREZ, Rafael, *El análisis documental: bases terminológicas, conceptualización y estructura operativa,* presentación José Ramón PÉREZ ÁVAREZ-OSSORIO, Granada, España, Universidad de Granada. Grupo de trabajo de Información y Documentación de la Comisión Española de Cooperación con la UNESCO, 1992, pp. 19-21.

SLYPE, Georges van, *Los lenguajes documentales de indización: concepción, construcción y utilización en los sistemas documentales*, Traducción del francés Pedro Hípola [y] Félix de Moya. Madrid, Fundación Germán Sánchez Ruipérez, 1992, 198 p.

SOTO, Susana, "La recuperación de la información: ¿Lenguaje natural *vs*. Lenguaje controlado?", en *Seminario Dilemas de la Biblioteca Actual*, pp. 1-7, Buenos Aires. http://eprints.rclis.org/archive/00006029/; http://eprints.rclis.org/archive/00006029/01/semdilemas-soto.pdf.

UREÑA LÓPEZ, L. Alfonso, *Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos,* Alicante, España, Club Universitario, 2002, 156 p.

VILLAR FLECHA, José Ramón, *Sistema soporte para la clasificación de documentos de texto utilizando razonamiento basado en casos*, [León], Universidad de León, Secretariado de Publicaciones, 2005, 165 p.

Wikipedia. *La enciclopedia libre.* http://es.wikipedia.org/wiki/Corpus_ling%C3%BC%C3%ADstico.