1
2 y

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Filosofía y Letras
Colegio de Bibliotecología

LA RELEVANCIA DE LA INFORMACIÓN BIBLIOGRÁFICA EN LA DOCUMENTACIÓN DE UN DICCIONARIO

T E S I S

QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN BIBLIOTECOLOGÍA

P R E S E N T A

GILBERTO ANGUIANO PEÑA

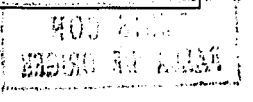
México, D.F., 1991

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

PA433
A54

Anguiano Peña, Gilberto.
La relevancia de la información bibliográfica en la documentación de un diccionario / Gilberto Anguiano Peña. -- México, D.F. : G. Peña A., 1991. -- 196 p.
Tesis (Licenciado en Bibliotecología) -- Universidad Nacional Autónoma de México, Facultad de Filosofía y Letras, Colegio de Bibliotecología, 1991.

1. Lexicografía—documentación. I. Universidad Nacional Autónoma de México. Colegio de Bibliotecología.



ÍNDICE

PRÓLOGO

La tesis que se presenta en las siguientes páginas nace de la necesidad de cumplir con los requisitos que establece la Universidad Nacional Autónoma de México para conceder el título de licenciado en bibliotecología. Ante esta situación y esperando resolverla satisfactoriamente, de acuerdo con mis posibilidades, preparación y capacidades personales, he seleccionado el tema *La relevancia de la información bibliográfica en la documentación de un diccionario*, ya que con éste puedo organizar y estructurar las experiencias y los conocimientos adquiridos en la carrera de bibliotecología, en las actividades realizadas en el Centro de Documentación del Instituto Nacional para la Educación de los Adultos y en mi colaboración como becario de investigación en las actividades de documentación que se realizan en el Diccionario del Español de México, de El Colegio de México.

Así pues, fue precisamente al realizar las funciones establecidas para documentar con información bibliográfica el uso y significado de las palabras solicitadas por los redactores del diccionario para su posterior análisis lexicográfico, como logré identificar algunos fenómenos bibliográficos que se encuentran asociados con fenómenos sociales del habla de los mexicanos, una vez que éstos son procesados documentalmente y luego obtenidos por un sistema de recuperación, en donde la información se encuentra almacenada y organizada temáticamente.

La presente investigación parte básicamente de los conocimien-

tos pertenecientes a tres áreas del conocimiento, las cuales son la bibliotecología, la documentación y la lexicografía con el objeto de demostrar la asociación que se da entre la información considerada como pertinente en la realización de diccionarios y las distribuciones bibliométricas que se pueden considerar como modelos del comportamiento de la información bibliográfica. Esto se efectuó con la intención de que los resultados sirvan como punto de apoyo en la comprensión de algunos fenómenos bibliográficos existentes en la lexicografía, lo que permitirá, además, prever y pronosticar el comportamiento de la información bibliográfica usada en dicha disciplina, para su mejor control y aprovechamiento.

INTRODUCCIÓN

Como se afirma en el prólogo, en esta investigación se estudian principalmente los conceptos que se estructuran bajo el título de la tesis, los cuales son: la relevancia, la información bibliográfica, la documentación y los diccionarios. Todos éstos son organizados de tal manera que sirven para esclarecer un problema específico como en el que se encuentra involucrada la información bibliográfica como un fenómeno social. El problema consiste en demostrar la asociación existente entre la información juzgada como "pertinente", que es la que satisface las necesidades de información de los usuarios de un sistema de recuperación de información, y las distribuciones bibliométricas que se pueden considerar como modelos del comportamiento de la información bibliográfica.

En la primera parte del estudio, utilizando el planteamiento de la teoría de la relevancia como el marco que contiene la teoría esencial de esta tesis, se trata de explicar lo que es la *relevancia* para este estudio, al describir los diferentes factores que intervienen en ella y los fenómenos que la afectan, con lo cual se llega a la conclusión de que la relevancia es la ponderación que se realiza sobre las respuestas obtenidas por un sistema de recuperación, esto en cuanto a que satisfagan la necesidad que originó la formulación de una consulta. En otras palabras, se busca que en el proceso de comunicación, la información que se transmite de un archivo fuente sea comunicada de manera efectiva a un archivo destinatario. Para determinar dicha efectividad tiene que existir una noción de relevan-

cia, que puede ser una idea o conocimiento elemental de lo que debe satisfacerse con una respuesta, o también puede existir un juicio de relevancia en el que se utilicen criterios bien determinados de lo que se pretende encontrar en la información buscada.

Bajo estas consideraciones, en la segunda parte del estudio se intenta comprobar con una experiencia real lo que la teoría plantea, por lo que se opta por identificar la información que resultase relevante en la elaboración de un diccionario, en cuya realización desempeño las actividades de documentación. Para esto resulta indispensable explicar las relaciones que existen entre la documentación y la lexicografía, ya que ésta es la disciplina en que se orienta la creación de la obra de consulta llamada diccionario. Para esta explicación se necesita describir lo que es un diccionario y lo que se puede entender en su proceso de elaboración como *documentación*. Se deriva de aquí mismo la necesidad de explicar lo que es la información que se utiliza en los diccionarios, dónde se halla esta información, la técnica de análisis que se utiliza para recuperarla y las características que deben tener los sistemas de recuperación que almacenan esta información.

Una vez aclarado lo anterior, se utiliza la experiencia del Diccionario del Español de México (DEM) para describir por medio de éste, a manera de ejemplo, los procedimientos documentales que se efectúan en la elaboración de diccionarios y el tipo de sistemas de recuperación de información que se utilizan para obtener información pertinente para sus labores.

En la tercera parte del estudio se describen los fundamentos de los juicios que se siguen en la selección de vocabulario en los diccionarios, y se elige de entre éstos el criterio de *distribución de unidades léxicas en diferentes textos utilizados*, para que funja en esta tesis como la noción de pertinencia que sirva para evaluar la información obtenida por el sistema de recuperación de información del DEM.

En la cuarta parte se procede a identificar la relevancia de la información bibliográfica en la documentación de un diccionario a

partir de la citada noción de pertinencia, basada en un principio de selección de vocabulario para diccionarios. Para detectar lo anterior se procedió a ponderar la información a partir del esclarecimiento de las necesidades que deberían de satisfacerse por parte de las respuestas obtenidas por el sistema del DEM. Posteriormente se delimitaron las preguntas que se formulan en forma de consultas, y se procede a realizar las búsquedas correspondientes. Luego, se hace el análisis de las respuestas obtenidas del sistema, utilizando como estrategia de búsqueda de la pertinencia, la operación de conjuntos denominada "intersección", con lo que resultan hasta nueve niveles de jerarquía decreciente, en los que se pondera la información a partir de la pertinencia, la relevancia, la no relevancia y la información no recuperada.

Por último, se establecen las asociaciones entre la información pertinente recuperada por el sistema del DEM, y las distribuciones bibliométricas que se desarrollan en esta misma información recuperada, esto al identificarse, por medio de un cuadro invertido de distribución de frecuencias basado en la Ley de Zipf, el núcleo de palabras donde se encuentra la información con mayor pertinencia. Al tener en cuenta la frecuencia en las respuestas, se establece la probabilidad promedio que tiene una palabra de llegar a ser pertinente en las búsquedas que se realizan en el sistema de recuperación del DEM, esto si existe como noción de pertinencia el criterio de selección de vocabulario llamado *distribución de unidades léxicas en diferentes textos utilizados.* De lo anterior, la conclusión más destacada en lo relativo a las asociaciones entre la información pertinente y las distribuciones bibliométricas es que, al incrementarse en una respuesta su frecuencia de aparición, se incrementa la probabilidad de que dicha información sea pertinente, e incluso se detalla hasta en el porcentaje de probabilidades que tiene una información para que en una búsqueda llegue a ser pertinente, si se toma como base su frecuencia y el lugar que ésta ocupe en el cuadro invertido de distribución de frecuencias aplicado en esta tesis.

Se concluye que los resultados de esta investigación se pueden sumar a la literatura empírica y descriptiva existente en la bibliotecología, en la que se esclarece la pertinencia y la relevancia de la información bibliográfica a partir de su existencia en textos ordenados temáticamente.

PRIMERA PARTE

ANTECEDENTES TEÓRICOS DEL ESTUDIO

UNA APROXIMACIÓN A LA RELEVANCIA

Para tratar el tema de la relevancia, antes hay que decir que el reconocimiento por parte de los científicos y la sociedad en general respecto al valor que tienen los conceptos *conocimiento* e *información* en la solución de todo tipo de problemas, ha convertido en requisitos esenciales para el progreso tanto a la comunicación efectiva del conocimiento como a la existencia de los sistemas de recuperación de información, ya que ambos le permiten al hombre tomar decisiones para así poder actuar en la solución de un sinnúmero de problemas.

Los sistemas de recuperación de información bibliográfica que se utilizan en distintas disciplinas se han desarrollado para facilitar la solución de problemas sobre el manejo y el aprovechamiento de la información, por eso mismo casi todo sistema de este tipo busca incrementar la efectividad de su comunicación en sus servicios con el propósito de recuperar información relevante para sus usuarios.

Es necesario entonces aclarar que en los estudios sobre la recuperación de información —que se puede entender como las acciones, métodos y procedimientos utilizados para recuperar datos que proporcionan información sobre un tema determinado— existe el

término especializado *relevancia de la información*, el cual indica la recuperación real de información bibliográfica, obtenida por un sistema de recuperación como respuesta a una pregunta específica formulada por un usuario.

También se puede distinguir en esta clase de estudios el término *pertinencia de la información*, que delimita de entre la información relevante a aquella seleccionada por el usuario del sistema de recuperación, bajo la condición de que dicha información satisfaga o dé solución a las exigencias y requisitos determinados por el usuario en su propio criterio de búsqueda.

Otro término que resulta importante conocer es el de *porcentaje de relevancia*, el cual expresa la proporción de información pertinente recuperada que da respuesta a una pregunta determinada, respecto del total de documentos recuperados bajo el criterio de búsqueda utilizado por un usuario.

En un sentido más estricto, la relevancia tiene que relacionarse con la efectividad de la comunicación y existe implícito un juicio o noción de relevancia en todo sistema de información, que hace posible interpretar su eficiencia; esto mismo nos permite entender por qué el significado de la relevancia tiene una gran variedad de enfoques, según sean los sistemas de información que les den origen; sin embargo, el objetivo terminal de la relevancia se centra en una comunicación efectiva entre la información proporcionada por un sistema, y un usuario.

El proceso de comunicación y la relevancia

El proceso de comunicación ayuda a entender mejor los diferentes enfoques que, sobre la relevancia de la información bibliográfica, existen vinculados a los estudios en bibliotecología, por lo que resulta oportuno destacar ciertos aspectos teóricos de este proceso que dan fundamento a este concepto.

Según Goffman,[1] la comunicación es un proceso en el cual algo llamado información se transmite de un organismo a otro. Al primero le llamó fuente o emisor y al segundo receptor. Entre el emisor y el receptor a veces puede ocurrir una retroalimentación dinámica e interactiva, en la que pueden invertirse los papeles.

Tratando este mismo aspecto, Shannon[2] afirmó que el proceso de comunicación requiere de cinco elementos: 1) una *fuente de información*, que es la que da origen al mensaje; 2) un *transmisor*, el cual usualmente codifica o modifica el mensaje, facilitando así su envío; 3) un *canal*, que es el medio por el cual los mensajes son transmitidos y el cual puede llegar a introducir ruido en el sistema; 4) un *receptor*, el cual decodifica o interpreta la forma del mensaje; y 5) un *destinatario*, que recibe la información.

Saracevic,[3] en este contexto del proceso de comunicación, considera a la relevancia como la medida de efectividad del contacto entre una fuente o emisor y un receptor.

El conocimiento que se comunica en el proceso aludido, es entendido en las ciencias de la información[4] como *información*, es decir, que lo que se transmite contiene: conocimientos, datos, información u objetos. Haciendo una generalización a partir de lo anterior, se podría decir que la literatura de una materia representa el conocimiento de la misma y que, por extensión, las palabras contenidas en textos o materiales bibliográficos, ordenados temáticamente, representan el conocimiento de los hablantes sobre el asunto del que hablan. Este conocimiento o información se encuentra en forma de palabras que a su vez están contenidas en textos, y son las que se comunican por medio de los sistemas de recuperación utilizados para resolver problemas de información, como sucede en la elaboración de diccionarios, aspecto que se estudiará en el desarrollo de esta tesis al argumentarse que la relevancia es la medida de efectividad del contacto entre la información proporcionada por un sistema de recuperación y un usuario del mismo.

Revisión de los puntos de vista sobre relevancia y pertinencia

Antes de proceder al estudio propuesto en la introducción, es oportuno profundizar sobre la noción de relevancia y sus interpretaciones más destacadas.

El primero en utilizar el término "relevante" fue Bradford,[5] al referirse a los artículos "relevantes de una materia", aproximadamente en los años treinta.

En el periodo de los años cuarenta y cincuenta, Mooers,[6] Taube[7] y Perry,[8] fueron los primeros en hablar sobre la relevancia de la información en relación con la recuperación real de información que les ofrecida por un sistema a su receptor: les preocupaba más la abundante recuperación de información no relevante, y ésta fue la causa principal que consideraron para realizar ajustes en los sistemas de información.

En 1958, durante la Conferencia Internacional sobre Información Científica, fue discutido minuciosamente el concepto de relevancia por los especialistas de la materia.[9]

En la década de los sesenta, aparecen varias definiciones de relevancia, principalmente por el hecho de que las definiciones elaboradas hasta entonces no eran muy claras en cuanto al significado del término.[10] Muchas de las definiciones de entonces se hacían parafraseando las definiciones anteriores de relevancia en una tentativa para explicar las relaciones entre éstas y los usuarios, los documentos y los sistemas de información.

Los científicos de la información, al especificar los factores, elementos y relaciones que la noción o juicio de relevancia comprende, formularon y desarrollaron muchos puntos de vista acerca de ella. Varios términos fueron propuestos por diversos autores para identificar diferentes aspectos de la relevancia: *pertinencia, relevancia del sistema, relevancia para el usuario*, etc., fueron éstos y más los términos conceptualizados para el sentido de relevancia.

Uno de los puntos de vista más destacados fue el de Saracevic,[11]

quien argumentó que la relevancia se podía tener desde el punto de vista de un sistema, desde el destino de la información, y a partir de la literatura temática.

Polushkin[12] afirmó que la relevancia es una característica del grado de correspondencia o correlación entre un documento recuperado y el contenido de una pregunta. Con respecto a la pertinencia, la definió como la característica del grado de correlación entre el contenido de un documento recuperado y la satisfacción a la solicitud de información, dándose esto último conjuntamente con las características puramente subjetivas de un usuario específico.

Saracevic[13] definió la relevancia como una medida de eficacia del contacto entre una fuente o emisor y un destinatario o receptor, durante el proceso de comunicación, y destacó además que el concepto de necesidad de información da origen a la noción de pertinencia.

Foskett[14] sugirió que la verdadera diferencia en el concepto de relevancia consiste en el sentido que ésta tiene respecto a un campo/materia/universo, y que está delimitada por los términos de la pregunta efectuada. En cuanto a la pertinencia, la definió como las respuestas que significan información nueva para un usuario, distinta a la ya existente en su mente, siendo ésta útil para el trabajo que desempeña. Este resultado es el requerido por el usuario y también lo que lo motiva a formular preguntas o consultas.

Lancaster[15] sugirió también una distinción entre la relevancia y la pertinencia. Distinguió la pertinencia como la relación entre un documento y la necesidad de información del usuario. Definió la relevancia como la relación entre un documento y una pregunta.

Harmon[16] distinguió la relevancia de un usuario de la del sistema de recuperación, donde la *relevancia del usuario* puede ser vista como la evaluación hecha por el mismo usuario a partir de la relación efectuada entre la pregunta bien realizada al sistema y sus propias necesidades de información. La *relevancia del sistema* puede ser vista como la evaluación hecha por el sistema de recuperación res-

pecto al grado de relación existente entre la información almacenada en el sistema y la pregunta formulada por el usuario.

Wilson[17] intentó explicar la relevancia no como una única noción sino como muchas. Definió lo que llama "relevancia situacional", como aquella que se da en una situación individual y específica, situación en la que el individuo ve a la relevancia desde un punto de vista particular y no como otros la ven o como puede ser realmente. Argumentó también que existe relevancia en relación con el grado de conocimientos de un individuo, relevancia que cambia en la medida en que ese conocimiento cambia.

Cuadra y Katter[18] destacaron también el aspecto "situacional" de la relevancia. Afirmaron estos autores que los juicios de relevancia habían sido tratados como si fuesen una "caja negra", cuyo contenido no tuviera interés, a no ser en cuanto que la caja negra desempeñase las funciones para las que fue creada. Para estos autores, la relevancia es más que un simple contagio de puntos y puede ser afectada por un sinnúmero de variables.

Experimentos realizados con juicios de relevancia

Los diversos experimentos realizados sobre la relevancia se han venido practicando de acuerdo a cinco clases de variables como lo ha expuesto Saracevic:[19] 1) documentos y representación de documentos, 2) preguntas o consultas, 3) situaciones y condiciones de los juicios, 4) formas de expresión y 5) condiciones humanas.

Para ejemplificar estas variables a continuación se describen algunos estudios, donde el interés principal era determinar la relevancia en los documentos y sus representaciones, encontrar la consistencia de los juicios sobre la relevancia y verificar la congruencia de estos juicios en relación directa con la medida en que éstos eran afectados por ciertas condiciones humanas.

Gifford y Baumanis[20] describieron una investigación empírica

sobre el papel que desempeñan los documentos en el enjuiciamiento de la relevancia. Su estudio se basó en una pregunta: ¿Qué es lo que en un documento provoca el juicio de un usuario? Esta pregunta fue directamente analizada, principalmente por la determinación que hacen las diferencias textuales significativas que pueden ser encontradas entre los documentos relevantes y los no relevantes. En la muestra examinada por estos autores, los documentos relevantes se distinguieron con regularidad de los no relevantes, no obstante, en la relación general de relevancia, es decir, lo que hace a un documento relevante de uno no relevante, ninguna evidencia de atributos generales fue encontrada. Los documentos relevantes para una pregunta específica generalmente tuvieron propiedades significativas, pero esto no se extendió de una pregunta a otra.

Tague[21] intentó clasificar las relaciones existentes entre las palabras presentes en las preguntas y las palabras para indizar documentos, y llegó a determinar la existencia de respuestas relevantes, relativamente relevantes y no relevantes. Aquí el autor concluyó que: 1) las palabras de los títulos son las más efectivas en términos de la evaluación; 2) las palabras contenidas en las preguntas son más útiles para localizar respuestas relevantes; 3) la frecuencia de las palabras utilizadas al formular una pregunta se presentan en mayor cantidad en documentos considerados relevantes; 4) el número de palabras obtenidas como respuestas en los documentos es mayor que las palabras que componen la pregunta original, caracterizándose así las respuestas relevantes.

Saracevic[22] en otro estudio, examinó aspectos relacionados con la evaluación de sistemas de recuperación, estudiando por un lado a los usuarios y sus preguntas y por otro a los juicios de la relevancia basados en distintos formatos de salida del sistema. Las medidas de evaluación aplicadas revelaron que hay una mayor coincidencia en lo que no es relevante que en lo que sí lo es. El autor también concluyó que los sistemas no atribuyen mayor relevancia que la que puede atribuirle un usuario, pero que pueden tener un desem-

peño tan bueno como el del usuario. Por último, afirmó que la utilización de diferentes formatos por parte de los responsables de los sistemas tienen un efecto significativo sobre los juicios de relevancia que emite un usuario.

Rath *et al*[22] realizaron un estudio comparativo entre cuatro tipos de indicadores de contenido léxico: títulos, dos tipos de resúmenes y textos. Concluyeron los autores que el uso de títulos sin resúmenes adicionales lleva a juicios de relevancia discordantes, y que no existe una diferencia significativa en las evaluaciones cuando se usan textos completos y resúmenes.

O'Connor[24] condujo diversos estudios para explicar los desacuerdos entre los juicios de la relevancia y los factores con los que están relacionados o que son los causantes de esos desacuerdos. El autor decidió que las causas básicas de estos desacuerdos sobre la relevancia se deben a las distintas interpretaciones sobre las solicitudes de información y los documentos recuperados, y no a factores tales como la educación, la experiencia, la función, etc., de los enjuiciadores.

Barhydt[25] realizó un estudio para testimoniar la eficiencia de los juicios de relevancia de los no usuarios sobre las preguntas y las respuestas en un centro de documentación. Concluyó que el usuario es el mejor árbitro sobre la relevancia de los documentos y que el especialista de una disciplina no es mejor *a priori* que el responsable del sistema de recuperación para evaluar documentos.

Resnick y Savage[26] desarrollaron un experimento controlado para comparar la relativa consistencia humana inter-intra-asunto, basándose en las citas, resúmenes, términos de indización y textos completos. Bajo las condiciones del experimento, los juicios fueron consistentes. Sin embargo, resultó un problema cuantificar los resultados basándose en los juicios humanos.

Conclusiones

Después de esta revisión se puede reconocer que existen tantos elementos en la comunicación del conocimiento (información) y su posible contacto efectivo con los usuarios, que la relevancia puede considerarse desde diferentes puntos de vista; éstos se pueden presentar en conjunto o por separado, en varias combinaciones y con diferentes jerarquías.

En general, la relevancia puede incluir factores y relaciones que a su vez se pueden considerar en varios aspectos:

1. Conocimiento de la materia.
2. Cualquier representación lingüística o simbólica del archivo del emisor o del archivo del receptor, o de los procesos.
3. Emisor, especialmente el o los archivos.
4. Receptor, especialmente el o los archivos.
5. Sistemas de información orientados a facilitar, mejorar, preservar o extender el proceso de comunicación.
6. Medio ambiente, realidades y funciones.
7. Reflejo de algún sistema de valores humanos (sociales, lingüísticos, etcétera).

Los factores que afectan a la relevancia son, de hecho, los juicios o criterios otorgados por los usuarios, lo que apoya el concepto de que la relevancia es una medida o evaluación, por lo que resulta también que es una relación y ésta es en todo caso relativa, ya que de manera general se considera la relevancia como dependiente de lo que ya se sabe o conoce, es decir, una información puede ser más relevante que otras que son consideradas irrelevantes, bajo determinadas circunstancias, en la medida en que el usuario conozca o no el tema del que se está informando. De lo anterior se puede interpretar que la noción de relevancia está sujeta a un tiempo y un espacio determinado, pues lo que ahora se considera como relevante, muy probablemente, al cambiar las condiciones que lo hacen enjuiciarlo así, dejará de serlo.

También conviene recordar lo que enuncia Saracevic,[27] respecto a que existirá una noción de relevancia cuando se desee un contacto productivo o efectivo en la comunicación —sea de manera consciente o no—, es decir, cuando se realice un cambio en un archivo receptor por medio de la información que se transmite de un archivo fuente, y esta noción será la medida de esos cambios. Los cambios podrán ser adiciones, eliminaciones o la reorganización de los mismos archivos, pero en cualquier caso deberán ser observables.

A partir de la ejemplificación de los experimentos realizados con los juicios de la relevancia, se puede argumentar que al aplicarse un criterio en la selección de información recuperada por un sistema, es posible hacer que los juicios sobre la relevancia sean consistentes, y, que al preestablecerse los requisitos y necesidades que el usuario debe cubrir con información, se podrá también establecer una ponderación referida al grado de satisfacción contenido en las respuestas recuperadas por un sistema.

Advertencia

Antes de terminar se puede argumentar que en la bibliotecología, como en cualquiera otra disciplina, al tratar de explicarse la realidad de los fenómenos que le competen, el especialista tiene necesidad de recurrir a los conocimientos esenciales ya existentes y directamente relacionados con el tema que se pretenda investigar; esta razón ha venido orientando la revisión y el análisis por parte de los bibliotecólogos sobre diversos conocimientos producidos en otras áreas. Esto mismo sucede en el presente estudio, en el que la utilización de ciertos conocimientos producidos por la lexicografía tiene como propósito enmarcar y explicar mejor un fenómeno real de la "recuperación de información". Dicho fenómeno se puede reconocer como *la relevancia de la información bibliográfica en el proceso de la documentación de un diccionario.*

[1] Goffman, W. "A general theory of communication", pp. 726-747, cit. por Saracevic, Tefko. *Introduction to information science*. New York, 1970.

[2] Shannon, Claude E. y Warren Weaver. *Matematical theory of communication*. Urbana, Illinois, University of Illinois Press. 1949.

[3] Saracevic, Tefko. *Relevancia: una reseña y una estructura para considerar el concepto en ciencia de la información*. México, Asociación de Bibliotecarios de Instituciones de Enseñanza Superior, 1978, 72 h., (Cuadernos del ABIESI; 7).

[4] La ciencia de la información es entendida por Saracevic, Tefko. *Op. cit.*, como un campo y una materia que se relaciona con los problemas que emanan de la comunicación del conocimiento general y con los registros en esa comunicación en particular, desde los puntos de vista técnicos y de aplicación. La bibliotecología, la documentología y la archivología comparten estos mismos intereses, éstos son el punto de relación y acercamiento entre todas estas materias.

[5] Bradford, S. C. "Sources of information on specific subjects", pp. 85-86, en *Engineering*, 137 (1934).

[6] Mooers, C. S. "Coding, information retrieval, and the rapid selector", pp. 225-229, en *American documentation*, vol. 1, núm. 4 (1950).

[7] Taube, M., *et al.* "Storage and retrieval of information by means of the association of ideas", pp. 1-17, en *American documentation*, vol. 6, núm. 1 (1950).

[8] Perry, J. W. "Suerimposed punching of numerical codes on Hansorted, punch cards", pp. 205-219, en *American documentation*, vol. 2, núm. 4 (1951).

[9] National Academy of Sciences. *Proceedings of the International Conference on Science Information*, Washington, National Academy of Sciences, 1959, 2 vol.

[10] Existen constantes confusiones sobre la relevancia, especialmente en pruebas y evaluaciones, en las que se le considera como una medida en el proceso de comunicación, con respecto a la efectividad del contacto entre el emisor y un receptor. El problema tiene su origen al no distinguir-

se si la relevancia es una medida, un instrumento de medida o la acción de la medición.

[11] Saracevic, Tefko. *Op. cit.* pp. 16-17.

[12] Polushkin, J. A. "Relevance and pertinence", pp. 52-54, en *Automatic documentation in mathematics linguistics*, vol. 7, núm. 1 (1973).

[13] Saracevic, Tefko. *Op. cit.* p. 14.

[14] Foskett. D. J. "A note on the concept of relevance", pp. 77-78, en *Information storage retrieval*, vol. 8, núm. 2 (apr. 1972).

[15] Lancaster, F. W. "Some notes on the distinction between pertinence an relevance", en su *Guidelines for the evaluation of information systems and services*, [s.p.i], 1977, Preparado para la Unesco bajo contrato.

[16] Harmon. G. *Op. cit.*

[17] Wilson, P. "Situational relevance", pp. 457-471, en *Information storage retrieval*, vol. 9, núm. 8 (1973).

[18] Cuadra. C. A. y Kartter, R. V. "Opening the black box of relevance", pp. 291-303, en *Jounal of documentation*, vol. 23, núm. 4 (1976).

[19] Saracevic, Tefko. *Op. cit.* p. 18.

[20] Gifford, C. y Braumains, G. J. "On understanding user choices: textual correlates on relevance judgements", pp. 21-26, en *American documentation*, vol. 20, núm. 1 (jan. 1969).

[21] Tague, J. "Matching of questions and answer terminology in an educational research file", pp. 26-32, en *American documentation*, vol. 16, núm. 1 (1965).

[22] Saracevic, Tefko. "Selected results from in inquiry into testing of IR systems", pp. 126-139, en *Journal of American Society of Information Science*, vol. 32, núm. 2 (apr. 1971).

[23] Rath, G. J., *et al.* "Comparisons of four tipes of lexical indicators of the content", pp. 126-130, en *American documentation*, vol. 12, núm. 2 (1961).

[24] O'Connor, J. "Relevance dissagreements and unclear request forms", pp. 165-177, en *American documentation*, vol. 18, núm. 3 (1967).

[25] Barhydt, G. C. "The effectiveness of non user relevance assessments", pp. 146-149, en *Jounal of documentation*, vol. 23, núm. 2 (1967).

[26] Resnick, A. and Savage T. R. "The consistency of human judgements of relevance", en *American documentation*, vol. 15, núm. 2 (1964)

[27] Saracevic, Tefko. *La relevancia... Op. cit.* pp. 12.

SEGUNDA PARTE

LA INTERRELACIÓN ENTRE DICCIONARIOS Y DOCUMENTACIÓN

La documentación en los diccionarios

Como se argumentó previamente en este estudio, se puede interpretar la relevancia desde distintos puntos de vista y es evidente que el punto de vista con mayor trascendencia para la bibliotecología está supeditado a la comunicación efectiva de la información bibliográfica que se da entre las fuentes, como pueden ser de manera concreta un sistema de recuperación de información, y los destinatarios o usuarios del sistema.

Ahora bien, en este estudio, donde se explica el fenómeno bibliográfico y particularmente la relevancia que tiene la información bibliográfica en la documentación de un diccionario, resulta esencial explicar el contexto donde se describe y analiza este fenómeno; por lo mismo, en esta segunda parte se describen específicamente los factores que influyen de manera directa en el tema estudiado, los cuales están bien determinados en el proceso documental que se desarrolla para elaborar la obra de consulta que nos ocupa. Así, se explica lo que es un diccionario de lenguaje natural y el tipo de información utilizada en él, donde se destaca el papel que desempeña la "palabra", como objeto que se ha de recuperar por los

sistemas bibliográficos. También se aclara por qué resulta más útil, en la elaboración de diccionarios, realizar la recuperación de las palabras por medio de la indización en lenguaje natural que por medio de la indización en lenguaje documental, esto al efectuarse una comparación entre ambos lenguajes. Así mismo, se habla ampliamente de lo que es la indización, por ser ésta la técnica de análisis más adecuada para recuperar y proporcionar información pertinente para los intereses de los responsables de hacer diccionarios. También se explica qué es, cómo es, cuáles son las metas, los objetivos y procedimientos en el Diccionario del Español de México, ya que este proyecto especial que se realiza en El Colegio de México se utiliza a manera de ejemplo para entender mejor los procedimientos documentales que se practican en la elaboración de los diccionarios modernos. Esto último sirve también para explicar que en este proyecto se cuenta con abundantes recursos bibliográficos, tanto manuales como automatizados, los cuales deben ser conocidos por los bibliotecólogos para entender mejor la interrelación que existe entre los diccionarios y la documentación.

Los diccionarios

La lexicografía es la disciplina que se encarga de elaborar diccionarios y está estrechamente relacionada con la lexicología; al respecto, Julio Casares[1] hace una diferenciación de ambas disciplinas. Para él, "la lexicología es el estudio científico, del origen, forma y significado de los términos", y la lexicografía "consiste en el arte de componer diccionarios". Aunado a estas definiciones, los especialistas de la materia argumentan el hecho de que no se puede pensar en la una sin la otra.

Para los lexicógrafos, la tarea básica es hacer diccionarios, y los han hecho de muchas maneras y para muy diferentes usos. Para esto emplean distintos niveles de lengua, dando mayor importancia

a ciertas entradas, seleccionando la definición más importante, etc., lo que ha llevado a la existencia de gran variedad de diccionarios:[2] los normativos, como el de la Real Academia Española; los de arcaísmos, como *el Vocabulario medieval castellano*; de frecuencias, como el *Diccionario estadístico del español de México*, aun no publicado; los inversos, como el *Diccionario inverso de la lengua española*; de técnicas, como el *Diccionario castellano con las voces de ciencias y artes y sus correspondientes en las tres lenguas, francesa, latina e italiana*; de oficios, como el *Diccionario general del periodismo*; de palabras cruzadas, como *The crossword anagram dictionary*; de vocablos actuales, como el *Dictionaire des mots contemporaines*; de una o más de una lengua (monolingües, bilingües, multilingües); o los automatizados, como los contenidos en discos compactos.

Especialmente para los mexicanos hispanohablantes, resultan importantes los diccionarios monolingües de tipo general que hacen un inventario exhaustivo del léxico, o conjunto de las palabras pertenecientes a la lengua española, la cual se utiliza casi en toda Latinoamérica. De entre este tipo de diccionarios monolingües se pueden destacar: el *Diccionario de la Real Academia Española*, de tipo normativo; el *Diccionario de uso del español*, que es muy parecido al anterior; el *Diccionario ideológico de la lengua española*, que funciona como si fuera un *thesaurus*; el *Diccionario de construcción y régimen de la lengua castellana*; el *Diccionario crítico etimológico de la lengua castellana*, que es de tipo histórico; y, para el caso particular de México, el *Diccionario del español de México*, que es de tipo descriptivo y regional, en el cual se basó este estudio.

Los diccionarios[3] no son glosarios ni vocabularios ni enciclopedias; de hecho son obras de consulta generalmente en forma de libro, que reúnen palabras para definirlas, explicarlas o traducirlas, según sea el tipo de diccionario que se pretenda hacer. Las palabras, o léxico, que contiene un diccionario, pertenecen a una lengua determinada, a cierto grupo social, a cierto trabajo o especialidad, a cierta región: a un habla individual; al léxico infantil, léxico

tabasqueño, léxico del hampa, léxico electrónico, léxico elegante. En esta obra de consulta las palabras son generalmente ordenadas alfabéticamente y están acompañadas por sus definiciones.

Si se acepta que un diccionario es un libro, aunque ya los hay automatizados contenidos en CD-ROM (Compact Disk-Read Only Memory), aceptamos que también éste puede ser objeto de estudio de la bibliotecología, cosa que resulta aún más adecuada para la investigación que en esta tesis se realiza, acerca de la relevancia de la información bibliográfica en la documentación de un diccionario.

Pero antes de proceder a iniciar la investigación mencionada, es necesario recordar ciertas particularidades de la obra de consulta donde se produce el fenómeno bibliográfico que nos ocupa, por lo que se puede iniciar la descripción de diccionario argumentando que si es un libro, no es un objeto natural, es decir, es un objeto cultural y tiene su origen cuando aparece la escritura y se puede de esta manera registrar la cultura de las palabras, lo cual es a su vez un fenómeno lingüístico; así el diccionario se convierte en el depósito de la memoria del conocimiento lingüístico de la sociedad.

El lexicógrafo, quien es el que hace los diccionarios, es el mediador entre el diccionario y el usuario al ser el responsable de hacer objetivas las realizaciones del habla y traducir este conocimiento lingüístico para el uso de la sociedad.[4]

Ésta por su parte, al reconocer que en el diccionario está depositada su propia memoria del conocimiento lingüístico, provoca que esta obra se constituya en un objeto público, al que se le atribuye veracidad, autoridad y normatividad.

La documentación e información utilizada en la elaboración de diccionarios

Tanto la bibliotecología como la documentación y la ciencia de la información, tienen como objeto de estudio la información y su re-

levancia, por considerarla como un factor clave en la resolución de los problemas que se presentan en la comunicación, ya que la efectividad del contacto entre el emisor y el receptor en el proceso de comunicación puede ser medida por la relevancia.

"Conocimiento" e "información" son conceptos imprescindibles en nuestros días, lo que hace claro que dentro de los sistemas de información se considere la efectividad en la comunicación del conocimiento como un requisito esencial para resolver los problemas que afectan al hombre. En la lexicografía, el hombre ha creado sus sistemas de recuperación de información, donde el objetivo es tener la mayor cantidad posible de información léxica documentada que sirva como testimonio o como prueba del conocimiento claro y seguro del uso de las palabras, cosa que permite a los lexicógrafos la elaboración de diccionarios.

La documentación en la disciplina lexicográfica está considerada como parte del conjunto de técnicas y criterios aplicados para la elaboración de léxicos y diccionarios. El principal objetivo de la documentación es asegurar el acceso a las fuentes (primarias y secundarias) que contienen los fragmentos de conocimiento o información con que trabaja el lexicógrafo, las cuales son principalmente los textos o materiales bibliográficos, que contienen lenguaje natural.

Actualmente, la realización de diccionarios está vinculada con las técnicas y criterios documentales, que básicamente son: la selección, la adquisición, la clasificación, el análisis y la recuperación de documentos. Esto se puede traducir en que los diccionarios tienen un sistema de información muy particular, ya que la unidad de información utilizada es la palabra.

En su tarea, el lexicógrafo tiene necesidad constante de la mayor 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 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.

Resulta oportuno describir aquí lo que se puede entender por documento. Éste, en el sentido más amplio de su significado, es cualquier material registrado gráficamente, cualquiera que sea su forma y naturaleza, que pueda proporcionar información; ésta puede tener forma de textos escritos, datos estadísticos, datos gráficos, registros, resultados, etc., que se encuentran depositados en diversos soportes, los cuales van desde el papel, que es el más tradicional, hasta los discos, las películas, las bandas magnéticas, las tarjetas perforadas, memorias electrónicas u otros.[5]

De acuerdo con lo anterior, es importante aclarar que actualmente la documentación que sirve como fuente a la lexicografía son palabras que se obtienen de un conjunto de textos escritos en lenguaje natural llamado *corpus*. La palabra es la unidad mínima aprovechable en la elaboración de diccionarios.

La palabra como información que se debe recuperar

La información o conocimiento utilizado básicamente en los diccionarios es la parte del conocimiento humano que llamamos "palabras", las cuales fueron expresadas por hablantes a través de la lengua primeramente hablada y posteriormente escrita, por lo que se encuentra depositada principalmente en textos. Para la recuperación de las palabras es necesario entender algunas generalidades del lenguaje natural, en el cual se encuentran ubicadas.

Respecto a su conceptualización, es difícil aclararla, ya que se pueden encontrar varias defin nes de lo que es *lenguaje*.

Para Martinet "hace medio siglo que las definiciones de lenguaje vienen presentando una cierta coherencia: parten todas del concepto de lengua como un sistema de signos utilizados para establecer comunicación".[6]

Además, establece Martinet una oposición entre lengua y lenguaje: "una lengua es definida tanto por su carácter vocal, como por su linealidad y doble articulación. El lenguaje corresponde a un concepto más amplio y responde a una definición del tipo: sistema de signos utilizados para la comunicación entre dos seres vivos".[7]

Saussure también distingue la lengua del lenguaje. Para él, la lengua y la palabra forman parte de un todo que es el lenguaje: "el lenguaje tiene un aspecto individual y un aspecto social, lo que hace imposible concebir el uno sin el otro."[8]

Al examinar la literatura existente sobre este tema, se encuentran una serie de constantes que se pueden enumerar así:

1. El lenguaje es un *sistema*, lo que significa que es un conjunto de unidades organizadas y relacionadas entre sí.
2. Al lenguaje es posible analizarlo en sus componentes.
3. El lenguaje es *arbitrario*, ya que no existe una necesidad intrínseca para que cada palabra signifique lo que significa, o para que cada lengua tenga la estructura que tiene.
4. El lenguaje es *vocal*, al estar formado por sonidos producidos por los órganos del habla de los seres humanos.
5. El lenguaje es *simbólico*, ya que por medio de él se tiene la capacidad de representar los conceptos, primeramente originados en la mente.
6. El lenguaje es un *vehículo del pensamiento*, al considerarse el pensamiento como algo distinto al lenguaje usado para expresarlo.
7. El lenguaje es *un producto de estímulos nerviosos del cerebro*, esto considerado por lingüistas y psicólogos que argumentan que lenguaje y pensamiento son lo mismo.

La similitud en estas constantes se encuentra principalmente al ser considerado el lenguaje como una capacidad humana de crear mensajes para comunicarse entre sí, y a la lengua como un sistema de signos vocales utilizados por las comunidades lingüísticas.

El lenguaje natural y su relación con el lenguaje documental

Existen muchas semejanzas y divergencias entre el lenguaje natural utilizado por los lexicógrafos, y el lenguaje documental con el que trabajan cotidianamente los bibliotecólogos, pero se podría decir que hay dos tipos de semejanzas principales: la *semejanza necesaria*, que se debe a los instrumentos utilizados por el lenguaje documental como resultado del lenguaje natural, y las *semejanzas contingentes*, que varían de acuerdo al tipo de lenguaje documental utilizado.

Los aspectos del lenguaje natural que se deben considerar en la recuperación de información respecto al lenguaje documental son:[9]

1. La parte oral es importante, ya que las formas escritas se derivan de él, a diferencia del lenguaje documental, que es fundamentalmente escrito.
2. Es de tipo general, es decir, no es especializado.
3. Es establecido y adaptado a través de largos periodos de tiempo y por consenso de la sociedad, es decir no lo hacen los especialistas ni los encargados de recuperar la información.
4. La sinonimia y la polisemia son factores naturales de este lenguaje, cosa contraria a la que sucede con los lenguajes documentales que buscan que un término tenga un solo significado.
5. Este lenguaje es aceptado y adquirido de manera natural por el usuario o hablante.
6. No es artificial, pues se produce de manera espontánea.
7. Tiene su propia estructura, de la cual se desprenden otras, como la estructura del lenguaje documental.
8. Es un instrumento de comunicación.
9. Tiene creatividad.
10. Es suceptible a cambios culturales.
11. Se caracteriza por tener una doble articulación.

12. Tiene su propia teoría, la misma de la lingüística.
13. No tiene funciones específicas.
14. Necesita respetar una secuencia de trazos, por ser inevitablemente lineal.
15. Las palabras gramaticales (artículos, conjunciones, preposiciones, etc.) son las palabras más utilizadas. A diferencia de lo que sucede en los lenguajes documentales, donde las más frecuentes son las palabras sustantivadas y los sustantivos.

Teniendo en cuenta lo anterior, es el lenguaje natural del cual se pretende recuperar la información o conocimiento (palabras) almacenado en archivos por los lexicógrafos, con el fin de elaborar diccionarios; también resulta comprensible que se utilice la indización del lenguaje natural como la técnica de análisis para los documentos, ya que esto permite al lexicógrafo recuperar y tener acceso, por medio de datos físicos y de contenido, a la información bibliográfica pertinente a sus necesidades.

La indización: una técnica de análisis documental aplicada en la elaboración de diccionarios

En los estudios interdisciplinarios entre la bibliotecología y la lexicografía, se encuentra el renovado interés en el procesamiento del lenguaje natural, a causa del incremento de la capacidad de memoria en las computadoras y al empleo de programas más sofisticados para recuperar la información.

La dificultad que significa el unir armónicamente las técnicas lingüísticas con los objetivos de la recuperación de la información fue explicada por la Federación Internacional de Documentación (FID)[10] en los años sesenta; sin embargo, la producción de estudios sobre este tema ha sido continua, aunque éstos sean de más interés para la bibliotecología, la documentación y la ciencia de la información, que para la misma lingüística. Sin embargo, a pesar de este he-

cho, en la lingüística en general y en la lexicografía en particular, se están desarrollando áreas del conocimiento con posibilidades de establecer "intersecciones"[11] con las disciplinas ya mencionadas. A las áreas de intersección tradicionalmente conocidas —la morfología, la sintaxis y la semántica— se suman otras, tales como la lingüística computacional, la terminología teórica y aplicada, etc. Y, tomando como base el acercamiento entre la indización y la recuperación de la información, se han llegado a establecer objetivos comunes entre la lingüística y las disciplinas que estudian la información, tales como: el control terminológico, la compatibilidad entre los lenguajes de indización, y la ponderación de términos en cuanto a la relevancia de la información.

La indización automatizada es el área más viable para realizar estudios en lo que respecta a este tipo de investigaciones interdisciplinarias, y es en países como Estados Unidos, Francia, Gran Bretaña, Alemania, Rusia y Japón, donde se está a la vanguardia en este tipo de investigaciones y proyectos. Sin embargo, en Hispanoamérica también se está trabajando al respecto de manera sobresaliente: en México, en el Diccionario del Español de México y el Centro de Estudios Lingüísticos y Literarios de El Colegio de México, así como en el Consejo Nacional de Ciencia y Tecnología y el Centro de Investigaciones Científicas y Humanísticas de la Universidad Nacional Autónoma de México; en Cuba, en el Instituto de Lingüística y el IDICT; en Puerto Rico, en la Universidad de Puerto Rico; en Venezuela, en La Universidad Simón Bolivar y la Universidad de Caracas; en Brasil, en el IBICT, el Lexicón y la Universidad de Sao Paulo; en Chile, en el PUC de Chile y el CLADES.

En cuanto a la indización misma, desde 1978, fecha en que Spark Jones[12] expuso la importancia de los sistemas en línea, la indización automatizada mantiene relaciones con la lingüística con miras primordialmente a la resolución de problemas para una mejor y efectiva recuperación de la información.

La indización es reconocida generalmente como una técnica de

análisis de contenido, que condensa la información significativa de un documento por medio de la asignación de términos, creando así un lenguaje intermedio entre los usuarios y el documento. Es uno de los procesos básicos de la recuperación de la información y puede ser realizada por el hombre (indización manual), o por programas de computadora (indización automatizada).

El Sistema Mundial de Información Científica (UNISIST),[13] en uno de sus grupos de estudio, elaboró un documento con los principios de indización, en el que se decía que la indización es la operación que describe e identifica el contenido de los documentos a través de términos, es decir, que los documentos pueden ser representados por términos seleccionados del lenguaje natural o del documental.

La indización, al estar directamente relacionada con la descripción física de los documentos, constituye un registro bibliográfico, que proporciona al usuario información física y de contenido de los documentos. Estos datos son organizados así en una forma más accesible para la recuperación de la información.

Las políticas de indización y de recuperación varían, respectivamente, de acuerdo con la exhaustividad y la precisión aplicada en el análisis de contenido que se practique. Estas políticas dependen de las necesidades que tengan los diversos usuarios a los que atiende un sistema de recuperación de información.

La indización puede ser realizada en un documento, ya sea en todo o en sus partes, y es una estrategia de búsqueda para la recuperación dentro de un sistema de información.

En lo que respecta a la indización automatizada, ésta comenzó al final de la década de los 50, cuando Luhn,[14] desarrolló la idea de que el vocabulario que existe en un documento debería constituir la base para el análisis de su contenido, y que ésta es la mejor manera para recuperarlo.

La primera aplicación realizada por Luhn fue la producción del sistema *key word in context* (kwic). Estos índices están elaborados a

partir de la rotación de palabras significativas de los títulos. De acuerdo con el mismo autor, este proceso puede identificar ... nos, pares de términos y hasta frases significativas que expresen el contenido de los documentos. Respecto a esto mismo, Steinacker[15] en 1974 propone un algoritmo para identificar frases o grupos de palabras significativas. El algoritmo produce cortes en el texto y los localiza; posteriormente los ordena alfabéticamente y construye un índice rotativo o giratorio de las diferentes combinaciones de las palabras de un mismo corte. Entre las aplicaciones de esa técnica pueden ser citadas: la creación de diccionarios; la elaboración de *thesaurus;* desarrollo, control y mantenimiento de enciclopedias; y la determinación de categorías gramaticales. El índice utilizado en la documentación de diccionarios es muy parecido al (kwic) y consiste en un conjunto de "concordancias". (Véase *Representación gráfica de la información*, en el final de esta parte).

La indización automatizada es, en general, un proceso u operación que identifica las palabras o expresiones significativas de los documentos, para describir en forma condensada el contenido de éstos, y utiliza para ello diferentes métodos desarrollados para programas de computadora. Esa operación, según Robredo,[16] es objetiva, pues se realizan siempre los mismos programas de extracción de términos significativos de los documentos, con lo cua se elimina la inconsistencia de la subjetividad que aparece en la indización manual, y es posible una mejor recuperación. El proceso de indización automatizada de la estrategia de búsqueda es realizado por los mismos programas, asegurando así la compatibilidad entre el lenguaje utilizado en la indización y el utilizado en la realización de la pregunta.

Vickery[17] menciona las siguientes funciones de los lenguajes de indización:

1. Recuperar los documentos con contenidos semejantes.
2. Recuperar los documentos relevantes sobre un asunto específico.

3. Recuperar documentos por grandes áreas temáticas.
4. Hacer la conversión de términos de indización entre diferentes lenguajes.
5. Servir de auxiliar en la selección del término adecuado para la estrategia de búsqueda.

Según Robredo[18] la indización puede ser realizada básicamente en tres niveles, partiendo del más general hasta llegar al más específico:

a) De *categorización*: representa el asunto que predomina.
b) *Superficial*: representa los conceptos principales de manera general.
c) *Profunda*: representa todos los conceptos fundamentales.

Los términos de la indización pueden ser expresados a través de distintos tipos de lenguaje:

Natural del libro o texto, utilizando los mismos términos del autor.

Controlado, que adopta términos valorados y definidos previamente.

Codificado, utilizando códigos anticipadamente establecidos para expresar los términos significativos.

Coordinado, que contiene implícitas las relaciones lógicas entre los términos, cuando éstas existen, estableciéndolas por medio de: equivalencia de la sinonimia entre los términos; subordinación de la jerarquía, partiendo de lo general a lo específico y viceversa; y coordinación o asociatividad, en la que los conceptos están relacionados a la idea de otro concepto.

En su reseña de 1982, Travis y Fidel[19] distinguen tres tipos de sistemas en la producción automática de índices y resúmenes:

1) *Sistemas de indización basados en diccionarios.* En cuanto a este sistema de indización automatizada, es, de hecho, el resultado de un análisis lingüístico sobre textos y se utiliza para atribuirle categorías gramaticales a las palabras que aparecen

en los textos seleccionados pertenecientes a un grupo de textos (*corpus*).

2) *Sistemas de indización basados en recursos estadísticos.* Destaca en este tipo de sistemas el *Método de frecuencia*, que tuvo su origen en 1957 y 1958 al ser propuesto por Luhn,[20] en donde enunciaba que la frecuencia de una palabra en los documentos está directamente relacionada con la capacidad de esa palabra para representar el contenido de un documento bajo un nivel de indización y de recuperación de información. El método de frecuencia trata del conteo automático de la aparición de las palabras, que se encuentran localizadas en textos y esto se realiza a través de la ocurrencia y de la co-ocurrencia de las palabras.

En este método la ocurrencia puede ser establecida, según Spark Jones,[21] al determinar: a) la frecuencia total de las palabras en el documento, donde la palabra es contada todas las veces que aparece. Posteriormente se hace la suma de las veces en que co-ocurre; b) la ocurrencia única de la palabra en el documento. Se cuenta una vez la palabra, independientemente del número de veces que ésta aparezca; c) la ocurrencia de las palabras en la colección. El conteo se realiza sumándose la aparición de cada palabra en la colección o archivo del sistema.

Hay que aclarar que a partir de los datos estadísticos se puede efectuar lo que Soergel[22] denomina el "conteo de conceptos", en donde se suman las frecuencias de todas las palabras o términos que determinan un concepto.

Existe otro método de indización basado en recursos estadísticos, llamado *Método de atribución de peso*, ha sido utilizado principalmente en cuanto a criterios de selección de palabras y es además en el que, en la cuarta parte de estudio, se finca la investigación empírica realizada.

Según Salton,[23] este método es una forma de atribuirle valores

semánticos a las palabras para hacerlas más precisas, sin que disminuya su capacidad de recuperación. Para esto se basa en la frecuencia de cada palabra. Luhn[24] fue nuevamente el precursor de este método, al proponer un modelo relacionado directamente con la frecuencia de una palabra completa, de una palabra truncada o una raíz de palabra y el valor de esa palabra para expresar el contenido de los documentos, o sea que en cuanto mayor sea la frecuencia de una palabra, mayor peso recibirá.

El peso puede ser atribuido de acuerdo con varios aspectos:

La frecuencia total o frecuencia única, donde la palabra recibe el mismo valor de su número de su frecuencia.

La fuente. Si una palabra se encuentra en un documento reconocido como relevante, recibirá un peso mayor que el que pueda recibir una existente en un documento menos relevante.

Por *la fuente y el usuario*. El usuario es quien juzgará si un documento es relevante o no.

Respecto a lo anterior, existe el antecedente de que Salton y Yang[25] analizaron la teoría de la relevancia para el usuario como método de atribución de peso, al utilizar la frecuencia de ocurrencias de palabras de un documento en la colección. Requiere una retroalimentación constante, pues utiliza el juicio de relevancia de los usuarios para la atribución de pesos.

Por último, respecto a este método también se puede añadir que Robertson y Sparck Jones[26] propusieron fórmulas matemáticas para la atribución de peso basadas en la teoría de la relevancia.

3) *Sistemas de indización basados en resúmenes.* Un resumen es un discurso sobre un discurso, en el que se convierte un documento primario en uno secundario, así se presenta de manera condensada el contenido de un documento. Por medio de la indización se traduce el lenguaje del resumen que se analiza a las expresiones del lenguaje natural o documental para que posteriormente sean utilizadas en la recuperación del documento original. En los resúmenes se puede indizar de manera automatizada por medio del lenguaje natural,

el cual es entresacado del mismo texto indizado, o, se puede indizar con lenguaje documental que supone la previa existencia de un vocabulario estereotipado elejido antes de comenzar la indización.

La indización bibliográfica y los sistemas de recuperación

La investigación documental y la recuperación de la información bibliográfica y de documentación por métodos electrónicos tienen una creciente importancia para manejar los datos que se precisan en todas las áreas del conocimiento humano, como también ocurre en la elaboración de diccionarios.

En general, un sistema de recuperación de información de tipo bibliográfico[27] presenta las siguientes características:

1. Almacena información bibliográfica.
2. Usa descriptores o palabras clave para la descripción y el acceso a los contenidos.
3. Debe minimizar el tiempo de búsqueda entre los enormes volúmenes de información disponible.
4. Debe dar el máximo de información con el mínimo de esfuerzo y duplicación.
5. Produce índices bibliográficos como método de recuperación.
6. Permite la recuperación específica de documentos que satisfagan los requerimientos elaborados por búsquedas lógicas de palabras clave.
7. Debe tener en cuenta las zonas de contacto existentes entre las distintas especialidades.
8. Debe tratar de ser, en lo posible, compatible con otros sistemas de recuperación para facilitar el intercambio de datos.
9. Debido a la gran cantidad de datos que deben manejar los sistemas de recuperación bibliográfica y la exigencia de rapidez de dicha recuperación, los sistemas deben estar orientados hacia el adecuado proceso mecánico o electrónico que pueda

favorecer el uso de diversos programas de procesamiento de datos, actualización de archivos, mantenimiento, etc., en fin, todo lo necesario para poder tener operando el sistema sin contratiempos.

Los tipos de respuesta en un sistema de recuperación de información[28] pueden tener bien definidas sus características y como un sistema de recuperación de información es cualquiera de los servicios que tienen como principal ocupación el producir una adecuada respuesta a los requerimientos que se le hagan para proporcionar información, se pueden distinguir cuatro tipos generales:

Recuperación por referencias, que es la recuperación de cada uno de los documentos sustitutos; por ejemplo, un resumen, o una referencia bibliográfica de un documento o conjunto de documentos, los cuales pueden contener información relevante para la solicitud realizada. En el caso de diccionarios, esta recuperación se realiza por medio de listas bibliográficas de un conjunto cerrado de textos (*corpus*), y por fichas y tarjetas lexicográficas con referencias bibliográficas obtenidas por medio de la indización de documentos fuente secundarios.

Recuperación real, que es propiamente la recuperación de un ítem (artículo, noticia, párrafo, definición, etc.), localizado en un texto. Para el caso de los diccionarios, esta recuperación se realiza por medio de: concordancias, que son muy parecidas a las palabras en su contexto (kwic); hojas de diccionarios testigo, donde se documenta la información contenida en diccionarios distintos al que se está elaborando; fichas lexicológicas y lexicográficas; información proporcionada por consultas hechas a especialistas de diferentes áreas temáticas.

Pregunta-respuesta, en la que, con base en una respuesta entregada a la solicitud original de información, se presenta una nueva pregunta inferida del material documental presentado como respuesta. En el sistema de recuperación de información del DEM, este servicio no se practica, por lo menos no con regularidad.

Recuperación de datos, que es la recuperación de ítems en tablas que contienen datos, detalles, etc. En el trabajo de los diccionarios se recuperan datos estadísticos obtenidos del *corpus* por medio de la indización de palabras, que se presentan en forma de índices automatizados o impresos.

El Diccionario del Español de México (DEM)

Antecedentes

El Sr. Antonio Carrillo Flores, director en 1972 del Fondo de Cultura Económica, realizó una invitación a El Colegio de México para que hiciera un diccionario mexicano del español, con el objeto de que no se perdieran o ignoraran tantas palabras necesarias y queridas de México. El Sr. Víctor Urquidi transmitió esta inquietud al Centro de Estudios Lingüísticos y Literarios (CELL) por conducto del Sr. Antonio Alatorre y éste a su vez al Dr. Luis Fernando Lara,[29] actual coordinador del proyecto especial denominado Diccionario del Español de México (DEM). Meses más tarde apareció publicado en *La gaceta* del Fondo de Cultura Económica, el dictamen elaborado por el Dr. Lara, que sirvió como base para que la junta de gobierno del mismo Fondo aprobara el inicio de las actividades del DEM, como un diccionario integral de la lengua española elaborado fuera de las fronteras de España y basado en los usos de México, una de las regiones de la lengua española que tiene más hablantes del mundo.

Fue durante el mandato presidencial de Luis Echeverría Álvarez cuando, a solicitud de don Víctor Bravo Ahuja, secretario de educación pública, a fines de 1972 se estableció un fideicomiso para secundar a El Colegio de México en la realización de este diccionario mexicano.

Una vez que se constituyó el DEM, el equipo lexicográfico dirigi-

do por el Dr. Luis Fernando Lara quedó dividido en cuatro secciones, las cuales se denominaron: de redacción, de revisión y corrección, de documentación, y el equipo de investigaciones estadísticas y computacionales; cada una de éstas contribuyó a la obtención de ambiciosos objetivos del proyecto. Actualmente, el DEM es, por sus importantes logros, por sus recursos instalados y por sus proyectos, un verdadero tesoro cultural del pueblo de México.

Precisamente la documentación, uno de los aspectos de esa riqueza con que cuenta este diccionario, es la que hizo posible realizar este estudio de interés para la bibliotecología. El aspecto más destacado se circunscribirá a las actividades que se realizan para documentar o testificar la información utilizada en un diccionario. Sin embargo, para establecer un contexto adecuado que permita valorar la importancia de estas actividades, es necesario explicar ampliamente qué es el Diccionario del Español de México (DEM) y los aspectos que lo caracterizan.

Las pautas que delimitan la elaboración del DEM son básicamente las siguientes[30]:

1. El DEM es un diccionario de tipo sincrónico, es decir, contiene el estado de la realización lingüística de un determinado momento, por ello el material documental utilizado se compone de textos (hablados o escritos en México), que se hayan producido entre 1970 y 1973. En el caso de las obras literarias o científicas que formaron parte de la documentación utilizada, se aceptaron textos publicados después de 1921.
2. El DEM es descriptivo, o sea que ordena e interpreta los datos originales, perceptibles y medibles (documentación), de tal manera que no fuerza el significado de los mismos.
3. El DEM es selectivo por aspectos de limitaciones en tiempo y en dinero, lo que obligó a una selección estricta de los vocablos que habría que incluir, por lo que se basa en criterios predeterminados, pero ajustables a las necesidades que se vayan presentando en la labor lexicográfica.

4. El DEM es una obra lexicográfica cuyo objetivo fundamental es ser un diccionario regional de la lengua común española hablada en México, y así refleja el léxico del español utilizado actualmente en el país, en cuanto a sus modalidades escritas y orales, cultas y coloquiales, urbanas y rurales.
5. El interés del DEM es mostrar el léxico (conjunto de palabras) o subsistema lingüístico del español que se utiliza entre las fronteras políticas de México.
6. El trabajo efectuado busca tener una finalidad práctica, que es la de servir a un hablante mexicano medio, como una obra de consulta y como punto de referencia en su apreciación del idioma.

De todo este cúmulo de pautas y objetivos, surge una gran cantidad de necesidades y de problemas relativos al mismo quehacer lexicográfico, tales como: la necesaria objetividad de la descripción lingüística; la adecuación entre los métodos utilizados y la realidad de los fenómenos léxicos; el volumen de datos que requiere una definición lexicográfica completa y los resultados que ha alcanzado hasta hoy la lexicografía española, en particular la hispanoamericana, de tal manera que pudieran influir en las actividades que realiza el DEM.

Las tareas lexicográficas del DEM

Para que un diccionario pueda llegar a manos de sus usuarios, es necesario efectuar varias tareas fundamentales, éstas son:

a) El registro o atestación, que comprende la organización del trabajo para compilar y registrar los materiales lexicográficos respecto a la información de las fuentes, como son: el grupo de textos (*corpus*) de donde se obtendrán las palabras, apuntes hechos sobre las palabras, aportaciones del público (usuarios), aportaciones de otros lexicógrafos, etc.; es decir, es el registro de la información obtenida en fuentes primarias.

b) *La documentación*, que es donde se reciben los materiales provenientes del registro y se les da el formato y orden necesarios. Como primera tarea de la documentación, se presenta la *normalización*, aquí se toman decisiones acerca de conservar o no los distintos registros ortográficos de una misma palabra, de si una palabra se refiere a cosas distintas o a lo mismo, y si una palabra se toma como lema o entrada de diccionario o si sólo se considera como una variante. Para efectuar lo anterior, se consideran las variantes ortográficas, las citas, las fuentes, los informantes, etc., razón por la cual se crean, para facilitar este proceso, los formatos de documentación que lleguen a ser necesarios para cada tipo de información, aunque generalmente en éstos se anota de dónde proviene la palabra, de quién y de dónde se tomó la cita, etc. Como segunda tarea de la documentación se considera la *exploración* de fuentes secundarias, como las revistas filológicas y de tipo lingüístico, de las que por medio de la indización, la transcripción y el registro bibliográfico se obtiene la documentación que se espera mantenga al día la información sobre los vocablos. La tercera tarea de la documentación, es la preparación de *monografías léxicas* para cada palabra o vocablo; para ello se reúne toda la información a propósito del vocablo que se va a redactar.

c) *La redacción* es el núcleo del diccionario. Es aquí donde se realiza el análisis semántico de los datos que se encuentran en la monografía léxica, y se interpretan todos los registros de fuentes primarias y fuentes secundarias con el objeto de redactar el artículo del diccionario, esto teniendo en cuenta siempre las características que se hayan previsto que contengan, tales como: marcas gramaticales, niveles de lenguaje, ejemplos, etcétera. Por experiencias se sabe que es mejor que un pequeño grupo de redactores participe en esta actividad para uniformar estilos y hacer correcciones lo menos posible.

d) *La revisión*. Esta tarea debe ser de dos tipos: *la interna*, hecha por una sola persona del equipo lexicográfico, que a su vez no haya

hecho la redacción del vocablo y que tenga facilidad para esta actividad. Debe revisar la redacción y además verificarla; de hecho, aquí se hace el análisis íntegro de nuevo, para verificar la redacción y señalar o, en su caso, resolver los problemas; es un proceso tan largo como la redacción. La *revisión externa*, que se realiza en una gran cantidad de vocablos en los cuales no basta una revisión interna y se tiene que pasar a un arbitraje para tomar decisiones sobre las palabras. Para esta revisión conviene contar con el auxilio de personas ajenas a la lexicografía, pero cuyos conocimientos especializados ayuden a tomar una decisión.

e) La edición. Por lo general esta etapa es un ciclo que comprende la transcripción tipográfica del diccionario y su revisión, hasta lograr su impresión.

La documentación en el DEM

Por lo antes expuesto, se puede comprender que el trabajo de documentación es muy importante para el DEM, ya que de él depende la calidad que pueda adquirir la obra en cuanto al [illegible] y la utilización de fuentes informativas. Se requiere un cuidado extremo en el asentamiento de referencias y de datos, un rigor muy elevado para dar solidez a los documentos, y un acervo amplio de conocimientos para indagar referencias en áreas que a primera vista pueden no tener relación con el documento en cuestión, por lo que las personas de esta sección trabajan exclusivamente en la documentación y no se les exige redactar.

Archivos documentales

Son los archivos de donde se recupera la información lexicográfica. El DEM tiene distintos tipos de archivos y por esto mismo, utiliza va-

rios tipos de recuperación: por referencia, real y de datos, para proporcionar una adecuada respuesta a los requerimientos de información sobre las palabras usadas en el español de México, por lo que resulta oportuno explicar cuáles son los archivos de donde se obtiene esta información y cómo se recupera.

Archivo de palabras. Son aproximadamente dos millones de palabras que se encuentran almacenadas en la memoria de una computadora y son el resultado de la indización automatizada en lenguaje natural practicada sobre textos, además del tratamiento por medio del análisis lingüístico y del análisis estadístico aplicado a dichos textos. El nombre que se le dio a este archivo es el de *Corpus del español mexicano contemporáneo* (Cemc). La computadora hace posible recuperar la información a través de palabras completas, truncadas o raíces de palabras en su contexto, es decir por medio de concordancias, por lo que es una recuperación de tipo real.

Archivo bibliográfico. Está compuesto por 966 fichas bibliográficas, cuyos códigos representan igual cantidad de textos que componen el *corpus.* Esta información indica además la clasificación del nivel de lengua a la que pertenece una palabra con base en su origen dentro de la literatura temática que compone el *corpus* explotado por los lexicógrafos, y se puede utilizar comparando el código numérico que conlleva cada concordancia con la lista bibliográfica de las obras que componen el *corpus*; es decir, es una recuperación por referencia.

Archivo de hojas de diccionarios testigo. Es la información obtenida y registrada gracias a la documentación que realizan los lexicógrafos sobre las definiciones contenidas en diccionarios considerados documentos secundarios, a los cuales se les estima como indicadores del uso de las palabras en el español. Esta documentación se realiza en cada palabra requerida para consulta, o previendo su próxima utilización. El tipo de recuperación aquí es manual y del tipo real.

Archivo de fichas lexicográficas. Está formado por fichas pequeñas que contienen el registro de las palabras y que se ordenan alfabéticamente en ficheros dispuestos específicamente para este tipo

de información. En el registro de cada palabra también se incluye la definición, si la tiene, y la fuente bibliográfica donde se localizó. Su elaboración es manual y la recuperación es manual y del tipo real.

Archivo de tarjetas lexicográficas. Compuesto propiamente por tarjetas media carta que contienen: el registro de las palabras en orden alfabético, el significado que les dan los hablantes, el nombre del informante y su localización geográfica. La recuperación en este caso es manual y del tipo real.

Archivo de datos estadísticos. Son los datos principalmente de frecuencias y de relaciones porcentuales estadísticas respecto a las palabras que se encuentran en el *corpus*. Al instrumento por el que se consiguen dichos datos se le conoce como *Diccionario estadístico del español mexicano*. La recuperación en este caso es automatizada, aunque también hay índices ya impresos de estos mismos datos.

Existe otro tipo de información que acompaña a algunos vocablos, y es la que proporcionan los especialistas de diferentes áreas del conocimiento, previa solicitud de algún redactor del DEM. Esta información, aunque de suma importancia, no constituye un archivo formal.

Los archivos descritos pueden encontrarse físicamente independientes como se han descrito o como integrantes de la *monografía léxica*, la cual contiene todos los datos que habrá de interpretar el redactor para elaborar los artículos que han de aparecer en el diccionario.

El sistema de recuperación automatizado se encuentra en el *archivo de palabras* o *Corpus del Español de México* y con base en la información que es recuperada por éste, es que en la cuarta parte de este trabajo se pondera la información considerada relevante.

Sistema de recuperación de información

Durante los primeros cuatro años de existencia del DEM, los lexicógrafos se dedicaron a la investigación, pues se requería reunir una

gran cantidad de datos que presentaran todas las maneras de hablar en el país, con el objeto de discernir a partir de ellas la realidad de nuestro vocabulario, y para ello resultó necesario construir el sistema de recuperación de información léxica llamado *Corpus del español mexicano contemporáneo (1921-1974)*,[31] el cual es llamado también Cemc, que contiene todos los textos escritos y hablados de géneros de la lengua, de distintos niveles sociales y de todas las regiones del país. Construir el Cemc implicó seleccionar materiales representativos en bibliotecas y fonotecas, transcribirlos en tarjetas perforadas, ordenarlos con la computadora electrónica y analizarlos para obtener de ellos los resultados documentales esperados. Aquí se reunieron cerca de dos millones de palabras, correspondientes a 966 textos diferentes.

Distribución de textos en el Corpus *del Español Mexicano Contemporáneo (fuente de información documental)*

Clasificación	*Textos analizados*	*Índice de concordancias*	*Clave de referencia bibliográfica*
Literatura	95	190 000	000-094
Cuento y ensayos	55	110 000	095-149
Periodismo	176	352 000	150-325
Ciencias	180	360 000	326-505
Técnicas	77	154 000	506-582
Del hogar	25	50 000	583-607
Discurso político	18	36 000	608-625
Religión	12	24 000	626-637
Habla de la ciudad de México	30	60 000	638-667
Novela rosa	18	36 000	668-685

Telenovela	8	16 000	686-693
Fotonovela	15	30 000	694-730
Historieta	10	20 000	709-718
Novela popular	12	24 000	719-730
Habla media de la ciudad de México	30	60 000	731-760
Lírica popular	24	3 000	761-784
Textos dialectáles	130	000	785-914
Antropológicos	33		915-947
Jergas	12	24 000	948-959
Vocabulario del hampa	6	12 000	960-965
Totales	966	1 932 000*	966

* De esta cantidad de concordancias el analizador utilizado por el DEM al agruparlos sintácticamente obtuvo 64 193 tipos, de los cuales se esperaba que documentaran 30 000 palabras aproximadamente.

Índice temático y de códigos para identificar la información bibliográfica que es recuperada por medio de las concordancias

LENGUA ESTÁNDAR

LENGUA CULTA

(1) Literatura

000-094 Obras de literatura

095-149 Cuentos y ensayos aparecidos en revistas y suplementos culturales

(2) Periodismo

150-171 Reportajes de autores mexicanos
172-206 Editoriales
207-241 Reseñas politicas
242-249 Reseñas sociales
250-284 Reseñas culturales
285-309 Reseñas deportivas
310-317 Reseñas policiacas
318-325 Reseñas taurinas

(3) Ciencias

326-327 Bibliotecología
328-332 Filosofía
333-338 Historia
339-342 Culturas indígenas
343-344 Educación y pedagogía
345-350 Psicología
351-354 Antropología
355-356 Arqueología
357-362 Derecho
363-367 Economía
368-370 Geografía
371-374 Política
375-378 Sociología
379-383 Astronomía
384-388 Matemáticas
389-391 Electrónica y electricidad
392-395 Física
369-399 Geofísica
400-403 Computación
404-415 Biología
416-428 Química
429-435 Administración
436-448 Contabilidad
449-453 Comercio
454-457 Medicina veterinaria
458-478 Medicina humana

479-483 Arquitectura
484-486 Artes coreográficas
487-492 Artes plásticas
493-494 Artes gráficas
485-486 Arte dramático
497-501 Música
502-505 Cine y fotografía

(4) Técnicas

506-507 Correos y filatelia
508-509 Periodismo
510-511 Publicidad
512-513 Radio y televisión
514-515 Transporte
516-517 Mercadotecnia
518-522 Ingeniería civil
523-524 Ingeniería industrial
525-527 Ingeniería química
528-530 Ingeniería automotriz
531-532 Ingeniería aérea
533 Ingeniería de ferrocarriles
534 Ingeniería naval
535-536 Ingeniería de minas
537-540 Carpintería
541-544 Electricidad
545-547 Mecánica
548 Dibujo técnico
549-550 Enfermería
551 Corte y confección
552-553 Albañilería
554-555 Plomería
556 Herrería
557-575 Agropecuarias
576-579 Caza y pesca
580-581 Ejército
582 Charrería
583-607 Textos el hogar

(5) Discursos políticos

608-625 Discurso político

(6) Religión

626-637 Religión

(7) Habla culta

638-667 Habla de la ciudad de México

LENGUA SUBCULTA

(8) Literatura popular

668-685 Novela rosa
686-693 Telenovela
694-708 Fotonovela
709-718 Historieta
719-730 Novela popular

(9) Habla media

731-760 Habla media de la ciudad de México

(10) Lírica popular

761-774 Habla media
775-784 Habla regional

LENGUA NO ESTÁNDAR

785-914 **(11) Textos dialectales**
915-947 **(12) Documentos antropológicos**
948-959 **(13) Textos jergales**
960-965 **(14) Textos del hampa**

Junto con la estructuración del Cemc, hubo que crear un sistema de análisis computacional de la lengua española que permitiera analizar lingüística y estadísticamente los materiales, para obtener resultados cuantitativos y cualitativos útiles para la lexicografía. Este programa tomó el nombre de *analizador gramatical*[32] del DEM, el cual es un conjunto de programas de computación que leen, reconocen, cuentan y clasifican las palabras que se encuentran en el *corpus*. Las metas que perseguía este sistema son:

Producir listas de las diferentes palabras que se encuentran en el *corpus*, asociando a cada palabra su frecuencia de aparición en la muestra, y algunos valores estadísticos que sirvieran para medir la representatividad de su frecuencia.

Llevar el conteo de frecuencias en forma adecuada por medio de la computadora, lo que hizo necesario diseñar un algoritmo que permitiera resolver dos problemas básicos:

1. La diferenciación entre homógrafos, por ejemplo, "amo" del verbo amar y "amo" sustantivo, de tal manera que cada ocurrencia se contase donde correspondiere y así evitar confusiones.
2. La agrupación de todas las diferentes formas de un lexema, por ejemplo, "comimos", "comí", "comeré", son diversas formas del verbo "comer" y, por lo tanto, cuando ocurren en los textos, deben contarse como ocurrencias del verbo "comer".

Para resolver estos problemas, los lingüistas y matemáticos, con el apoyo de Roberto Ham Chande,[33] obtuvieron como resultado el planteamiento de las bases lingüísticas del sistema de análisis gramatical.

Los problemas básicos que debían ser resueltos para que en la computadora se realizara el conteo de frecuencias de los vocablos en forma adecuada, era la diferenciación entre homógrafos y la

agrupación de todas las diferentes formas de una palabra (tipos). La diferenciación entre homógrafos podía llevarse a cabo si las palabras del *corpus* tuvieran asociada su categoría gramatical. La solución planteada contemplaba como requisito indispensable la asociación automática de categoría gramatical a cada ocurrencia del (Cemc). Se decidió que algunas reglas de reconocimiento de la morfología de los vocablos, así como algunas otras basadas en las relaciones de procedencia entre vocablos de un sintagma y las reglas de concordancia que permitían manejar las mofológicas y las de precedencia de manera simultánea, resultaban ser lo suficientemente consistentes como para pensar en diseñar algoritmos computacionales que, haciendo uso de tales reglas, produjeran el análisis sistemático del *corpus*, es decir, el etiquetamiento gramatical de todos los vocablos contenidos en él.

Estas categorías asignadas son convencionales y fueron preestablecidas en los programas de recuperación por los lexicógrafos del DEM y por Isabel García Hidalgo[34] de la Unidad de Cómputo de El Colegio de México. Las categorías gramaticales asignadas fueron:

(0) = ambigua
(1) = adverbio
(2) = adjetivo
(3) = conjunción
(4) = preposición
(5) = pronombre
(6) = artículo
(7) = contracción
(8) = nominal
(9) = verbo

Para el procesamiento de datos se contó con el apoyo del Centro de Procesamiento y Evaluación Dr. Arturo Rosenblueth de la Secretaría de Educación Pública, donde se pudo hacer uso de sus computadoras, particularmente de la *Univac 1106*, que tenía una gran capacidad para manejar datos y analizarlos.

Una vez que fue estructurado el Cemc y capturados los materiales en la computadora durante más de un año, al cabo de los primeros cinco años del proyecto ya se pudo contar con una base de datos, de 24 *megabytes* de dimensión, que presenta un estado del español de México entre 1921 y 1974; un "diccionario estadístico del español de México", de 25 *megabytes* de dimensión, en el que se ha cuantificado cada una de las palabras encontradas en el Cemc, bajo los términos de frecuencia y distribución de su uso en México, y cientos de miles de contextos de uso de las palabras, las llamadas concordancias, que forman la materia prima del trabajo del DEM y de muchas investigaciones más.

Representación gráfica de la información

El deseo original del equipo de lingüistas del DEM era producir, con la ayuda de la computadora, las listas de las diferentes palabras que se encontraran en el Cemc, y se asociara a cada palabra su frecuencia de aparición y algunos valores estadísticos que sirvieran para medir su representatividad. Además de esto, se deseaba que para cada una de las palabras encontradas en el *corpus* se pudiera recuperar un número razonable de contextos, llamados concordancias que son muy parecidas al kwic, acompañadas con los diferentes usos del vocablo en cuestión; esto último se considera documento imprescindible para redactar los artículos que expliquen el significado de cada una de las entradas que tienen los diccionarios.

La forma o representación gráfica en que se recupera la información en la lexicografía depende en gran parte del tipo de la indización automatizada que se efecúe sobre los textos y su grado de profundidad. Aunque las palabras aprovechadas exitosamente con el uso de la computadora generalmente contienen estas propiedades:

1. Tipográficamente está separada cada palabra de otras por

blancos o signos de puntuación, así se le conoce como palabra textual.
2. Los nombres propios no son objeto del estudio lexicográfico, por lo que no se toman en cuenta.
3. Las palabras son las unidades mínimas en la documentación con significado para el lexicógrafo.
4. Se reconocen como unidades léxicas.
5. Al ser recuperadas por la computadora tienen forma gráfica de palabras completas, palabras truncadas y raíces de palabra, también conocidas como "tipos". Las palabras truncadas y raíces de palabra se utilizan principalmente para disminuir el ruido, así se evita la aparición de las mismas palabras con diferentes desinencias (terminaciones) de tipo gramatical.
6. Las palabras gráficas iguales que tienen distinto significado (homógrafos) se distinguen y son tratadas de manera independiente.
7. Como listas de "concordancias", junto a cada unidad léxica se recuperan todos los contextos donde aparece la unidad léxica, junto a sus datos bibliográficos o sus códigos, que son necesarios para que tengan el valor de documento o testimonio, y que las localizan en el tema de la literatura correspondiente a un determinado nivel de lengua.
8. Al recuperarse la palabra con el texto original, o la línea anterior y la posterior de donde aparece, se convierte en una palabra en su contexto, lo que permite tratarla como un sintagma, al que por su ubicación dentro de una frase o de una oración se le puede asignar una etiqueta de su categoría gramatical.
9. Las palabras donde se agrupan manualmente las diferentes variantes gráficas de un vocablo y que cumplen funciones sintácticas y gramaticales idénticas se llama lema.
10. Los "tipos" pueden ser variantes gráficas de otra palabra re-

conocida como la unidad canónica del lenguaje, al que se le llama vocablo, en cuyo caso se agrupan en una carpeta monográfica léxica bajo la representación gráfica del vocablo, que es una palabra indizada manualmente por el lexicógrafo. A esta acción los lexicógrafos la han llamado lematización.

El tratamiento o análisis que se efectúe en las palabras también puede dar como resultado, si así se ha previsto, información de tipo estadístico, al que se puede tener acceso directamente por la computadora o a través de índices o listados estadísticos. De donde se obtienen datos de concordancias o de frecuencias sobre las palabras.

El Cemc cuenta con una colección representativa de datos ortográficos, morfológicos, sintácticos, léxicos y semánticos del español de México y de él se nutre el juicio que se realiza para incluir los vocablos que aparecen en el DEM, también a partir del Cemc se efectúa el reconocimiento de niveles y regiones de uso; y las constantes decisiones sobre los hechos ortográficos, morfológicos y sintácticos ocurridos en el español mexicano.

Terminada la obtención de resultados globales, los lexicógrafos iniciaron el análisis y redacción de cada una de las palabras, esto, a partir de prioridades, criterios y delimitaciones, y para hacer el análisis semántico previo a la redacción de artículos, se utilizaron como base las concordancias de vocablos.

El proyecto del DEM; en cuanto a su forma de investigación y a la promoción de la lengua misma, se ha convertido en uno de los más importantes centros de documentación y estudio de la lengua española en el mundo, es un importante colaborador con diferentes proyectos afines que se realizan en Hispanoamérica, el sur de Estados Unidos y en Europa, además apoya el conocimiento de la terminología hispánica sobre ciencia y técnica.

- Información solicitada: la palabra o vocablo AYATE.
- Lugar: archivos de palabra.
- Soporte documental: El folder que corresponde a su *monografía léxica*.
- Información recuperada: un tipo con dos concordancias y otro con una concordancia, lo que significa que la palabra AYATE tiene tres de frecuencia absoluta.
- Pertenencia: léxico común.

ANÁLISIS DOCUMENTAL DE LA INFORMACIÓN

Concordancias

AYATE

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.

761001314 AH (SIC), = AHI) CONOCI AL CAIMAIN CON UN PESCADO EN L'**AYATE** AYER QUE ME FUI A BA÷AR

AYATES

576026022 QUE VA COSIDA UNA BOLSA DE FORMA PIRAMIDAL DE MALLA MUY CERRADA (GENERALMENTE EMPLEAN **AYATES**), POR ESTA RAZION DEL DIAMETRO DE LA MALLA NO PASA DEL CENTIMETRO. SU MANIOBRA

Referencias bibliográficas de los textos a los que se refieren las concordancias:

834 Parsons, Elsie Clews. "Folklore from Santa Ana Xalmimilulco, Puebla, México". *The Journal of Amencan Folklore*, 45 (1932), 318-362. [Lengua hablada. México, D.F., 1929. Un informante, 105:22 a., M., medio.]

761 *La Azucena* (versión 27) Cintas Colegio, Huejutla, Hgo., 1967.

576 Mercado Sánchez, Pedro. *Breve reseña sobre las principales artes de pesca usadas en México*. Secretaría de Industria y Comercio, México, 1959. 79 pp.

Interpretación documental

Concordancia 1: palabra clave o "Tipo" AYATE. Referencia bibliográfica núm. 834. Clasificación temática: (111), textos dialectales. Pertenece a la lengua no estándar.

Concordancia 2: palabra clave o "Tipo" AYATE; Referencia bibliográfica núm. 761. Clasificación temática: (110), línea popular (habla media). Pertenece a la lengua estándar (lengua subculta).

Concordancia 3: palabra clave o "Tipo" AYATES; Referencia bibliográfica núm. 576. Clasificación temática de los textos: (4), técnicas caza y pesca).

Notas

[1] Casares y Sánchez, Julio. *Cosas del lenguaje, etimología, lexicología, semántica*, Madrid, Espasa Calpe, 1961, 236 p. (Colección Austral; 1305).

[2] Véase al respecto Alvar Esquerra, Manuel. *Proyecto de lexicografía española*, Barcelona, Planeta, 1976, 271 p. (Ensayos/Planeta. Lingüística y crítica literaria).

[3] Tratado ampliamente en Haench, G., *et al. La lexicografía: de la lingüística teórica a la lexicografía práctica*, Madrid, Gredos, 1982, 563 p. (Biblioteca románica hispánica III. Manuales; 56)

[4] Información basada en los apuntes tomados durante el *Curso de lexicografía* impartido por el Dr. Luis Fernando Lara en El Colegio de México en el periodo 1990-1991, y de Lara, Luis Fernando "El objeto diccionario", pp. 21-38, en su *Dimensiones en la lexicografía: a propósito del Diccionario del Español de México*, México, El Colegio de México, 1990. 249 p. (Jornadas; 116).

[5] Véase Meyriat, Jean y Micheline Bauchet. *Guía para establecer centros de documentación en ciencias sociales en los países en vías de desarrollo*, México, Universidad Nacional Autónoma de México, Instituto de Investigaciones Sociales, 1973, 128 p.

[6] Martinet, André. *Elementos de lingüística general*, 2a ed., Madrid, Gredos, 1972, 244 p. (Biblioteca románica hispánica III. Manuales; 13).

[7] *Ibid.*

[8] *Saussure, Ferdinand. Curso de lingüística general*, Buenos Aires, Losada, 1945, 378 p. (Filosofía y teoría del lenguaje).

[9] Véase Nocetti, Milton A. "Línguas naturais e linguagens documentárias: traços inerentes e ocorrências de interação", en *Revista biblioteconomia*, vol. 6, núm. 1 (1978).

[10] Sparck Jones, Karen. *Linguistic and information science*, New York, Academic Press, 1973, 244 p. (Library and information science).

[11] Al respecto existe el antecedente en Perales Ojeda, Alicia. "La intersección lingüística en la ciencia de la información", pp. 184-192, en Congreso Internacional sobre el Español de América (2o: 1986: México). *Actas*, México, Universidad Nacional Autónoma de México, Facultad de Filosofía y Letras, 1986.

[12] Sparck Jones, Karen. *Op. cit.*

[13] "The Unisist draft on indexing principles: test and coments", pp.

29-34, en *International classification*, vol. 4, núm. 1 (may 1977).

[14] Luhn, H. P. "A statistical approach to mechanized encoding and searching of literary information", pp. 309-317, en *IBM journal*, vol. 1, núm. 4 (oct. 1957); "The automatic creation of literature abstracts", pp. 159-165, en *IBM journal of research and development*, 2 (1958); y "Keyword-in-context index for technical literature (kwic index), en *IBM advanced systems development division report*, 12 (1969).

[15] Steinacker, Ivo. "Indexing and automatic significance analysis", pp. 237-241, en *Journal of American Society of Information Science*, vol. 25, núm. 4 (jul./ago. 1974).

[16] Robredo, Jaime. "Documentação de hoje e de amanhã", *Brasília ABDF*, 8 (1982).

[17] Vickery, B. C. "Structure and function in retrieval languages", pp. 69-82, en *Journal documentation*, vol. 27, núm. 2 (1971).

[18] Robredo, Jaime. "A indexação automática de textos: o presente já entrou no futuro", pp. 236-274, en Machado, U. D. *Estudos avançados em biblioteconomia e ciência da informação*, Brasília, ABDF, 1982, vol 1.

[19] Travis y Fidel, R. "Subject analysis", pp. 123-157, en *Annual review of information science and technology*, 17 (1982).

[20] Luhn, H. P. "A statistical... *Op. cit.*

[21] Sparck, Jones, Karen. "Indexing term weighting", pp. 619-633, en *Information storage and retrieval*, vol. 9, núm. 11 (nov. 1973).

[22] Soergel, Dagobert. "Automatic and semi-automatic methods as an aid in the construction of indexing languages and thesauri", pp. 34-39, en *International classification*, vol. 1, núm. 1 (may 1974).

[23] Salton, G. "Automated language processing", pp. 169-199, en *Annual review of information science and technology*, 3 (1968).

[24] Luhn, H. P. "A statistical... *Op. cit.*

[25] Salton, G. y Yang, C. S. "The specification of term values in automatic indexing", pp. 351-372, en *Journal of documentation*, vol. 29, núm. 4 (dec. 1973).

[26] Robertson, S. E. y Sparck Jones, J. "Relevance weighting of rese-

arch terms", pp. 129-146, en *Journal of the American Society for Informati[illegible]cience*, 27 (1976).

Véase Martínez Márquez, Alejandro. *Revisión del estado actual de la atización de los procedimientos de almacenamiento y de recuperación [illegible]rmación documental*, [s.p.i.], *ca.* 100 h.

Véase Kochtanek, Thomas Richard. *A general method for identifying r documents sets from a know relevant document*, Michigan, University [illegible]films International, 1984, 97 p.

[illegible]tos que corresponden al DEM se obtuvieron de Lara, Luis Fernando [illegible]oticia del diccionario del español de México", pp. 63-66, en *Boletín editorial de El Colegio de México*, núm. 33 (sept./oct. 1990).

[30] Véase detalladamente en Lara, Luis Fernando *et al. Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, pp. 7-39, (Jornadas; 89).

[31] Consúltese El Colegio de México. Diccionario del Español de México. *Corpus del español mexicano contemporáneo (Cemc)*, México, El Diccionario, 1975, *ca.* 100 h.

[32] Esto lo detalla su autora en García Hidalgo, María Isabel. "La formación del analizador gramatical del DEM", pp. 85-115, en Lara, Luis Fernando *et al. Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, (Jornadas 89).

[33] Ham Chande, Roberto. "Del 1 al 100 en lexicografía", pp. 41-43, en Lara, Luis Fernando *et al. Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, (Jornadas; 89).

[34] García Hidalgo, María Isabel. *Op. cit.*

TERCERA PARTE

PERTINENCIA EN LA INFORMACIÓN PARA LA LEXICOGRAFÍA

JUICIOS EN LA SELECCIÓN DEL VOCABULARIO QUE SE INCLUYE EN LOS DICCIONARIOS

Considerando que la información pertinente será aquélla obtenida del sistema de recuperación que satisface las necesidades de información de los usuarios y que la información relevante es la información que corresponde semánticamente a la pregunta formulada, es necesario identificar *lo que se busca de la información* en el presente estudio, razón por la cual resulta importante recordar lo expuesto por Luz Fernández[1] acerca de que

> todo diccionario monolingüe, y por tanto también el diccionario general de lengua, se puede definir como un catálogo sistemático de unidades léxicas pertenecientes a una lengua, presentadas de acuerdo a un orden codificado que permite su consulta. Cada unidad léxica sirve de *entrada* al enunciado que desarrolla la información gramatical y semántica sobre ella.

y también lo que esta misma autora argumenta respecto a que

> El conjunto de entradas ordenadas y sometidas a una lectura vertical se denomina *nomenclatura*. La nomenclatura es por lo tanto, un conjunto de elementos lingüísticos

que ha sido seleccionado de acuerdo con sistemas establecidos de antemano: es decir, constituye una estructura a la que llamaré *macroestructura* traducción del término francés *macrostructure* empleado por Rey-Debove.[2]

Las anteriores consideraciones hacen comprender que es necesario cumplir con una serie de requisitos para que una palabra sea incluida como entrada en un diccionario, lo que también presupone un criterio o juicio de selección para poder decidir de entre la ilimitada cantidad de palabras que se manifiestan en una lengua, cuáles corresponden y deben integrarse al tipo de diccionario que se está elaborando y cuáles tendrán que dejarse pendientes o, en su caso, omitirse definitivamente.

La pertinencia de los vocablos como requisito de selección

La diferencia cuantitativa que se observa en la nomenclatura, o número de entradas que tienen las obras de consulta llamadas diccionarios, se debe fundamentalmente a tres factores:[3]

1) diversos criterios de selección del vocabulario;
2) distintas maneras de presentar las unidades lexicográficas; y
3) diferentes interpretaciones del concepto de entrada lexicográfica.

El léxico o vocabulario de una lengua tiene el problema de que, aunque por una parte sea una estructura, por otra, el intento de registrar y describir las relaciones existentes en esa estructura se ve frustrado, porque el léxico es una estructura abierta que crece constantemente, convirtiéndose al vocabulario en un inventario ilimitado.

Este carácter ilimitado del léxico determina la falta de su registro completo (inventario incompleto).

El lexicógrafo se encuentra entonces ante un problema principal: ¿Qué parte de ese caudal léxico reunir? sin olvidar que todo diccionario presenta su nomenclatura como una estructura, o sea una macroestructura, y no como una yuxtaposición de elementos léxicos arbitrariamente elegidos.

Como el léxico total de una comunidad lingüística resulta ilimitado, los lingüistas han propuesto distinguir un conjunto léxico opuesto a éste, que llaman léxico común, constituido por el vocabulario común a todos los hablantes, o mejor, como el léxico usado por una comunidad más o menos extensa en la cual las necesidades de comunicación se imponen sobre los particularismos que representan un vocabulario muy reducido; recordemos que se calcula que el vocabulario de uso activo y diario de una persona culta no pasa de cinco mil vocablos. A este nivel, el léxico fundamental o común tiene la propiedad de poder ser inventariado.

El registro limitado de este vocabulario común y su descripción caracterizan al tipo de diccionario *básico*, a diferencia de otros diccionarios, que incluyen este conjunto léxico y, ampliándolo con el léxico del total de lengua que emplean los usuarios a los que va destinado, dan como producto los diccionarios *generales de la lengua*.

Criterios para la selección de vocablos pertinentes en los diccionarios

Hay cuatro criterios que determinan de manera decisiva la selección de entradas de un diccionario. A los tres primeros se les ha llamado "externos":[4]

1) Su finalidad (descriptiva, normativa, etcétera)

2) El grupo de usuarios al que va destinado (especialistas, traductores, alumnos de bachillerato, público culto, etcétera)

3) Su extensión

4) El método de selección de unidades léxicas según principios lingüísticos, pero siempre de acuerdo con los otros tres criterios. Este criterio es de índole "interna".

1) El criterio de la finalidad descriptiva y normativa del diccionario a través de la clasificación de usos de la lengua o "estratificación"

Según Antonio Alcalá,[5] la función cognoscitiva y la función comunicativa entre los hombres son realidades cotidianas, y a esto añade que se puede afirmar que elaboramos algo en la mente y lo comunicamos fuera de nosotros. Tradicionalmente se ha asignado al lenguaje la función de comunicar las percepciones organizadas por el conocimiento, es decir, que en el lenguaje humano se presupone un emisor y un receptor de mensajes. El primero lo elabora, lo codifica y además lo emite; el segundo lo recibe y lo decodifica, además de que en algún momento puede realizarse esto en sentido opuesto, de tal manera que el emisor se convierta en receptor del que antes era oyente.

Por otro lado, el ser humano aprende a hablar, posteriormente a leer y escribir, es decir, primero inconsciente y paulatinamente va imitando los sonidos que escucha de boca de los mayores, articulando palabras y uniéndolas unas con otras. Después, su información lingüística consiste, al ir a la escuela, en aprender a leer y a escribir, con esto aprende, por medio de un sistema de signos gráficos, a manejar la lengua,[6] que es, según la definición del *Diccionario básico del español de México*, el uso del sistema de signos fónicos o gráficos con el que se comunican los miembros de una comunidad humana o que determina cierta situación de comunicación.

La *lengua hablada* tiene un fin determinado: ser por excelencia el instrumento comunicativo del hombre, y la *lengua escrita* es el vehículo más apto para la conservación del pensamiento y para la transmisión del mismo. En esencia, estas dos clases de lengua son lo mismo, manifestadas por signos de diferente naturaleza: sonidos y grafías. La lengua escrita está supeditada a la lengua hablada.

El hombre, desde la invención de la escritura, ha podido atesorar las más variadas experiencias a través de su historia, por medio de la lengua, que es un producto cultural. Sin embargo, no todos los hablantes tienen la habilidad, conocimiento o destreza en el uso de la lengua: es más, se asegura que no hay persona alguna que conozca a p[illegible]cud este sistema.

El nivel lingüístico de un hablante está determinado por el uso de la lengua; es decir, al hablar, el hombre no utiliza la totalidad del sistema, sino solamente fragmentos, los necesarios para establecer la comunicación o interacción; se vale entonces de una parte del sistema que satisface sus necesidades de hablante, que responde a su grado de habilidad y conocimiento de la lengua. Cuando el hombre se apropia solamente de aquellos elementos del sistema lingüístico que le son útiles, integra sectores sociales con intereses afines que en su conjunto se pueden distinguir, identificar y en donde se localiza su nivel lingüístico.

En la sociolingüística, que es la disciplina que estudia las relaciones entre la lengua y la sociedad, se pueden apreciar tres niveles lingüísticos básicos: el popular, el familiar y el culto, aunque puede haber tantos niveles como actividades humanas existan, y de hecho hay más niveles mientras más necesidades tiene el hombre.

Los niveles de lengua o etiquetas de uso social en el DEM. Con ciertas diferencias entre los distintos diccionarios,[7] se puede decir que todos ellos hacen distinciones entre la lengua culta la cual no marcan, y los otros niveles de lengua, que sí marcan: lengua informal, lengua familiar, lengua popular, lengua vulgar, etc. Estas calificaciones sociales de uso del léxico se denominan *niveles de lengua* y se refieren a la función propia de la lengua dentro de la sociedad, por lo que no tienen relación alguna con las clases sociales.

Las marcas o etiquetas de uso corresponden a las distinciones que se hacen sobre el cultivo del idioma, más que a una determinada clase social. Es decir, que la estratificación a que se refieren las marcas de niveles en un diccionario, son jerarquizaciones que parten del dominio de la lengua culta, el cual a su vez se considera el modelo de corrección, por lo que no corresponde a distinciones socioeconómicas.

En el DEM, se entiende como *lengua culta*[8] "el uso de un idioma en la comunicación intelectual de sus hablantes, uso lo suficiente-

mente fijo como para permitir un amplio entendimiento entre usuarios." Los usuarios de la lengua culta puden pertenecer también a clases sociales bajas que, aunque no la lleguen a utilizar am[illegible]mente, la utilizan para comunicarse. La lengua culta, además es lo suficientemente flexible como para aceptar todas las innovaciones que impone la vida cultural de la comunidad. La lengua culta es para el DEM, en ese sentido, "el registro sociolingüístico de la lengua española en que: *a*) predomina la función referencial sobre las otras funciones del lenguaje, y *b*) se efectúan las comunicaciones lingüísticas de la mayor parte de los hispanohablantes educados."

Por ser la *lengua culta* un concepto lingüístico de suma importancia para todo diccionario general de lengua y por su valor respecto a su jerarquía o peso entre los niveles de la lengua, fue tomada como punto de partida por los lexicógrafos del Diccionario del Español de México (DEM), con objeto elaborar un modelo de usos sociales del español mexicano, es decir, un modelo que albergara todos los posibles registros en que se realizan en el español mexicano a partir del conocimiento de la comunidad lingüística mexicana y de los niveles de la lengua que tienen como etiquetas las de: popular, elevado, coloquial, etcétera.

La lengua culta para el DEM se equipara al concepto de lengua estándar propuesto por Paul L. Garvin y que la Escuela de Praga denominaba lengua literaria o lengua escrita. Sin embargo, los lexicógrafos del DEM hicieron una distinción aún más detallada entre la lengua estándar y la lengua culta para lograr calificaciones de uso del español de México mejor definidas con respecto a una visión global de los usos sociales de la lengua.

Al *español estándar* se le consideró como un español uniforme en todo el país, resultado de la poderosa influencia no sólo de la educación, sino también de la radio, la televisión y la prensa. Un español que se caracteriza por ser *general* en todas las regiones de México, el cual, aunque se origina en las ciudades, lo que los antropólogos llaman cultura urbana, se difunde ininterrumpidamente a partir de

los principales centros de irradiación del país, entre los que destaca la ciudad de México.

Al nivel elevado del español estándar se le consideró como la *lengua culta*, en el que se ubicó a los textos de literatura, textos científicos, de periodismo, ciencia, técnica, discurso político, religión y habla culta.

Al nivel del español mexicano estándar que se desvía de la lengua culta y es considerado más familiar, más del dominio popular, lo llamaron *lengua subculta*, en donde se ubicaron los textos de literatura popular, habla media y lírica popular.

Por otro lado en contraposición a la lengua estándar, se determinó el nivel denominado *lengua no estándar*, que corresponde a textos del uso del español poco extendido, limitado a ciertas regiones geográficas (dialectos del español mexicano), documentos antropológicos o a ciertos grupos sociales cerrados, como son las jergas del hampa, de algunas profesiones, etc., y, por último, también se incluyeron textos de habla popular.

Según lo anterior, se puede decir que de un esquema general se pasó a la implementación específica de cada tipo de lengua en géneros y de éstos a su representación en textos.

De acuerdo con esta estratificación cultural de la lengua, se tiene como primer nivel de clasificación el de la *lengua culta*, que sirve como marco de referencia y modelo de corrección, el cual no es marcado por el diccionario. El segundo nivel de clasificación correspondió a la *lengua subculta*, conformada por el léxico que se caracteriza por no constituir un marco de referencia prestigioso para los hablantes y que es considerado como una "incorrección cultural" en los diccionarios normativos; es conocida como *nivel popular* y se marca generalmente con las etiquetas de popular, informal, coloquial, etc. El tercer nivel de clasificación resultó ser la *lengua no estándar*, donde se encuentran principalmente dos tipos de léxico correspondientes a grupos sociales cerrados o de constante movimiento con tendencias a los llamados lenguajes secretos, donde se

Niveles de lengua en el Corpus del español mexicano contemporáneo (Cemc), que tiene una función referencial.

Lengua		*Nivel de lengua*
ESTANDAR 1. General (geogr.) 2. Urbana (Social.) 3. Irradiadora	culta	a. vocabulario intelectual y rico. b. sintaxis rica. c. modelo de corrección.
	sub-culta	a. vocabulario no intelectualizado. b. sintaxis limitada. c. desviaciones del modelo de corrección.
NO ESTANDAR 1. Limitada (geogr.) (sociol.) 2. rural (regional) urbana (grupos cerrados) 3. poco irradiadora	dialectal	a.vocabulario no intelectualizado, pero rico. b. sintaxis regional. c. modelos propios (?)
	jergal	a. vocabulario limitado (terminologías) b. sintaxis pobre. c. sujeta a modas.

FUENTE: Lara, Luis Fernando. *Investigaciones lingüísticas en lexicografía*, p. 24.

Niveles de lengua en el Corpus del español mexicano contemporáneo (Cemc), que tiene una función referencial.

Lengua		*Nivel de lengua*
ESTANDAR 1. General (geogr.) 2. Urbana (Social.) 3. Irradiadora	culta	a. vocabulario intelectual y rico. b. sintaxis rica. c. modelo de corrección.
	sub-culta	a. vocabulario no intelectualizado. b. sintaxis limitada. c. desviaciones del modelo de corrección.
NO ESTANDAR 1. Limitada (geogr.) (sociol.) 2. rural (regional) urbana (grupos cerrados) 3. poco irradiadora	dialectal	a.vocabulario no intelectualizado, pero rico. b. sintaxis regional. c. modelos propios (?)
	jergal	a. vocabulario limitado (terminologías) b. sintaxis pobre. c. sujeta a modas.

FUENTE: Lara, Luis Fernando. *Investigaciones lingüísticas en lexicografía*, p. 24.

tiene en cuenta que los dialectos como las jergas resultan generalmente poco capaces de irradiar sus características a grandes zonas del país, y aunque no dejó de considerarse este inconveniente, se pensó que este nivel de lengua puede formar marcos de referencia para el sentido de la corrección lingüística de los hablantes de este tipo de lengua.

El análisis de los tipos de textos que se producen en México y su estratificación interna en niveles de lengua, fue el punto de referencia para la formación del *corpus*, de donde se obtendría la documentación útil para la elaboración del Diccionario del Español de México.

2) El grupo de usuarios al que va destinado

Otro criterio de selección de vocablos de suma importancia lo representa el hecho que el diccionario que se elabore va a estar orientado a satisfacer las necesidades de un sector determinado de la población, el cual puede estar conformado por público en general, público culto, por estudiantes de diversos niveles de educación, o por especialistas de alguna área determinada del conocimiento. Es muy importante este criterio de selección, pues a partir de este objetivo se puede establecer mejor: el nivel de lengua que se presenta en los artículos, los vocablos que contendrá la macroestructura, las marcas o etiquetas de uso de la lengua, los ejemplos, la profundidad con que se trate cada definición, etcétera.

3) La extensión

Este criterio de selección de vocablos suele ser un factor limitante, principalmente originado por carencias de espacio, de tiempo y de dinero. El grado de resolución que se tenga al respecto de estas carencias, repercutirá seguramente en la cobertura y en la buena o mala calidad que se pueda observar en un diccionario terminado.

4) Los principios lingüísticos de selección[9]

Teniendo siempre presente los tres criterios externos (la finalidad del diccionario, sus usuarios y el espacio disponible), la selección de palabras o unidades léxicas, según principios lingüísticos, serán esencialmente determinadas por la frecuencia de uso, la distribución de frecuencias y la disponibilidad de las unidades léxicas.

Frecuencia de uso. Este tipo de selección de entradas está basado en el criterio de frecuencia, y ésta se puede determinar por el análisis estadístico de un *corpus* o conjunto cerrado de textos. El análisis estadístico de las palabras recogidas por un *corpus* indicará qué palabras se usan con frecuencia suficiente para incluirlas en el diccionario, especialmente en el caso de neologismos y tecnicismos. Sin embargo, los lexicógrafos reconocen que la frecuencia establecida de acuerdo con un *corpus* tiene sus puntos débiles, ya que está determinada por la temática utilizada, y que un corpus es una muestra representativa de tipo estadístico, la cual seguramente contendrá lagunas que cubrir.

La nomenclaturas de los diccionarios, además de corresponder a prescripciones sobre el buen uso del idioma, pueden formarse con base en el mayor arraigo que presentan las palabras y locuciones en el uso de los hablantes de una comunidad, es decir, se puede contar con diccionarios cuyas macroestructuras dependan del uso de las palabras. Con este criterio en la selección de unidades léxicas, el diccionario general tiende a reflejar mejor la realidad del léxico de una lengua (diccionario descriptivo de vocablos frecuentes) que aquéllos limitados únicamente a criterios puristas y casticistas (diccionario normativo).

Para conocer la frecuencia de los elementos léxicos y hacer así una selección entre los más frecuentes, se practican dos métodos: uno subjetivo y parcial, que puede denominarse tradicional, y otro objetivo, el análisis estadístico.

El *método tradicional* consiste en reunir las voces consideradas

como las empleadas por el hablante de cultura media en su lenguaje hablado y escrito. Este léxico se reúne de los diccionarios ya existentes y a veces, también, de materiales lexicológicos y lexicográficos.

El *método estadístico*[10] *de las frecuencias del vocabulario* es considerado por muchos lingüistas como el único que puede proporcionar objetivamente las unidades léxicas empleadas por los usuarios de la lengua.

Como la estructura léxica tiene un carácter ilimitado y abierto, el examen estadístico se puede efectuar a partir de un conjunto limitado de textos llamado *corpus*. El *corpus* elaborado por los lexicógrafos es una muestra de textos producidos por los hablantes que, para ser realmente efectiva, debe cumplir con los requisitos siguientes: como primer paso, ser representativa del habla y tener suficiente riqueza de material léxico, asegurando de esta manera la aparición de una cantidad considerable de vocablos, no sólo frecuentes, sino también de baja frecuencia pero de gran utilidad para cualquier hablante; es decir, que contenga los vocablos llamados disponibles, así como el vocabulario técnico y científico común; y eliminar dentro de lo posible, el léxico de orden temático y estilístico que se caracteriza por su uso restringido.

El siguiente paso consiste en medir la frecuencia que tiene cada palabra en todo el *corpus* y la frecuencia de la misma palabra dentro de un género o subconjunto de textos similares, así como la dispersión o porcentaje de ocurrencias de un vocablo entre géneros distintos. De esta manera se obtienen las cifras que distinguen los vocablos que sólo se emplean en relación directa con estilos y temas particulares, de aquellos utilizados en varios campos y temas. La presencia de este último tipo de vocablos es mucho más importante para los lectores de un diccionario general de lengua —como pueden ser los adultos— que las palabras de uso restringido a ciertos temas o estilos. Este razonamiento fundamenta la importancia de la presente investigación.

Para corregir los resultados unilaterales del método estadístico,

se puede utilizar el *criterio de distribución de las unidades léxicas en los diferentes textos aprovechados*. Este método se tratará ampliamente en la cuarta parte de este estudio, y es el que se tomó como base para establecer las relaciones entre la lexicografía y la bibliometría.

Por otro lado, para corregir las deficiencias de la pura estadística matemática y también del criterio de la distribución, se ha introducido el *criterio de la disponibilidad*, es decir, el recurrir a la selección convencional de cierto número de palabras que se consideran constituyentes del discurso de los hablantes en una situación tipo en la que se desarrolla un tema. La noción de disponibilidad se opone a la de frecuencia. Se denomina vocabulario disponible al conjunto de palabras con una frecuencia baja y poco estable, pero usuales y útiles, que están a disposición del hablante.

Las palabras disponibles son palabras usuales, importantes y necesarias desde el punto de vista temático, real y situacional, es decir, dependen del tema de conversación. Estas palabras principalmente son sustantivos o palabras sustantivadas, que tienen una significación concreta o material, por lo que tienen poca frecuencia si se les compara con palabras gramaticales (artículos, pronombres, etc.), o con los verbos; sin embargo, el hablante necesita disponer de ellas en determinadas situaciones; por esta razón los lexicógrafos seleccionan e incluyen palabras de este tipo en la integración del vocabulario básico correspondiente a una lengua o idioma.

Conclusiones

Los lexicógrafos realizan una selección de textos, para luego clasificarlos con fines lexicográficos por medio de agrupamientos de literatura temática; esta clasificación está fundamentada en los niveles de la lengua que la sociolingüística acepta como básicos. Posteriormente se procede a almacenarlos junto con sus representaciones, en la memoria de una computadora para, finalmente, poder recupe-

rarlos bajo búsquedas específicas, es decir, con la idea de que la información recuperada transmita información relevante al lexicógrafo o a cualquier usuario que llegue a utilizar el sistema de recuperación diseñado para ese fin.

Delimitación del juicio de pertinencia utilizado en esta tesis

En cualquier búsqueda de información intervienen muchos factores para determinar si la información recuperada es pertinente, relevante o irrelevante. En el caso de un diccionario, la pertinencia corresponde a la información recuperada que da satisfacción a la necesidad que originó la búsqueda. En cuanto a la relevancia, será considerada como una medida de efectividad cuando la información que se trasmita del archivo llamado Cemc, provoque cambios en el archivo del receptor (lexicógrafo); dichos cambios serán adiciones, eliminaciones o reorganizaciones del diccionario a partir de la pertinencia de la información recuperada del sistema.

Tomando en cuenta lo anterior, es imprescindible para este estudio el significado del léxico para la estadística lingüística, ya que de éste se desprende el juicio de pertinencia manejado en este estudio, que posteriormente se asociará a las distribuciones bibliométricas de la literatura temática, por lo que a continuación se describe:

> Para la estadística, el léxico de una lengua es el resultado de la unión de los léxicos individuales de los hablantes... el léxico individual depende de una multitud de fenómenos que van desde la edad y el sexo del hablante, hasta las diferentes peculiaridades de su educación y de su actividad diaria, por lo que cada hablante conoce un léxico distinto. El resultado de esto es que, en estadística lexicológica, el conjunto del léxico no solamente no se puede identificar en su totalidad, sino que además puede variar de acuerdo con el tipo de hablantes cuyos léxicos particulares se han investigado. Dadas las circunstancias, solamente se puede definir el *léxico común del español mexicano* como una intersección de léxicos individuales.[11]

La información bibliográfica pertinente en el DEM ayuda a do-

cumentar o testimoniar el uso y frecuencia de palabras o vocablos utilizados por los hablantes mexicanos, como pertenecientes al *léxico común del español mexicano*, principalmente a través de textos que al haber sido ordenados por temas, posibilitan el hacer estudios bibliométricos que tienen su fundamento para evaluar la información desde aspectos cuantitativos como:

a) la información recuperada relevante, que es la que tiene mayor frecuencia, de donde se explica el hecho de poder aplicar el estudio bibliométrico de la ley de Bradford/Zipf;

b) la información recuperada relevante, que es la que tiene mejor distribución en la literatura temática, también utilizando la ley de Bradford/Zipf;

c) La información biliográfica, estadísticamente pertinente, puede resolver documentalmente el problema de la selección de palabras que se incorporen como entradas en un diccionario de tipo básico. Esto al establecerse como criterio de selección el requerimiento de que las palabras pertinentes pertenezcan al léxico común del español usado en Mexico, o sea, de una intersección de léxicos individuales de los hablantes.

Notas

[1] Fernández Gordillo, Luz. *La problemática de las macroestructuras en el dicionario general*, México, La autora, 1982, 207 h. Tesis (Licenciada en Letras Hispánicas) Universidad Nacional Autónoma de México, Colegio de Letras Hispánicas.

[2] *Ibid.*

[3] *Ibidem.*

[4] Véase Haench, G. *et al. La lexicografía: de la lingüística teórica a la lexicografía práctica*, Madrid, Gredos, 1982, 563 pp. (Biblioteca románica; hispánica III. manuales; 56).

[5] Alcalá, Antonio y Huberto Batis. *La comunicación humana y literatura*, México, ANUIES, 1973, 47 p. (Temas básicos. Área: lengua y literatura).

[6] *Ibid.*

[7] Fernández Gordillo, Luz. *Op. cit.*

[8] La información correspondinte al DEM, se tomó de Lara, Luis Fernando y Roberto Ham Chande. "Base estadística del Diccionario del Español de México", pp. 7-39, en su *Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, (Jornadas; 89).

[9] Véase Haench, G. *et. al. Op. cit.*

[10] Respecto a la utilización de la estadística véase ampliamente en Muller, Charles. *Estadística lingüística*, Madrid, Gredos, 1973, 116 p. (Biblioteca románica hispánica II. Estudios y ensayos; 201).

[11] Lara, Luis Fernando y Roberto Ham Chande "Base estadística del Diccionario del Español de México", en su *Investigaciones lingüísticas en lexicofrafía*, México, El Colegio de México, 1979, p. 15.

CUARTA PARTE

ASOCIACIÓN ENTRE LA INFORMACIÓN PERTINENTE Y LAS DISTRIBUCIONES BIBLIOMÉTRICAS

SECCIÓN 1. PONDERACIÓN DE LA INFORMACIÓN

Para demostrar la asociación que existe entre la información juzgada como pertinente y las distribuciones bibliométricas de la literatura temática como modelos del comportamiento de la información obtenida por un sistema de recuperación, resulta indispensable delimitar primero la información pertinente para luego establecer dicha asociación.

Por lo antes dicho, es que en la primera sección de esta cuarta parte se realiza una experiencia de recuperación de información pertinente, utilizando como instrumento del ejercicio al sistema de recuperación de información que pertenece al Diccionario del Español de México (DEM), con el objeto de ponderar la información recuperada y así poder en la segunda sección de esta parte efectuar el estudio bibliométrico correspondiente.

Necesidades de información

Los lexicógrafos, al seleccionar las palabras que han de integrar en

la macroestructura del tipo de diccionario en que trabajan, necesitan fundamentar su elección con principios y criterios sólidos que ayuden a resolver su problemática de manera imparcial y objetiva. Buscando este fin se han utilizado los métodos de indización basada en recursos estadísticos, como un instrumento que les ayuda a tomar decisiones en tan difícil tarea; así lo hace el equipo lexicográfico del Diccionario del Español de México (DEM) al utilizar el método llamado de *frecuencia*, para seleccionar las entradas o información pertinente que compuso la macroestructura de su primer diccionario, el *Diccionario fundamental del español de México*. Para este diccionario se eligieron las palabras más usadas por los mexicanos, al sumar la frecuencia de las palabras definidas con mayor cantidad de concordancias del sistema de recuperación de información llamado *Corpus del español mexicano contemporáneo* (Cemc). Para realizarse esto se aplicó el procedimiento estadístico de la sumatoria de frecuencias al 73% percentil, es decir que del 100% de información se eligieron las palabras a partir de una mayor frecuencia, que al irse sumando llegaran a comprender hasta el 73% del total existente, y posteriormente cada palabra con frecuencia elevada fue contada todas las veces que apareció en la colección de textos, obteniéndose como resultado su frecuencia absoluta, la cual tuvo una tendencia a ser igual o mayor a 90 frecuencias.

Con respecto al otro diccionario editado por el DEM, que se llamó *Diccionario básico del español de México*, sus entradas fueron las palabras del diccionario fundamental más otras palabras seleccionadas bajo distintos requerimientos y criterios lexicográficos internos, previamente establecidos, tales como:

1. Tener una frecuencia elevada en el *corpus*.
2. Ser palabras definientes de otras usadas.
3. Ser palabras propuestas por el Consejo de Redacción del DEM.
4. Ser palabras propuestas por el propio equipo de redacción del DEM.
5. Ser palabras que hubiesen aparecido como terminología en textos de educación primaria.

Aunado a estos requerimientos respecto a la selección de palabras, éstas debían tener documentación fidedigna de fuentes primarias y secundarias, con la que se comprobara su uso en México. De esta manera, se puede decir que se incorporó al método de frecuencia el criterio de disponibilidad léxica, para la selección y ponderación de la pertinencia de las palabras.

Por otro lado, es sabido que en la lexicografía se utiliza otro método que es llamado *criterio de distribución de unidades léxicas*, en la selección de palabras, el cual también está basado en recursos estadísticos, y además es una medida de corrección de la utilización exclusiva del criterio de "mayor cantidad de frecuencias". Este método de distribución usado en el DEM tiene sus fundamentos en el hecho de que la información o documentación referente a palabras esté bien distribuida en la literatura temática existente en el *corpus* que les sirve para documentar el uso real de las palabras en México.

Delimitación de las preguntas de consulta

La idea acerca de que la información bibliográfica pertinente es útil para resolver el problema de selección de vocabulario que debe aparecer en un diccionario básico, surgió cuando el Instituto Nacional para la Educación de los Adultos (INEA), que depende de la Secretaría de Educación Pública, sugirió que incluyeran 850 palabras en el *Diccionario básico del español de México*, para que esto fuera una ayuda útil en la alfabetización de adultos.

La sección de documentación del DEM se encargó de recibir y de controlar estas solicitudes. Aquí fue donde se verificó que estas palabras en su mayoría (no todas) no habían sido seleccionadas para su inclusión en el *Diccionario fundamental del español de México* ni en el *Diccionario básico del Español de México*, por no haber cumplido con los requisitos impuestos para la selección del vocabulario que les competía respectivamente, y por haber sido propuestas en

la última etapa de preparación y edición del *Diccionario básico del español de México*, pero que ahora, en una nueva edición de este último, muy probablemente fueran seleccionadas. Esta razón fue la que incrementó mi interés por verificar qué tan *relevante* es la información existente en el sistema de recuperación de información del DEM, para esas 850 palabras, y de esta manera tener una idea de qué tan *relevante* es esta información para responder a preguntas del léxico, al margen de las que programe el DEM con sus propios criterios.

Las palabras solicitadas habían sido seleccionadas y consideradas relevantes para la educación de los adultos por los responsables de los proyectos educativos del INEA, que propician sobre todo la alfabetización, la educación comunitaria, la participación social y la capacitación para el trabajo. Tomé a manera de consultas estas solicitudes y las evalué bajo el criterio de distribución de unidades léxicas en diferentes textos de la literatura temática a través del sistema de recuperación del DEM.

Resulta importante esclarecer por qué fueron precisamente las 850 solicitudes del INEA que se tomaron a manera de consultas para esta evaluación, por lo que a continuación se exponen las razones de esta decisión.

El INEA, entre otros objetivos, busca "alcanzar, mediante la enseñanza de la lengua nacional, un idioma común para todos los mexicanos, sin menoscabo de las lenguas autóctonas"[1] y "ofrecer a todos los individuos de 15 años y más, la oportunidad de alfabetizarse y utilizar la escritura y el cálculo básico en la vida cotidiana".

La metodología utilizada por el INEA en la alfabetización es el llamado *método de la palabra generadora*, original de Paulo Freire, que se basa en los vocablos utilizados frecuentemente por los adultos analfabetos para referirse a sus problemas, necesidades e intereses. Alrededor de ellos gira la alfabetización y de esta manera se pretende relacionar permanentemente el aprendizaje de la lectura y escritura en la discusión de los problemas que les atañen.

El método de trabajo del INEA en la alfabetización se divide en dos etapas: preoperativa y operativa.

En la etapa preoperativa, que tiene más trascendencia para este estudio, se realizan cuatro actividades:

1) Investigación del universo temático. Es decir, la detección del vocabulario utilizado por los adultos analfabetos en su entorno cotidiano, así como de sus principales problemas, necesidades e intereses, para agruparlos por temas como la salud, el trabajo, etc.; de acuerdo con cada tema seleccionado, se plantean los puntos y se tratan en una etapa de discusión.
2) La selección de palabras. Para identificar las palabras que reúnan las siguientes características:
   — Que las palabras se refieran a problemas o necesidades del adulto.
   — Que las palabras posean riqueza fonética y silábica, y que a partir de ellas se puedan generar nuevas.
3) El arreglo de palabras de acuerdo con el grado de dificultad que implique su manejo.
4) La codificación, que implica la búsqueda de palabras que pueden representarse en forma objetiva y gráfica, ya sea por medio de una fotografía o un dibujo.

En la etapa operativa se realizan las actividades de discusión, aprendizaje de la lectura, la escritura y el aprendizaje de las matemáticas.

Este perfil de los intereses del INEA, sobre las palabras y su función en la educación de los adultos, se ha presentado para aclarar el origen y fundamentos de las palabras que posteriormente fueron utilizadas como consultas, para evaluar la información pertinente y la relevante en el sistema de recuperacón de información del DEM, ya que en estos fundamentos se presentan objetivos afines a los del mismo DEM en lo relativo a la lengua común.

La existencia y utilización de palabras pertenecientes al léxico común es esencial para las metas y objetivos que se buscan tanto

en la elaboración de un diccionario básico —lo cual ocurre en el DEM— como en la educación de los adultos —lo cual busca el INEA. Al detectar esto y considerarlo útil y funcional, elegí como un juicio de pertinencia el de que la información obtenida por el sistema de recuperación de información debía pertenecer al léxico común, para así satisfacer una necesidad real, y tan bien fundamentada —está basada en principios lingüísticos de selección de vocabulario—, que resultara posible ponderar todas las respuestas a las consultas formuladas al sistema, de tal forma que no sólo se identificara claramente la información que fuese, pertinente, relevante recuperada, relevante no recuperada, no relevante recuperada, y no relevante no recuperada, sino que además a estas mismas respuestas se las pudiese ubicar detalladamente en el grado en que satisficieran o no la necesidad de información planteada desde la formulación de preguntas al sistema. De esta manera queda establecida la relevancia para el usuario (documentalista).

Al poder considerar la respuesta de las 850 consultas de información como una muestra representativa del funcionamiento del sistema de recuperación del DEM, pienso que la ponderación de la información antes explicada servirá también para establecer el porcentaje o nivel de efectividad que se alcanza en el sistema de recuperación de información del DEM, para responder con precisión sobre preguntas de léxico. De esta manera queda establecida la relevancia para el sistema.

La necesidad que dio sustentación al juicio de pertinencia estuvo originada por un principio de selección lingüística para elegir las palabras que se incluirían en un diccionario, llamado *criterio de distribución de unidades léxicas*, el cual además es una manera de corregir la selección basada exclusivamente en las palabras que tienen frecuencias altas.

En esta cuarta y última parte de la tesis se efectúa una aplicación del método de distribución de unidades léxicas utilizado como un criterio o juicio para seleccionar palabras; para esto se realizar

un análisis cuantitativo sobre la distribución de la información bibliográfica referente a palabras, contenidas éstas en textos organizados temáticamente y archivados en un sistema de recuperación. A partir de lo anterior se identifican las palabras pertinentes y las relevantes, documentalmente hablando. La pertinencia en este estudio es la información (El contenido de documentos) que satisface la necesidad de información del usuario, y la relevancia es el grado de correspondencia o correlación entre la información (un documento) recuperada y el contenido de una pregunta, es decir, se evalúa la eficacia del contacto entre una fuente y un destinatario en el proceso de comunicación.

La información pertinente con base en los principios lingüísticos de selección de vocabulario debe estar constituida por las palabras "documentadas": que usan los mexicanos en su lengua, que tengan un uso lo suficientemente fijo para propiciar la comunicación y el entendimiento entre los hablantes, que en ellas predomine una función referencial por sobre las otras funciones del lenguaje y que su uso sea uniforme en todo el país. Es decir, las palabras que tienen estas características forman parte del léxico común del mexicano, en el cual, estadísticamente hablando, los léxicos individuales de todos los mexicanos se intersectan. A partir de esta razón, las palabras recuperadas tendrían que encontrarse bien distribuidas en los textos, o sea, a lo largo de la literatura temática archivada en el *Corpus del español mexicano contemporáneo* (Cemc), para ser consideradas pertinentes. En esta etapa se utilizan sólo los datos cuantitativos de las respuestas.

Para asignar la pertinencia y la relevancia a la información recuperada en general, se utiliza el también método de indización llamado "atribución de peso", que se refiere a que si una palabra se encuentra documentada en un texto o material bibliográfico reconocido como relevante, recibe un peso mayor que el que pueda recibir una palabra existente en un documento menos relevante. Para esto se atribuirá el mayor peso a los textos pertenecientes a la lengua culta.

La exigencia o requerimiento que deben satisfacer las respuestas del sistema, es que para que la información obtenida resulte pertinente, la palabra buscada debe estar distribuida no sólo en uno de los géneros o clasificaciones de lengua antes citados, sino que además exista documentación de por lo menos una frecuencia de las palabras en cada uno de estos tres géneros —los tres a la vez—, es decir, que tenga una buena distribución en la literatura temática en que está organizada la información obtenida de textos almacenados dentro de los archivos que componen el Cemc.

Planteamiento general y estrategia

Para conseguir el objetivo propuesto, la primera problemática fue efectuar una evaluación de la información obtenida a manera de respuestas por el sistema de recuperación de información del DEM. Esta evaluación se realizó teniendo presente que dichas respuestas se dieron a partir de la solicitud de información formulada por parte del que esto escribe, bajo el carácter de consultas, y con la intención de que las respuestas mostraran la efectividad del SRI.

Las respuestas que ofrece el sistema son recuperadas en forma de concordancias que son agrupadas bajo grafías de palabras completas, palabras truncadas y raíces de palabras. Todas ellas están agrupadas en una monografía léxica para cada palabra o vocablo y de esto se pueden obtener resultados sobre la relevancia o relación entre la información recuperada y el contenido de una pregunta.

En cuanto a la estrategia de búsqueda de la información pertinente se aplicó la teoría de los conjuntos para encontrar las respuestas que se necesitaban, por lo que se interpretó la información bajo otros términos: el conjunto universo son todas las palabras almacenadas en el Cemc (1 932 000), que fueron ordenadas en 14 géneros o clasificaciones de literatura temática y que equivalen al 100%; este conjunto universo a la vez está dividido en tres subconjuntos:

la *lengua culta* como el subconjunto A (con 1 336 000 palabras), que abarca los géneros de 1 a 7 y que equivale al 65.6416%; la *lengua subculta* como el subconjunto B (con 234 000 palabras), que abarca los géneros de 8 a 10 y que equivale al 12.2776%, y la *lengua no estándar* como el subconjunto C (con 362 000 palabras), que abarca los géneros de 11 a 14 y que equivale al 22.0808 por ciento.

Considerado esto así, la operación que se hizo necesario realizar en la búsqueda de la información perteneciente al léxico común —que es la considerada pertinente en este estudio— fue la llamada "intersección".

Una vez resuelta la primera problemática, la siguiente consistió en efectuar un ordenamiento y clasificación de la información (palabras) con base en la distribución de la información de unidades léxicas en la literatura temática (textos) obtenida por medio del sistema de recuperación, pues de esto dependería la selección estrictamente documental de las palabras que se pueden incluir en un diccionario básico. En esta selección prevaleció el criterio de pertinencia y de relevancia documental que tenían las palabras, asignado éste por medio del método de indización llamado *método de atribución de peso*. Se equipara esta acción a una ponderación, donde a la información se le dio nivel de importancia según respondiera o no a la necesidad que se había planteado.

Una vez realizados estos pasos, se pudo efectuar la asociación entre la información pertinente y las distribuciones bibliométricas, objetivo principal de esta investigación.

Búsqueda de la pertinencia

A continuación se presentan los pasos que se siguieron para recuperar y delimitar la información considerada como pertinente:

- Identificación de las 850 palabras utilizadas como consultas.

- Búsqueda manual de la documentación sobre cada palabra, en los archivos del DEM, la cual fue proporcionada por el sistema de recuperación de información, y se encuentra impresa en forma de índices de concordancias, muy parecidas a los que conocemos como "kwic", integrados a cada una de las *monografías léxicas* que documentan el uso de las palabras en el español mexicano.
- Identificación de las palabras que tuvieron o no documentación (no hubo en 119).
- Verificación de información cuantitativa, que tiene como fundamento la certificación de la existencia y validez de los registros bibliográficos, en cada una de las palabras con documentación. Es decir, se cotejó la recuperación real de ítems o concordancias (véase la documentación de un diccionario en la segunda parte), al confrontarlas físicamente contra su respectivo registro de información, previamente asentado en una hoja de datos estadísticos por los documentalistas del DEM, en la cual se realizaron cambios cuando fue necesario, o, en el caso de no existir la hoja de datos estadísticos, se elaboró, anotándose primero los tipos o palabras clave indizadas que pertenecieran a la palabra o vocablo buscado, y posteriormente a cada tipo se le agregó el registro del número de concordancias que se encontraban agrupadas bajo el mismo. El registro estadístico completo hizo posible la identificación de cuántos tipos y cuántas concordancias conformaban cada palabra o vocablo en el momento de la búsqueda, lo que hizo clara también la frecuencia absoluta de cada vocablo, aunque resulta oportuno mencionar que en el futuro puede haber modificaciones respecto a esta información, por diferentes factores, de entre los que destacan: la corrección de manera manual de posibles errores en la digitalización de los textos; la incorporación de tipos que resultaron amiguos en un primer análisis automatizado y la incorporación como entradas en un diccionario de algunas de las variantes de palabras que ahora están agrupadas y forman parte de un único vocablo.

- Obtención de la frecuencia absoluta de cada palabra. Para esto fue necesario sumar las concordancias de los tipos o palabras clave que conforman cada vocablo buscado.
- Investigación con respecto a en qué clasificaciones de textos habría sido documentada cada una de las palabras. Se procedió a identificar la existencia y distribución del total de frecuencias pertenecientes a una palabra a lo largo de las 14 clasificaciones de textos en que está ordenado el archivo del sistema. Aquí se aprovechó el índice estadístico impreso llamado *Corpus del español mexicano contemporáneo: frecuencias absolutas*, que se basa en la identificación del código de la referencia bibliográfica de cada concordancia para presentar de cada tipo o palabra clave que conforma a un vocablo, la distribución de sus apariciones en las 14 clasificaciones de textos o géneros. El resultado de esta acción dio a conocer cuántas veces aparecía documentada una palabra por clasificación, y esto a su vez hizo posible identificar a qué nivel de lengua pertenecía cada concordancia recuperada perteneciente a cada palabra.

EJEMPLO DE COMPILACIÓN DE DATOS
(véase muestra completa en Apéndice I)

Documentación: "Distribución de frecuencias de los vocablos por clasificación de textos"

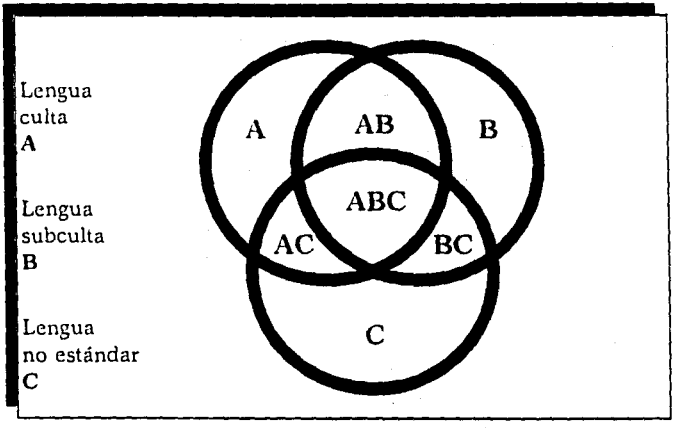
	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
abanico	4	1	3	1				2	1						12	AB
aborto*	3	1	8	4									2		18	AC
abrevar	5														5	A
abridor				1											1	A
abrojo	2														2	A
abstracto	8	6	16	2			1				1				34	AC
acaecer	2	2						1							5	AB
acalorado	1	1							1						3	AB
acanaladura			1												1	A

- Identificación de la tendencia temática, así como del nivel de lengua de las concordancias que conforman cada una de las palabras. Para esto es necesario sumar el total de frecuencias aparecidas en cada una de las 14 clasificaciones de textos. En los cuadros de resultados se puede observar la tendencia temática acumulada de la información perteneciente a 731 palabras recuperadas. (Véase el cuadro 1 y la gráfica 1 de resultados de ponderación.)
- Evaluación de la capacidad de respuesta del sistema para las palabras solicitadas desde el aspecto de la literatura temática. Para esto fue necesario sumar las palabras documentadas por cada una de las clasificaciones de textos. En los cuadros de resultados se puede observar la capacidad de respuesta por cada una de las 14 clasificaciones de textos para 731 palabras recuperadas en consulta, en las cuales, por cierto, se asemejan sus proporciones temáticas a la información temática almacenada en el sistema desde su creación. (Véase el cuadro 2 y la gráfica 2 de resultados de ponderación.)
- Clasificación de las palabras dentro del nivel de lengua llama lengua culta, si pertencen a los textos de 1 a 7.
- Clasificación de las palabras dentro del nivel de lengua llamado lengua subculta, si pertenecen a los textos de 8 a 10.
- Clasificación de las palabras dentro del nivel de lengua llamado lengua no estándar, si pertenecen a los textos de 11 a 14.
- Documentación de la palabra en los tres niveles de lengua en que está organizada la información archivada en el sistema de recuperación, lo cual, según el criterio establecido, la haría pertenecer al léxico común. Por esto mismo se hizo necesario realizar una operación de conjuntos llamada intersección, en la que bastaría que una palabra buscada existiera por lo menos con una frecuencia de aparición por cada uno de los tres niveles de lengua en que está organizado el archivo del sistema de recuperación; con esto quedaría satisfecha a la necesidad que originaba la búsqueda.

Teniendo en cuenta lo anterior, a la documentación localizada

en las clasificaciones de la lengua culta se la consideró como el subconjunto A; a las clasificaciones que comprendía la lengua subculta se las consideró el subconjunto B, y a las clasificaciones que comprendía la lengua no estándar se las consideró el subconjunto C. (Véase cuadro 3 de resultados de ponderación y apéndice I.)

Modelo de búsqueda en el Sistema de recuperación de información



Búsqueda de pertinencia de 731 vocablos en un archivo de aproximadamente 2 000 000 palabras llamado *Corpus del español mexicano contemporáneo* (CEMC).

- Aunque la operación de intersección permitiría identificar sólo un pequeño núcleo de información pertinente, la otra información (relevante e irrelevante) sirvió para ordenar las palabras en cuanto a su grado de importancia respecto a la documentación de un diccionario. Partiendo de este hecho, estableció las siguientes

jerarquías para la información relevante recuperada: 1) la primera jerarquía correspondería a la intersección de los tres niveles de lengua (léxico común). En esta jerarquía está la información pertinente que satisface las necesidades que originaron la búsqueda; 2) la segunda jerarquía correspondería a la intersección de los niveles culto y subculto; 3) la tercera jerarquía correspondería a la intersección de lengua culta con lengua no estándar; 4) la cuarta jerarquía, aunque no fuera una intersección, correspondería a la lengua culta, que es la fuente con mayor peso en la elaboración de un diccionario; 5) la quinta jerarquía correspondería a la intersección de lengua subculta y lengua no estándar.

- Para la información recuperada no relevante, asigné las siguientes jerarquías: 6) la sexta jerarquía correspondería a la lengua subculta; 7) la séptima jerarquía correspondería a la lengua no estándar.
- Respecto a las palabras no documentadas por el sistema de recuperación, con el objetivo de asignarles relevancia y no relevancia, realicé una investigación bibliográfica en 19 diccionarios de consulta que son utilizados como fuentes secundarias en el DEM. Lo anterior permitió establecer las dos últimas jerarquías de información: 8) la octava jerarquía correspondió a las palabras documentadas en diccionarios del español general y en diccionarios mexicanos, también se las consideró relevantes no recuperadas. La novena jerarquía correspondió a palabras que no tenían documentación en diccionarios, y palabras que sólo tenían documentación en diccionarios de mexicanismos, también se las consideró no relevantes no recuperadas. (Véase cuadro 3 de resultado y apéndice II.)

Resultados de la ponderación de palabras en cuanto a la relevancia de la información

Los resultados muestran el análisis realizado sobre 850 preguntas

realizadas al sistema de recuperación del DEM, las cuales representaban el 100% de la información solicitada por el usuario y a las cuales el sistema respondió de la siguiente manera:

Palabras con relevancia

- 824 palabras por sus características —intersecciones— se consideraron relevantes, lo que equivale a un 96.94% del total. De éstas se puede hacer un desglose en tres aspectos:
  - 166 palabras recuperadas pertenecen al léxico común —intersección de tres niveles de lengua— y por lo tanto son pertinentes. Equivalen al 19.52% del total. En estas 166 palabras se concentran 5 350 concordancias que es más del 50% del total de documentaciones emitidas por el sistema como respuesta.
  - 544 palabras fueron relevantes recuperadas —intersección de dos niveles de lengua o pertenecer a la lengua culta. Equivalen al 64% del total de palabras solicitadas.
  - 114 palabras consideradas relevantes, por tener registros en diccionarios del español general como en diccionaros mexicanos, no fueron recuperadas por el sistema. Equivalen al 13.41% del total de palabras solicitadas.

Palabras sin relevancia

- 26 palabras por sus características —sin intersección, y no pertenecer a la lengua culta— resultaron no relevantes, lo que equivale a un 3.05% respecto al total y se pueden dividir en dos aspectos:
  - 21 palabras no relevantes que fueron recuperadas por el sistema. Equivalen al 2.47% del total.
  - 5 palabras no relevantes no fueron recuperadas por el sistema. Equivalen al 0.58% del total.

Consideraciones

- Para que exista información documentalmente pertinente, primero deberá existir la información relevante.

- La información bibliográfica pertinente resolvió la necesidad de información que existía para identificar entre las palabras solicitadas por el INEA aquellas pertenecientes al léxico común de los mexicanos, que son las que tienen un uso lo suficientemente fijo como para propiciar la comunicación y el entendimiento entre los mexicanos. Las palabras con pertinencia documental pueden ser seleccionadas como entradas para un diccionario básico, bajo el principio lingüístico de selección de vocablos llamado "distribución".
- La información bibliográfica relevante resuelve varios problemas sobre el conocimiento de las palabras: *a*) sobre si la palabra solicitada en búsqueda al sistema, tiene documentación que sirva como testimonio de su uso y frecuencia en el habla de los mexicanos; *b*) si pertenece a un nivel de lengua determinado o a varios; *c*) en qué lugares geográficos se utiliza; *d*) qué etiquetas gramaticales puede tener una misma palabra gráfica; *e*) si tiene polisemia; etcétera.
- Desde el punto de vista de un usuario, como lo es el autor de esta tesis, los resultados sobre la información pertinente son altos, al considerar que las palabras habían sido propuestas por una institución que tiene una metodología propia para seleccionar palabras utilizadas por los adultos, aunque esto mismo hacía esperar una mayor cantidad de información pertinente en estos resultados. Por esto mismo, al analizar las palabras en un aspecto global, resulta que no es el sistema el que se muestra ineficiente, sino que la mayoría de las palabras solicitadas más bien están basadas en un criterio de disponibilidad que en uno de uso, es decir, son palabras que se pensaron útiles en circunstancias determinadas o conversaciones temáticas, y no necesariamente que fueran las que más usan los adultos.
- En cuanto a la recuperación de información, hubo una respuesta del 83.52% con información relevante sobre el total de información solicitada al sistema, lo que hace notar que el sistema tiene

un excelente alcance y nivel de información en sus archivos, que le permite cumplir su cometido de documentar palabras, aun sin ser las más frecuentes.

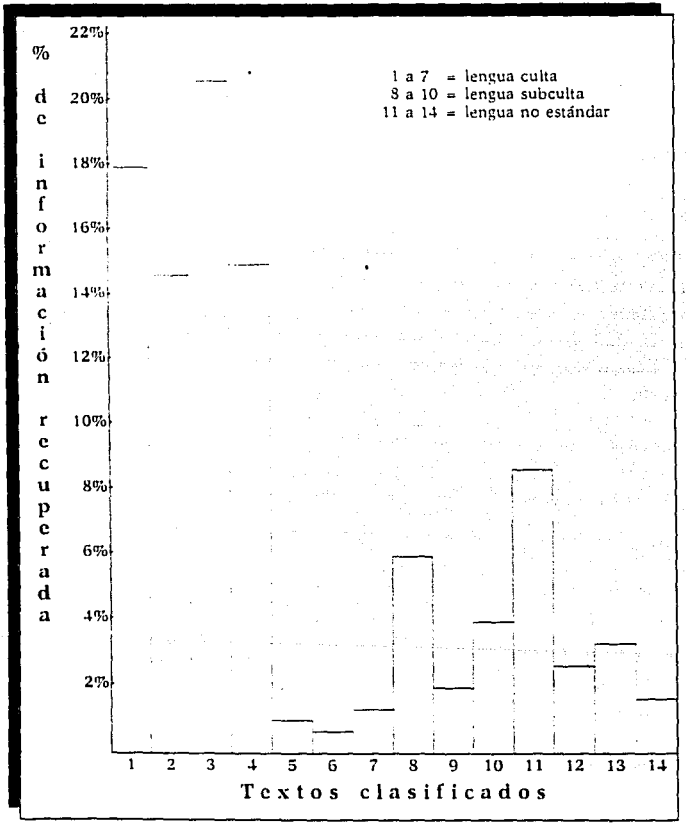
- En cuanto a la información no relevante, el sistema recuperó muy poca y con respecto a la que no recuperó, es notorio, al cotejarla con diccionarios tanto del español general como de mexicanismos, que no es información propia para un diccionario como el DEM, sino para diccionarios especializados o para enciclopedias.

Cuadro 1. Resultados de ponderación

(Porcentaje de respuestas en forma de concordancias y clasificación de textos)

Género	*Concordancias*	%	*Nivel*
1 Literatura	1 802	17.93	
2 Periodismo	1 463	14.56	
3 Ciencias	2 123	21.13	Lengua oculta
4 Técnicas	1 570	15.62	73.32
5 Discursos políticos	156	1.55	
6 Religión	84	.83	
7 Habla culta	171	1.70	
8 Literatura popular	630	6.27	
9 Habla media	199	1.98	Lengua subculta
10 Lírica popular	384	3.82	12.07
11 Textos dialectales	858	8.54	
12 Doc. antropológicos	204	2.03	Lengua no estándar
13 Jergas	232	2.30	14.55
14 Habla popular	169	1.68	
Totales	10 045	100.00	

Gráfica 1. Porcentaje de información recuperada (concordancias) con base en la clasificación de la literatura temática

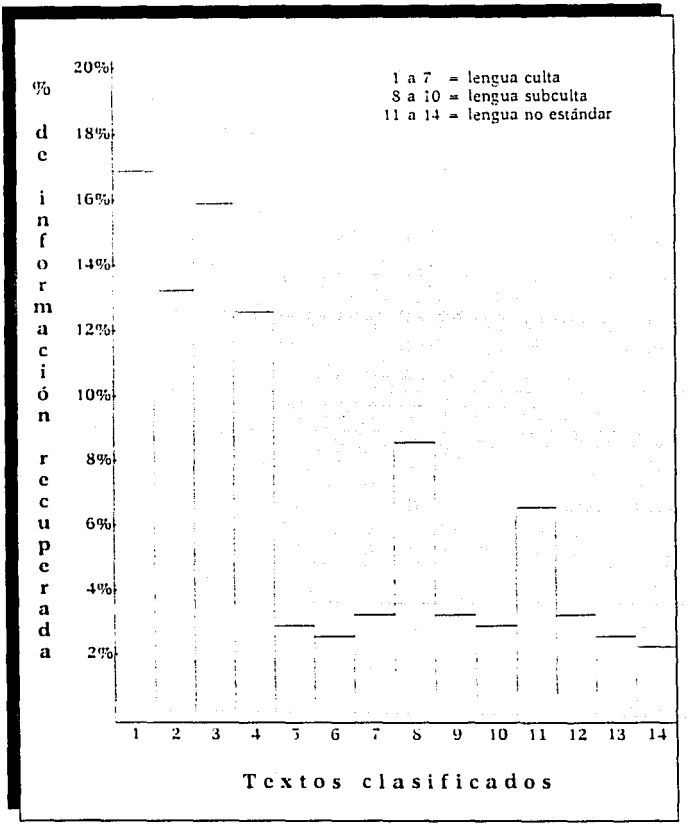


Cuadro 2. Resultados de ponderación
(Porcentaje de respuestas en forma de palabras que se obtuvieron de la búsqueda en el corpus* por clasificación de textos)

Género	*Palabras*	%	*Nivel*
1 Literatura	457	17.33	
2 Periodismo	354	13.42	
3 Ciencias	424	16.07	Lengua culta
4 Técnicas	325	12.32	67.77
5 Discursos políticos	84	3.18	
6 Religión	61	2.31	
7 Habla culta	83	3.14	
8 Literatura popular	234	8.87	
9 Habla media	93	3.52	Lengua subculta
10 Lírica popular	80	3.03	15.42
11 Textos dialectales	197	7.47	
12 Doc. antropológicos	98	3.71	Lengua no estándar
13 Jergas	75	2.84	16.75
14 Habla popular	72	2.73	
Totales	2 637	100.00	
		99.94	

* Estos resultados afectan a 731 palabras, de una solicitud total de 850, ya que de 119 palabras no hubo respuesta.

Gráfica 2. Porcentaje de información recuperada (palabras) con base en la clasificación de la literatura temática



Cuadro 3. Resultados de ponderación
(Resultado de la búsqueda de información bibliográfica y ponderación de la misma desde el aspecto de su relevancia)

Jerarquía decreciente	*Niveles de lenguaje (intersecciones)*	*Claves*	*Vocablos*
1°	culto + subculto + no estándar	ABC	166
2°	culto + subculto	AB	115
3°	culto + no estándar	AC	84
4°	culto	A	338
5°	subculto + no estándar	BC	7
6°	subculto	B	13
7°	no estándar	C	8
8°	Relevante sin respuesta	—	114
9°	no relevante sin respuesta	—	5

Evaluación del servicio de consulta del DEM

Respecto al servicio de consulta, la American Library Association[2] distingue entre *servicio directo* y *servicio indirecto*. La mayor parte del servicio directo que se brinda tanto en bibliotecas como en centros de documentación en México es del tipo de "ayuda a la consulta", que incluye la respuesta a preguntas de naturaleza factual —por teléfono, en persona o por correo— y la ayuda a los usuarios para encontrar los materiales bibliográficos que necesite.

Para este estudio, la búsqueda de información a manera de consulta directa significa la actividad en que se desarrolla una estrategia para indagar la información solicitada por el usuario, con el objeto de encontrar materiales bibliográficos o ítems sobre una palabra determinada.

En la búsqueda de información se tienen como fuente los datos sobre las palabras depositadas en textos, que son emitidos por el sistema de recuperación de información de el DEM.

La medición del éxito en el trabajo de consulta es una tarea que requiere ser efectuada con rigor y debe contar con objetivos muy determinados e instrumentos de medida apropiados, por lo que resulta oportuno recordar lo expuesto por F. W. Lancaster[3] respecto a la medición de los servicios de consulta, donde se señala que los servicios pueden evaluarse en tres niveles: efectividad, costo-efectividad y costo-beneficio.

El nivel de efectividad es la medida que se aplica en el presente estudio sobre las respuestas emitidas por el sistema de recuperación del DEM, y el criterio esencial de medición es el grado en que el sistema maximiza la accesibilidad de información, ya que de esta manera se expresa en qué medida un servicio satisface las demandas de sus usuarios, es decir, si ofrece información relevante y pertinente.

Este tipo de evaluación podría hacerse de manera subjetiva, como por ejemplo, recogiendo opiniones a través de cuestionarios o entrevistas, o de forma objetiva, como se practicará en este estudio para determinar el éxito del contacto entre el sistema y el destinatario en términos cuantitativos, partiendo de la recuperación real de la información por medio del sistema de recuperación del DEM.

Métodos y medidas de la pertinencia y de la relevancia

Existen varios métodos y medidas para formarse una opinión de los juicios de la pertinencia, de la relevancia, y para determinar las conexiones semánticas entre las preguntas y los textos relevantes.

Resulta oportuno para este estudio recordar lo que mencionó Harmon[4] respecto a la diferencia que existe, por un lado, en la *relevancia para usuario del sistema*, donde el usuario hace una evaluación a partir de la relación entre la pregunta realizada al sistema y la satisfacción de sus propias necesidades de información, y por otro lado, la *relevancia del sistema*, que es la evaluación hecha por

el sistema sobre el grado de relación entre la información contenida en el sistema y la pregunta realizada por el usuario. En esta tesis se obtienen ambas.

Como mencionó Saracevi,[5] las primeras unidades cuantitativas para medir la relevancia en sistemas fueron: *la recuperación*, que es el porcentaje de respuestas relevantes recuperadas sobre el total de respuestas relevantes existentes en un sistema, y la *precisión*, que es el porcentaje de respuestas relevantes recuperadas sobre el total de respuestas recuperadas.

Barhydt[6] y Saracevic[7] presentaron algunas medidas de evaluación de desempeño de sistemas: la *sensibilidad* es el porcentaje entre el número de documentos recuperados y el total de documentos relevantes en el archivo; la *especificidad* es el porcentaje entre el número de documentos no relevantes no recuperados y el total de documentos no relevantes en el archivo; y la *efectividad*, que es una suma de la sensibilidad y la especificidad menos una constante 1.

A partir de los anteriores conceptos se pueden identificar las variables que se presentan al evaluar un sistema de recuperación:

a) = documentos relevantes recuperados del archivo
b) = documentos no relevantes recuperados del archivo
c) = documentos relevantes no recuperados del archivo
d) = documentos no relevantes no recuperados del archivo
e) = total de documentos recuperados

Las fórmulas de las medidas de eficiencia de un sistema de recuperación son:

$$\textit{recuperación} = a \div a + c = R$$

$$\textit{precisión} = a \div a + b = P$$

$$\textit{eficiencia} = a + c \div a + b \div c + d = Ef$$

$$\textit{especificidad} = d \div b + d = Es$$

$$\textit{efectividad} = R \div Es$$

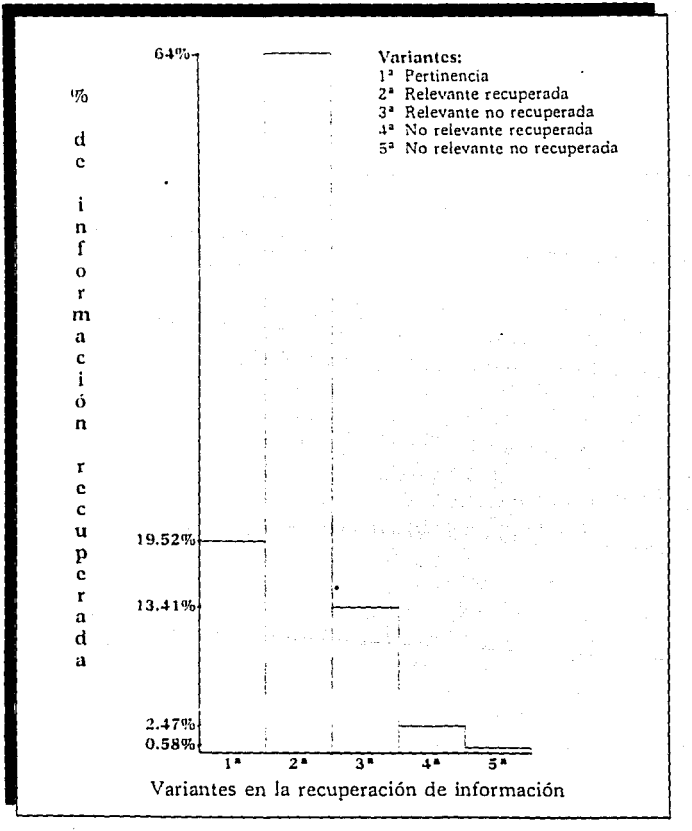
La relevancia de la *precisión* es definida por Saracevic[8] como la medida del contacto efectivo entre la fuente y el destinatario, y puede ser calculada al dividir el número de documentos relevantes recuperados sobre el número total de documentos recuperados.

Medidas de eficiencia del sistema de recuperación del DEM

Recuperación. Al dividir el número de documentos relevantes recuperados por el sistema (710), entre el número resultado de la suma de documentos relevantes recuperados, más los documentos relevantes no recuperados (710 + 14), se identifica que el sistema tuvo una *medida de recuperación* del 86.16% respecto a 850 consultas formuladas.

Precisión. Al dividir el número de documentos relevantes recuperados por el sistema (710), entre el número resultado de la suma de documentos relevantes recuperados, más los documentos no relevantes recuperados (710 + 21), se obtiene que el sistema tuvo una *medida de precisión* del 97.12%.

Gráfica 3. Pertinencia y relevancia asignada a 850 respuestas emitidas por el sistema de recuperación de información del (DEM)



El estudio bibliométrico y su asociación con la información pertinente a partir del enfoque de la distribucion en la literatura temática

Desde Bradford,[9] en la década de los treinta y luego en la de los cuarenta, ha existido interés por los artículos relevantes acerca de un tema o materia, y a partir de este interés en la bibliotecología se consolida la bibliometría como un importante instrumento de conocimiento sobre la *relevancia*, ya que por medio de la bibliometría se pueden estudiar ciertos fenómenos de comportamiento de tipo cuantiativo que ocurren acerca de la información bibliográfica, los cuales ayudan a delimitar cuál es la información relevante y la información pertinente, por un lado, y cuál no lo es, por el otro; si se toma la bibliometría como lo hace Faithone,[10] para el cual es el tratamiento cuantitativo de las propiedades del discurso escrito y el comportamiento inherente a él.

Los estudios bibliométricos, por ser de carácter cuantitativo, han descubierto leyes empíricas como la ley de Bradford, la ley de Lotka y la ley de Zipf; de éstas han surgido teorías y se han realizado observaciones cuantitativas. En estos estudios, como en los de control bibliográfico, se nota como interés fundamental el aspecto de la relevancia.

Kozachov,[11] en un intento por relacionar algunos hechos estudiados en la bibliometría con la noción de relevancia, estableció una relación entre la noción de relevancia y el proceso cognoscitivo científico, y descubrió varios aspectos de la literatura científica como: el crecimiento, la dispersión y la obsolescencia, en términos de su relación con la relevancia.

Bradford[12] también afirmó, en su *Ley de la dispersión de la literatura*, que estaba interesado por medir la capacidad de las revistas

científicas para contener artículos relevantes de una materia, es decir, estaba interesado en el modelo de una distribución estadística que describiera la relación entre cierta cantidad de revistas científicas y un rendimiento o subordinación de artículos. La observación de esto lo llevó a la idea de que la dispersión de los artículos en una materia o tema, incluidos en las revistas que los contienen, forma un patrón o modelo regular de rendimientos decrecientes, y que si las revistas científicas se ordenaran por la productividad decreciente de los artículos que contienen un determinado tema, éstas podrían ser divididas en un núcleo de revistas especializadas sobre un tema, y en varios grupos o zonas que contuvieran el mismo número de artículos que el núcleo.

Saracevic[13] sintetizó algunas distribuciones y descubrimientos en la bibliometría y los interpretó en términos de la relevancia llamándolos distribuciones relacionados con la relevancia. Partiendo de este razonamiento, la distribución de las palabras en diferentes textos temáticos, como se vio anteriormente, es una manifestación de comunicación de conocimientos, semejante a las distribuciones de los estudios bibliométricos.

Saracevic[14] también estudió la distribución de documentos recuperados, como respuestas hechas por un sistema de recuperación de información experimental, y encontró que la distribución sigue la ley de Bradford. Tomando todas las preguntas juntas, un reducido número de documentos se recuperaban constantemente como respuestas, llegando a formar un núcleo, y el resto seguía el modelo de Bradford. Luego se estudió el juicio de relevancia de los usuarios sobre las mismas respuestas recuperadas, donde la distribución de los documentos que se juzgaron relevantes se comportaron también como en la ley de Bradford.

La fuerza de los trabajos bibliométricos radica en la conexión directa entre las leyes y las teorías, por un lado, y la observación, por otro. Al respecto, Zipf[15] investigó la distribución de las palabras en un texto, y observó que la frecuencia de uso en un pequeño núcleo

de palabras tiene un patrón regular. La información contenida en estos dos últimos párrafos, sirve como antecedente directo al presente estudio.

Después de este breve resumen de literatura sobre la bibliometría, se puede utilizar una alternativa adicional que puede incorporarse en la aplicación del análisis biliométrico, que es la investigación de los problemas específicos en que se encuentra involucrada la información bibliográfica como fenómeno social en la elaboración de diccionarios —la lexicografía—, cuestión que se ejemplifica en este estudio, el cual se ubica en un determinado tiempo y espacio, y resulta factible a partir del hecho de que los lexicógrafos, al desempeñar su tarea de elaborar diccionarios, utilizan recursos bibliográficos en la documentación de diccionarios; estos recursos pueden estar sujetos a la bibliometría, entendiendo a ésta según Pritchard,[16] como la aplicación de métodos estadísticos a los libros y otros medios de comunicación escrita.

De lo anterior se puede deducir que al utilizar un análisis bibliométrico, como una adaptación a la ley de Zipf, se pueden obtener generalidades en la economía de la información, útiles como apoyo y referencia para la comprensión de los fenómenos existentes en la disciplina lexicográfica, área donde se ubica y desarrolla este estudio, ya que la palaba impresa llega a tener un papel que, si no es decisivo, por lo menos sí es indicativo con respecto a la aparición y la frecuencia de uso de las palabras cuando éstas están documentadas por medio de textos, ya que siguen un patrón regular respecto a su utilización. Estas mismas distribuciones bibliométricas están asociadas con cierto juicio de pertinencia y de relevancia asignado a la información sobre las palabras, el cual tiene su origen en la teoría sociolingüística de la estratificación social del uso del lenguaje y, por extensión, también de las palabras que lo conforman. Este hecho permite jerarquizar la información bibliográfica obtenida a partir de textos, en relación con la buena o mala distribución que tengan las palabras recuperadas, dentro de los distintos niveles de

lengua existentes en los léxicos individuales de hablantes depositados en determinados textos organizados de manera temática. De estos fundamentos y juicios se establece que la información bibliográfica pertinente es la que satisface la necesidad de información al estar documentada y tener buena distribución en las clasificaciones —ser del léxico común— que propone la sociolingüística con respecto a tres niveles de la lengua: la lengua culta, la lengua subculta y la lengua no estándar. A la información que está contenida en estas tres se le atribuye el juicio documental de *pertinencia*, el cual resulta un principio estadístico necesario para incorporar entradas de palabras en un diccionario de tipo básico.

Tomando como fundamento esta teoría, los lexicógrafos del Diccionario del Español de México (DEM) formularon el algoritmo estadístico del programa que se utilizó para recuperar el léxico archivado del sistema de recuperación que han desarrollado, lo que a su vez permite afirmar que existe insertado un juicio de relevancia de la información en el propio sistema de recuperación, juicio que es exclusivamente *documental* ya que en la práctica se evalúan más elementos para tomar decisiones concernientes a las palabras que han de incluirse o no en un diccionario.

La dispersión

Siguiendo la idea de Bradford/Zipf y adaptándola a los intereses de este estudio respecto a la información relevante en la lexicografía, se puede pensar que un número relativamente pequeño de palabras recuperadas por un sistema de información tiene las mayores posibilidades de ser pertinentes en relación con el total solicitado en búsqueda y que a partir de este núcleo esencial se darán otros estratos de información donde las palabras tengan cada vez menos probabilidad de ser pertinentes. Este es el caso que se pretende ejemplificar para demostrar la asociación que existe entre la infor-

mación (palabras) considerada documentalmente pertinente en la elaboración de un diccionario básico y las distribuciones bibliométricas propias de la tarea bibliotecaria y documental a través de un cuadro invertido de distribución de frecuencias.

Cuadro 4. Dispersión
(Comportamiento de la información pertinente según su frecuencia y su asociación con el criterio de pertinencia "lengua común" en la literatura temática)

Rango	*Frecuencia*	*Vocablos*	*Inf. Pertinente %*
1	360	1	100
2	175	1	100
3-4	93	2	50
5	88	1	100
6	86	1	0
7	84	1	100
8	83	1	100
9	82	1	100
10	80	1	100
11	76	1	100
12-13	74	2	100
14-15	68	2	50
16	67	1	100
17-18	66	2	100
19	65	1	100
20	64	1	100
21	63	1	100
22-23	62	2	50
24	61	1	100
25	57	1	100
26	55	1	100

Cuadro 4. (*Continuación*)

Rango	*Frecuencia*	*Vocablos*	*Inf. Pertinente* %
27-28	54	2	100
29-30	53	2	100
31	52	1	100
32-33	51	2	50
34-35	50	2	0
36	49	1	0
37-41	48	5	60
42	47	1	0
43	46	1	100
44-45	45	2	50
46-47	44	2	50
48-51	43	4	75
52-54	42	3	70
55-58	41	4	75
59	40	1	0
60-64	39	5	60
65-67	38	3	100
68-70	37	3	100
71	36	1	100
72-73	35	2	100
74	34	4	25
78-83	33	6	66
84-87	32	4	25
88-92	31	5	80
93-97	30	5	80
80-101	29	4	50
102-111	28	10	80
112	27	1	0
113-118	26	6	67
119-123	25	5	40

Cuadro 4. (*Conclusión*)

Rango	*Frecuencia*	*Vocablos*	*Inf. Fermíneli* %
124-134	24	11	55
135-136	23	2	0
137-145	22	9	44
146-154	21	9	56
155-162	20	8	25
163-170	19	8	13
171-181	18	11	73
182-193	17	12	50
194-199	16	6	17
200-203	15	4	50
204-214	14	11	36
215-228	13	14	29
229-250	12	22	27
251-273	11	23	26
274-285	10	12	8
286-299	9	14	29
300-334	8	35	11
335-369	7	35	9
370-402	6	33	18
403-443	5	41	7
444-494	4	51	0
495-560	3	66	4
561-631	2	71	0
632-731	1	100	0

En el cuadro invertido de distribución de frecuencias se introduce la noción de rango para clasificar las palabras por frecuencia decreciente, como se podrá observar más adelante.

En el cuadro se puede observar que cuando una palabra tiene

el número de frecuencia más elevado, se le asigna el rango número uno, y que a las palabras de frecuencia decreciente se les va asignando el número de rango posterior inmediato, procediendo de esta manera hasta llegar al último rango, que corresponderá también a la última palabra documentada por el sistema, que tendrá invariablemente la frecuencia uno.

Después de tener esta distribución, se procedió a establecer 6 zonas, tratando de que en cada una de ellas estuviera la misma cantidad de palabras consideradas pertinentes para interpretar el modelo regular de rendimiento decreciente en la información, y así determinar el núcleo de palabras que tienen pertinencia y las otras zonas que contengan la misma cantidad de información:

- La primera zona, en sentido decreciente, correspondió a la información contenida entre el rango 1 y el rango 31 y que a su vez contenía a las frecuencias de 360 a 52.
- La segunda zona correspondió a la información contenida entre el rango 32 y el rango 73, que contenía las frecuencias de 51 a 35.
- La tercera zona correspondió a la información contenida entre el rango 74 y el rango 118, que contenía las frecuencias de 34 a 26.
- La cuarta zona correspondió a la información contenida entre el rango 119 y el rango 180, que contenía a las frecuencias 25 a 18.
- La quinta zona correspondió a la información contenida entre el rango 181 y el rango 270, que contenía a las frecuencias de 18 a 11.
- La sexta zona correspondió a la información contenida entre el rango 271 y el rango 731, que contenía a las frecuencias de 11 a 1.

Resultados de la distribución

La identificación de la información pertinente, se puede asociar con la frecuencia de uso de las palabras, y éstas con las distribuciones bibliométricas en la literatura temática. Sin embargo, existen excepciones en dichas asociaciones que justifican la existencia de

cada método utilizado para identificar y valorar la información bibliográfica. De esto tenemos dos ejemplos, uno de frecuencia elevada y otro de frecuencia baja veámoslos:

- De frecuencia elevada existe la palabra *índice*, que tuvo frecuencia 93 y no perteneció al léxico común, por tener una mala distribución en los textos de literatura temática al igual que otras palabras; en éstas se nota que tienden a pertenecer a un sólo nivel de lengua y muy particularmente a un área temática. *Índice* tuvo 93 ocurrencias, de las cuales 45 se hallan en textos científicos, 45 en otros cinco temas de la lengua culta y sólo una en un tema de la lengua subculta. Según la distribución bibliométrica practicada en este estudio, *índice* tenía un promedio de 89.58% de probabilidades de pertenecer al léxico común y de ser pertinente; sin embargo, no lo fue, aunque por frecuencia de uso hubiera tenido pertinencia si se hablara de un método de selección de palabras basado en las frecuencias elevadas.
- Como ejemplo de baja frecuencia existe la palabra *ayate*, que tuvo tres de frecuencia, pero al estar documentada en los tres niveles de lengua establecidos con función referencial por el DEM, resultó del léxico común y por lo tanto pertinente. Según la distribución bibliométrica *ayate* tenía un promedio de 9.65% de probabilidades de ser pertinente y lo fue. Desde el punto de vista basado estrictamente en frecuencias, no hubiera sido pertinente.

No obstante estas dos excepciones, las distribuciones bibliométricas se asocian muy estrechamente con la información considerada pertinente, bajo el criterio de pertenencia para el léxico común, como también con las frecuencias que este léxico conlleva por su uso, por lo que estos resultados podrían ser un valioso auxiliar económico en futuras búsquedas del léxico común si se consideraran los siguientes porcentajes obtenidos del comportamiento de la infor-

mación recuperada por el Siste de Recuperación de Información del Diccionario del Español de México:

- Cada palabra con frecuencia 52 y más, tiene la probabilidad de ser pertinente en 89.50%, en las búsquedas.
- Cada palabra con frecuencia de 35 a 51, tiene la probabilidad de ser pertinente en 64.28%.
- Cada palabra con frecuencia de 26 a 34, tiene la probabilidad de ser pertinente en un promedio de 62.22%.
- Cada palabra con frecuencia de 18 a 25, tiene la probabilidad de ser pertinente en 43.54%.
- Cada palabra con frecuencia de 11 a 17, tiene la probabilidad de ser pertinmente en 30%.
- Cada palabra con frecuencia de 3 a 10, tiene la probabilidad de ser pertinente en 9.65%.
- Las palabras con frecuencia de 2 o menos no tendrán pertinencia alguna.

[1] Instituto Nacional para la Educación de los Adultos (México). *Memoria 1982-1988*, México, El Instituto, 1988, pp. 17-46.

[2] American Library Association. En el informe presentado por Shores, L. "The measure of reference", pp. 297-302, en *Southeastern Librarian*, 11 (1961).

[3] Lancaster, Frederick Wilfrid, M. J. Joncich. *Evaluación y medición de los servicios bibliotecarios*, México, UNAM, Dirección General de Bibliotecas, 1983, pp. 1-2.

[4] Harmon, G. "Information need transformation during inquiry: an interpretation of user relevance", pp. 41-43, en *Proceedings of American Society of Information Science*, 7 (1970).

[5] Saracevic, Tefko. *Relevancia: una reseña y una estructura para considerar el concepto en ciencia de la información*, México, ABIESI, 1978, 72 h. (Cuadernos de ABIESI; 7).

[6] Barhydt, G. C. "The efectiveness of non user relevance assessments", pp. 146-149, en *Journal of documentation*, vol. 23, núm. 2 (1967).

[7] Saracevic. *Op. cit.*

[8] *Ibid.*

[9] Bradford, S. C. "The documentary chaos", pp. 106-121, en su *Documentation*, London, Lookwood.

[10] Fairthorne, R. A. "Empirical hiperbolic distributions (Bradford-Zipf-Mendelbrot) for bibliometric description and prediction", pp. 321-343, en *Journal of documentation*, vol. 25, núm. 4 (1969).

[11] Kozachov, L. S. "Relevance in informatics and scientology", pp. 3-11, en *Nauchno-Teknicheskaya Informatsiya*, Serie 2, núm. 8 (1969).

[12] Bradford, S. C. "The documentary chaos". *op. cit.*

[13] Saracevic, Tefko. *Introduction to information science*, New York, Bowker, 1970, pp. 110-151.

[14] Saracevic, Tefko. *On the concept of relevance in information science*, Cleveland, Ohio, Case Western Reserve University, 1970.

[15] Zipf, G. *Human behavior and the principle of last effort*, Cambridge, Addison-Wesley, 1949.

[16] Pritchard, A. "Statistical bibliography or bibliometrics?", pp. 348-349, *Journal of documentation*, 25 (1969).

CONCLUSIONES GENERALES

A continuación se presentarán las conclusiones obtenidas como resultado de este estudio, aspectos que se observaron en la información recuperada:

- *La documentación recuperada (palabras)* es utilizada en los tres niveles de lengua organizados en el archivo del sistema de recuperación.
- *La representación gráfica o grafía* de las palabras tiene variantes, las cuales se pueden agrupar y sumar bajo un solo concepto, que es el que les da unidad; éste se le conoce con el nombre de *vocablo* o *palabra*.
- *Juicio o criterio*. La adecuada distribución en la literatura temática de la documentación recuperada por el sistema satisface la necesidad de contar con información testificada en los tres niveles de lengua que se establecieron como requisito, para contar así con información perteneciente al léxico común, es decir información pertinente.
- A partir del establecimiento de un criterio de selección es posible clarificar las relaciones existentes entre los juicios del usuario y las respuestas con *pertinencia*, *relevancia* y *no relevancia*, de aquí que el usuario pueda ponderar todos los niveles o jerarquías de la información a partir de que satisfagan o no sus necesidades.
- Teniendo un juicio de relevancia basado en un criterio o juicio de selección de información, se puede tener una mayor consistencia sobre lo que es pertinente y lo que no lo es.

La relevancia de los documentos puede ser enjuiciada eficazmente por los documentalistas a partir de criterios y principios de selección de información previos a las búsquedas de documentos.

- *La representación*. La recuperación se realiza por medio de la indización automatizada de palabras completas, palabras truncadas y raíces de palabra. Sin embargo, aunque esto ayuda en el tratamiento de la información, los lexicógrafos deben realizar actividades derivadas y complementarias de manera manual para reunir las concordancias correspondientes a las variantes de una palabra o vocablo, problemática que sigue sin ser resuelta por las computadoras, y que al parecer no tendrá solución, aunque se hayan desarrollado avances importantes en la inteligencia artificial. Fue hasta que los lexicógrafos realizaron el ajuste necesario de la documentación recuperada, cuando se pudo encontrar la frecuencia absoluta de una palabra.
- *Las preguntas*. La pregunta formulada en sí misma no le da pertinencia a las respuestas obtenidas, ya que es necesario que éstas además cumplan con los requisitos preestablecidos bajo un criterio de selección por parte del usuario.
- Respecto a las preguntas, el sistema automático responde proporcionando las palabras gráficamente correspondientes a la solicitud formulada, pero se ha notado la existencia de variantes en los registros de una misma palabra, en cuanto a aspectos ortográficos, variantes en las grafías y errores de captura, que afectan indudablemente la recuperación. Hay que considerar que la pregunta de un usuario puede estar elaborada incorrectamente o sin prever estas variantes, lo que puede modificar los resultados en las búsquedas; esto puede suceder, pues hay que tener en cuenta que para una buena descripción de las palabras del español hablado en México resulta en ocasiones importante, documentalmente hablando, respetar este tipo de variantes.
- Se ha podido constatar que la mayor parte de las preguntas planteadas al sistema buscándose la intersección de léxicos individua-

les, fueron contestadas satisfactoriamente, ya que existe un fenómeno lingüístico que lo ha permitido así; es un fenómeno llamado economía lingüística —el intento de conseguir con un mínimo de trabajo lingüístico un máximo de efecto comunicativo—, que es independiente al hecho de que el *Corpus del español mexicano contemporáneo* haya servido como una muestra estadística del uso del español mexicano basado en textos producidos entre 1970 y 1973. Lo anterior permite hacer generalizaciones vigentes sobre las palabras que tienen mayor uso y/o distribución en el español mexicano, aunque no sea así con palabras de baja frecuencia y mala distribución. Esta razón puede generar mayor acuerdo y dar validez al criterio de pertinencia utilizado en este estudio, que sobre el criterio de relevancia, como criterio de selección de palabras, pues el criterio de relevancia está sujeto a modificaciones derivadas de las mismas características de la lengua y su uso, por lo que resulta dependiente y relativo a una gran variedad de factores.

- *Documentos.* La abundancia de documentos recuperados como respuesta no garantiza que sean pertinentes para cubrir las necesidades que tiene un usuario, sin embargo sí existe al presentarse esta abundancia, un porcentaje elevado de probabilidades de que llegue a hacerlo.
- *La información recuperada por cada palabra.* Podría ser interpretada por los desconocedores de la lexicografía como información irrelevante, redundante o como provocadora de ruido, sin embargo, la riqueza de una palabra y su documentación pueden ofrecer un mayor conocimiento de la misma, al hacer posible el identificar su polisemia y el o los niveles de lengua en que se usa realmente.
- *Conocimientos.* La experiencia e ideología de los hablantes se ven reflejadas en las palabras que utilizan, por lo que el identificar palabras de léxico común, facilitará el conocimiento y comunicación de los mexicanos.

APÉNDICE I

Clasificación de la lengua en el Cemc

Género 1 = nivel de lengua culta
Género 2 = nivel de lengua subculta
Género 3 = nivel de lengua no estándar

Subclasificación de información contenida en el Cemc

Nivel	*Género*
Lengua culta	1 Literatura 2 Periodismo 3 Ciencia 4 Técnica 5 Discurso político 6 Religión 7 Habla culta
Lengua subculta	8 Literatura 9 Habla media 10 Lírica popular
Lengua no estándar	11 Textos dialectales 12 Documentos antropológicos 13 Jergas 14 Habla popular

Documentación: "Distribución de frecuencias de los vocablos por clasificación de textos"

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
abanico	4	1	3	1				2	1						12	AB
aborto*	3	1	8	4									2		18	AC
abrevar	5														5	A
abridor				1											1	A
abrojo	2														2	A
abstracto	8	6	16	2		1					1				34	AC
acaecer	2	2						1							5	AB
acalorado	1	1							1						3	AB
acanaladura			3												3	A
acatar	1	2	1		1										5	A
acedo															0	—
acera	12	2	1	3			1								19	A
acoso															0	—
acotar	1	1	3												5	A
acrónimo*															0	—
aculturación				4											4	A
acunar	1														1	A
adecuación				2											2	A
adeudo		3	5												8	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
adicción			1												1	A
adicto	2	2				1		1					1	1	8	ABC
aditivo			6	4											10	A
afear	1	1													2	A
ágata*	2														2	A
agio		2													2	A
agitar	17	2	20	6				4	1		1				51	ABC
agotar	14	12	12	4	2			1	1				2		48	ABC
aguardiente*	7	2						1		17	5				32	ABC
agudeza	3		6	1											10	A
ahínco	2														2	A
ahíto	1														1	A
ahogo	1		1								1				3	AC
ají		1													1	A
alabeado															0	—
alambrar	1														1	A
alcalino			25	1			1								27	A
alcohólico*	2	2	26		1			1							32	AB
alcoholismo*		2	12								1				15	AC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
alergia			1	3											4	A
alfalfa				22							1	1			24	AC
algebraico	1		5												6	A
alheli*															0	—
aliado															0	—
aligerar	1	1		1											3	A
alinear	2	7	5	4				3			1				22	ABC
alineado	1		2	2											5	A
alisar	4		1	1				3			3			1	13	ABC
almíbar	1			9							2				2	AC
aluminio	1		17	11				1	1		2				5	ABC
aluvión*	1															A
amado	10	2	2	4				4		12						AB
amainar	1														1	A
amargo	19	2	5	1		1		9			3	2			42	ABC
aminorar	3	2	1												6	A
amnistía		1													1	A
amo	35	10		5		1		4			5		3	1	64	ABC
ampliación	1	25	12	5	7										50	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
amplificación			7												7	A
amplificar	1	1	5	1				1							9	AB
anca*	2						1								3	A
andado		1													1	A
andanada	1														1	A
andino		1													1	A
anexo	1	4	7	10	1	1		2	1			5			32	ABC
animoso	1	2		1											4	A
anodino	2										2				4	AC
anona*									1	3	3				7	BC
anonadado					1			1	1						3	AB
anquilosis															0	—
anudar	3														3	A
anular	2	2	12	5		1		1	1					2	26	ABC
apelar				1											1	A
apañar													25		25	C
apear										1					1	B
apelar	5	4	9	2	1			1							22	AB
apilar								1				2			3	BC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
apiñar	1														1	A
aportación	11	14	15	6	2										48	A
arboleda	2	1	2						6						11	AB
armamentismo		1													1	A
arqueológico	1	9	7	1			8		1		3				30	ABC
arrear	1										3			1		AC
arrogar			1													A
arruga*	7		3	1							1					AC
artritis*			1								3					AC
as	2	6								5		3				ABC
asado		1		5			2	1			16					ABC
asco	3		1					2		2	1	2			12	ABC
ascua	1	1													2	A
asear	1	1	2	2	1			1			4	1			21	ABC
asedio	2	1	1		1										5	A
asenso	1														1	A
asimétrico	1	2	6	1											10	A
asintótico			2												2	A
asociatividad															0	—

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
asociativo		1	2												3	A
asomo	3		1					2							6	AB
asumir	6	18	17	1	7	2		3	1			1	1		57	ABC
atajo	1	1								1					3	AB
atañer	3	1	2												6	A
atar	14	2	4	4		1		2	2			2			31	ABC
ateo*	2	1	1				1			1					6	AB
aterrar	3												2		5	AC
atinar	1	6	2	1	1		1	1				1			14	ABC
atrofia*			5	1											6	A
atrofiar	2		1												3	A
audacia	4	9	2	1				3							19	AB
audición		8	6	3			2						1		20	AC
auditivo			13	5			2	2							22	AB
auge	6	3	9	1	1		3				1				24	AC
aula		13	2	2	5						1				23	AC
aullar	3									1					4	AB
aura*	1	1								6					8	AB
áureo	1														1	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
aurícula*			3												3	A
auscultación	1		1												2	A
auscultar	2		1	1											4	A
auto*	13	10	9	2				6	3		1		1		45	ABC
autoconsumo															0	—
autóctono	3	1	3	2	2	1	1				1				14	AC
avaricia*	2	1	3			1		1							8	AB
avaro	3														3	A
avieso			1												1	A
avío*		1	3												4	A
avivar	1	2	1	1				1							6	AB
avocar	2	1	2												5	A
axón			2												2	A
ayate				1						1	1				3	ABC
azada			1												1	A
azaroso	1	3					1								5	A
azimut			12												12	A
azimutal															0	—
azolve															0	—

	Lengua estándar															
	Lengua culta [A]							*Lengua subculta* [B]			*Lengua no estándar* [C]					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
azucena*	3									5					8	AB
azúcar*	14	5	8	77			1	7			54	3		6	175	ABC
baba*	2	1						1		1			4		9	ABC
babero*				2							1	3			6	AC
baboso*	1							1					1		3	ABC
balanceado		1	2	1											4	A
balín													2		2	C
balón*		11													11	A
barata		5	4	2			1	2	1	2	6	2		8	33	ABC
barato	6	12	4	1			1		11		20	1	1	4	61	ABC
barca	4							4		2	1		1		12	ABC
barreta				3											3	A
basamento			2												2	A
basuka															0	—
basurero	5		1					2							8	AB
bata*	7		1	3			3	5	2					1	22	ABC
batear		3	1												4	A
beato	2		1										1		4	AC
beca		3	1				5	2	2			1			14	ABC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
becerro*	1		1	8					3	10	16			3	42	ABC
berenjena*															0	—
beriberi															0	—
bibliográfico	1		9	4											14	A
bicoca															0	—
bifocal															0	—
bigote*	11	3						5	1			2			22	ABC
bípedo															0	—
bisección															0	—
bisector				1											1	A
bisectriz			1												1	A
bisexual															0	—
bizco									2	1					3	B
boa*				1										2	3	AC
bobina			2	4					5						11	AB
bobo	2							1							3	AB
bocina	1	1	3					7							12	AB
bocio															0	—
boina												1			1	C

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
bola*	6	6	7	15			6	5	7	4	17	6	2	1	82	ABC
bólido*	1	1													2	A
bolillo*				1				3	2		2	6	3	1	18	ABC
boludo															0	—
bota*	5	3		5				1	2						16	AB
botar*	2						3	1	1	3	18		2	1	31	ABC
bote	7	6		3					2		10	4	4	12	48	ABC
botija															0	—
bozo															0	—
brama	1														1	A
brete							1								1	A
breva	2														2	A
brócoli*				1					2						3	AB
bruja*	4	2	2	1				4	4	20	10	7			54	ABC
bruto	6	8	19	6	1			3			4	1			48	ABC
buey*	7	4	2	2		1		2		8	3	3	3	1	36	ABC
búho*	2		1												3	A
bujía*		1		2							2				5	AC
buque*	1		2	20			2			3	5				33	ABC

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
buzo		3									1		1		5	AC
cabina				2				2							4	AB
caca*									1						1	B
cacao	2	1	1	6	1			2			8				21	ABC
caduco	1	1		3		1	1								7	A
caída	10	7	22	21	3	1		1	1		1				67	ABC
caimán*	1		1							20					22	AB
calambre*			8												8	A
calcio			38	8			2	1							49	AB
calculadora			1												1	A
caldera	1	1	3	5				1								AB
calentamiento			1	17												A
calostro				8												A
callo*				1							1	1				AC
camada*			2							1						AB
can	3	1						3								AB
cana*	4							1			2	2	8		17	ABC
canevá															0	—
canícula	1														1	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
canoa*	1		1	1				3			11			1	18	ABC
canto*	27	12	26	5		2		2	6	4	4				88	ABC
canuto															0	—
caos	5	9	7			1		1					1		24	ABC
caótico	1	1													2	A
capilar		1	6	4											11	A
captura		3	6	30											39	A
cardinalidad															0	—
carie		1	2	1											4	A
carretilla*	1	1		2							2			2	8	AC
cauda	1	2	2												5	A
caudillismo	1		2		1										4	A
caudillo	4	7	11		1		1								24	A
cauto	2			1											3	A
cebo	1		1			1									3	A
cedazo*	2			3							2				7	AC
ceja*	11	1	1	8			1	4		2	1				29	ABC
celulosa	1	2		4											7	A
censura	1	5	2					3							11	AB

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
centeno			2						1						3	AB
cera	5	2	1	1				4		2	7	3		1	26	ABC
cerbatana															0	—
cerradura	2	1		2						1					6	AB
cervical*			7												7	A
cesar	25	8	10	7	2	2		4		1	1	2			62	ABC
cianuro			7	6											13	A
cicatriz*	3	2	11					1				1			18	ABC
ciego*	30	5	9	14		1		13			4	4			80	ABC
cieno			4												4	A
cigueñal*				2					2						4	AB
cilíndrico			5	6											11	A
cincel*	1	1	2	6											10	A
cisma															0	—
cisne*	1	1	3	2											7	A
cláusula*	2	4	5												11	A
clavícula*		1	1												2	A
climaterio															0	—
clítelo															0	—

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
cloroformo			4	1											5	A
coacción		1													1	A
coagulación			7												7	A
coagular	1		1												2	A
coágulo*	2		1												3	A
cocada															0	—
coco*	1	4	1					3		52	12	1			74	ABC
cocoa															0	—
cócono											1				1	C
cohibir	2							1							3	AB
colocación	1	4	10	11							2			2	30	AC
comedido	1	1													2	A
comezón	1			1						1	1	1			5	ABC
cónico	1	3	4												8	A
conmutativo				2											2	A
conserva		1			1						3				5	AC
convocar	3	14	2	1			1								21	A
convocatoria	1	11			1										13	A
copiar	6	1	9	2				3	1		5	1			28	ABC

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
corifeo															0	—
coto	1	1													2	A
cotona				2						1					3	AB
coz															0	—
croar*	1														1	A
cuaco				1						2					3	AB
cuartelazo	1	1	1												3	A
cuate*	12						3	8	4		5	1	36	7	76	ABC
cubicar															0	—
cúbito*															0	—
cueva*	11	4	2				3	2			29	1		3	55	ABC
cuna*	6	3	2	2	1		4	1	3	5	1				28	ABC
cuota	1	22	28	3	3		2		5		8	1		1	74	ABC
cursiva*															0	—
cúspide	5	3	3		1		1								13	A
cutáneo			15	3											18	A
chal	7						2				1	1			11	AC
chauvinismo*		2													2	A
che*		1							1		1				3	ABC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
chequeo		1													1	A
chícharo*				8							2			9	19	AC
chulo	6							7	1	3	4	3			24	ABC
dado	1		2	18											21	A
dañar	4	4	6	24				3	1				1		43	ABC
década*	8	22	23	15											68	A
decaer	3	2	2	1					3		1				12	ABC
deceso		2													2	A
declamar	4	1						1	1		1				8	ABC
declinación			2	2											4	A
declinar*	2	4	3	3											12	A
déficit	1	15	18	5											39	A
degeneración	2		11												13	A
deidad		1	5			1									7	A
delatar		1													1	A
deliberación	1	1	1												3	A
delinear		3	1	2	1	1									8	A
dendrita			1												1	A
denegar								1							1	B

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
denigrar		2				1									3	A
denuedo		1													1	A
denuncia	7	36	2	1		2		1	1						50	AB
denunciar	8	22	6	1	1			2			1	1	1		43	ABC
depauperar															0	—
derribar	10	5	1	2								1			19	AC
desenlace	3	3	2					3							11	AB
desgano								3							3	B
deshidratación*			5	1	1										7	A
desidia		1													1	A
desintegración	3	2	6					1							12	AB
desmedido	3	3	2								1				9	AC
desmenuzar	1		1	10		1					4				17	AC
desnudo*	36	3	3	3	1			14	1		4			1	66	ABC
desprendimiento	1		6	1											8	A
desunión	1	1										1			3	AC
desuso		1		2							1				4	AC
deudo	5		10					1							16	AB
devorar	17	5	2	1		1		4		1	2				33	ABC

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
diacrítico															0	—
diana	9	3		1						2					15	AB
dicción	2		7	1											10	A
didáctico*		2	13		1										16	A
diésel*		3	1	4											8	A
dimanar					1										1	A
dimisión		1													1	A
diptongo															0	—
dislocación			1												1	A
divisibilidad															0	—
divisible															0	—
dominante	3	6	18	5	2		1	2			1				38	ABC
dona*				2						2	2				6	ABC
donar	2	4	1	3	2						4	1			17	AC
duende*	1										11				12	AC
dúo	2	2	1												5	A
duque	13	2		1				4							20	AB
ébano*	1														1	A
ego	3		2												5	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
emanación	1		1												2	A
émbolo			2	10											12	A
eminencia	2		1					1							4	AB
enano*	8							1		6	1	1	1		18	ABC
encanto	12	4	5	1			1	10		3		1			37	ABC
encefalitis															0	—
encía*			3					1							4	AB
encono	2														2	A
enchiquerar															0	—
engorda				6							2				8	AC
enhiesto	2	1													3	A
enjambre*			1												1	A
enmienda	2														2	A
enmudecer	7			1											8	A
ente		2	17					1					1		21	ABC
entuerto															0	—
enunciación			1	1											2	A
envés			2	2											4	A
epígrafe*	1		1												2	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
equidistante	2			1											3	A
equívoco	6	3	1	1											11	A
erario*		1	2		1										4	A
erizo*	1		1					1					4		7	ABC
errata*	2														2	A
esbozo	1	1	1	3											6	A
escisión		2	2									1			5	AC
esclavismo		1													1	A
escroto			4												4	A
esguince*	1														1	A
especulación	2	13	1	5	1	1									23	A
espejismo*	1	2	3					1							7	AB
espinaca*				7					4						11	AB
establo	1	1	1	5							4			2	14	AC
estabular															0	—
estemón*			5	1				1							7	AB
estribo*	3	2	1	22											28	A
ética	9	7	14	7			1	2	3						43	AB
eufónico															0	—

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
eunuco	1	4		1											6	A
evocar	7	3	4			1		5							20	AB
exiguo	2	2	2												6	A
expiración				1											1	A
extensionista															0	—
extractivo															0	—
eyaculación			1												1	A
factura	1	1	8	1				2							13	AB
faena	4	39	2	3					1		4				53	ABC
fémur*		2	2												4	A
fértil	2	1	4	3				1							11	AB
fertilizar				4											4	A
fideicomiso*		9	2	1											12	A
fierro	5	2	5	19				3	4		17	3	3	2	63	ABC
flema								1							1	B
fofo															0	—
folio			4					1							5	AB
fractura*	2	2	14	4	1						1				24	AC
gaita*			1												1	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
gancho	3	6	3	19		1					3	2	3	4	44	AC
garrote								1	1		4		4		10	BC
garrucha	1														1	A
gástrico*	1		12	1											14	A
gatear	2														2	A
gestación*	1	1	11	6									1		20	AC
glaciación			3												3	A
gorra	5							2			1			3	11	ABC
gorro*			1	1			1	5			2		1		11	ABC
gota*	17	2	28	7		1	2	6		1	1	1		2	68	ABC
gotear	2		1					1							4	AB
gravar			1	1											2	A
graznar*	1							1							2	AB
gregario	1			1											2	A
grenetina				16											16	A
grotesco	7	4	1												12	A
guarnición	1	2		2							1				6	AC
guerrear		1													1	A
guetto		1													1	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
haba*				4				4	1		6			2	17	ABC
habano	1														1	A
hacendoso	1			1											2	A
hacinado		1													1	A
hado															0	—
halo	3			1											4	A
hato*				7											7	A
haz	9		7	4			4	3	6	3	1		1	1	39	ABC
hebra	1	3	11								2				17	AC
hediondo											1	1			2	C
hegemonía		6	2		1										9	A
hendido															0	—
heno*	1		3	4							3				11	AC
hidratación			1	1											2	A
hiel	2						1	1			1				5	ABC
higo*										2					2	B
hilar*	3	5	1	2							2				13	AC
hipérbaton*															0	—
hipo			2					1							3	AB

	Lengua estándar															
	Lengua culta [A]							*Lengua subculta [B]*			*Lengua no estándar [C]*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
hipocorístico															0	—
hipófisis			10												10	A
hispano		6						1							7	AB
hito				1											1	A
hojear	3	1						2	1			4			11	ABC
hola*	1		1					12			1				15	ABC
homicida	2	7	1												10	A
homicidio	1	5	4				1	2							13	AB
homología															0	—
homólogo	1		3												4	A
homotecia															0	—
horario	1	6	10	2	1		2		2		4			2	30	ABC
horror	18	3	3				2	5	2			3	1	1	38	ABC
hosco	3	1						1							5	AB
hoz*															0	—
huida	7		2	1				2			2				14	ABC
hulla		1	2												3	A
humanismo*	2	1			1										4	A
húmero*															0	—

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
hundir	31	8	12	2			2	8			3				66	ABC
hurto	1	5							1						7	AB
icono															0	—
ida	4	3	1				1	1			5	1		1	17	ABC
ideólogo		2	1		1										4	A
ido	2	2					1		2		1	2			10	ABC
idóneo		5	2	1	1										9	A
ilegítimo			1		1										2	A
imparcial	1	1	3		1						1				7	AC
improductivo		2	1	3	1										7	A
impugnar	1		5												6	A
inanición															0	—
inanimado	1		1												2	A
inca		1	1								1				3	AC
incauto	1							1							2	AB
incidir	1	5	11		2										19	A
incisión	1		6	9											16	A
inciso		2	8	9											19	A
incógnita	3	2	6					2	1						14	AB

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Voces	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
inconsistente			1		1										2	A
incubar*	1	1	2	5											9	A
indecisión	1		1		1										3	A
indeciso	8	1	2	2				2		1	1	2	1		20	ABC
índice	7	15	45	14	5		6	1							93	AB
indigente		1													1	A
indulto*	2	1	1												4	A
inevitable	14	6	16	7		2					1		1		47	AC
inflamable	1		1	2											4	A
infraestructura	1	19	8	8	3						1				40	AC
inhumano	3					2									5	A
inicio	1	12	5	5	2								1		26	AC
inicuo		2													2	A
inmediación	2	1	2	1							1		1		8	AC
inminencia	1														1	A
inmundicia	4	1	1												6	A
inmune	1		3												4	A
innato			1	2	1		2	2							8	AB
innocuo			2												2	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
innovar			2												2	A
inquina															0	—
inquisición*															0	—
insano	2														2	A
insidia															0	—
institucional	1	3	8	1	8	1									22	A
internacionalismo															0	—
intriga	2		1	1											4	A
invocar	8	1	6		2	1					2				20	AC
irracional	4	3	9	2	1	1									20	A
irrigar	1		2												3	A
irritar	4	2	1	1				1			1		1	1	12	ABC
itacate											1				1	C
jarabe*	1	1		1				2	2	16				3	26	ABC
jauría	3	1						1							5	AB
jeringa	1		2	7					1			1		1	13	ABC
jeta															0	—
kilométrico															0	—
kiosco*	1							5			3				9	ABC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
lagartija*	1								2		1			2	6	ABC
lamer	5	1		1							1				8	AC
lapa					1										1	A
laringe*			5	1											6	A
lecho	15	2	3					8	2	1					31	AB
lelo	1												11		12	AC
lía				1											1	A
liberación	8	10	8	5	1	10	2	1							45	AB
libertador						1		1							2	AB
librecambio															0	—
liderazgo*		1													1	A
ligamento*			3	1											4	A
lima*	2			9						14	6		7		38	ABC
lío	3		3					8	1	2	3	6	1	1	28	ABC
lítico			1												1	A
loa		1													1	A
localidad		7	9	13	2			2							33	AB
lubricación				4											4	A
lubricante			1	4											5	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
lucro		3	2		2	2					1		1		11	AC
macizo	6		2	8				2		1	9		2		30	ABC
mafia		1					1								2	A
malear	1	1									2				4	AC
mamar	1		1			1		1		2	2	2	6	2	18	ABC
mana								1				6			7	BC
mandado	4							3	1	1	7	3		3	22	ABC
manía	4			1				2			1				8	ABC
manido															0	—
manso	7	10	2					1		4	4	1			29	ABC
manuscrito	5	6	11	3		1									26	A
marginal	2		22	1											25	A
marrano								2			3	1			6	BC
materia	32	96	114	48	14	4	12	6	3		29	1	1		360	ABC
matriarcado															0	—
mazo'	2		2												4	A
mecenas		2													2	A
médano			2												2	A
mediano	4	28	20	24			1				6	1		2	86	AC

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
meneo			1												1	A
menopausia			1				1								2	A
menú	2	2		3				1							8	AB
mesón	3			3					1			2			9	ABC
metonimia															0	—
miedoso	3							1							4	AB
mies															0	—
mímica	1		1					1							3	AB
mimo		1													1	A
mira	2	8	3		2		1	1			2			2	21	ABC
miscelánea*	1	1	4	1				1							8	AB
moco*			3									2	1	1	7	AC
módico															0	—
modoso															0	—
mofar	1														1	A
mohín								1							1	B
mohíno										1					1	B
mohoso	1														1	A
momia								2							2	B

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
monada	1							2							3	AB
mongol			2												2	A
monta		2		2				1							5	AB
mora	2		2		1								4		9	AC
mucama		1													1	A
muda			1								1	1			3	AC
mudo	5		1												6	A
mueca	4							3							7	AB
mugir*				1											1	A
munición*				1										1	2	AC
musa	4	4	1					3							12	AB
muslo*	6	2	7	1							1				17	AC
mutación	2		4												6	A
nabo*				8					1		2				11	ABC
nacido	5	3	13	6					1		2			1	31	ABC
nado							1		1		1				3	ABC
nana	11		2					2	6	5	7			1	34	ABC
nato	2	2					1				1				6	AC
náutico		1									3				4	AC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
nave	7	10	1	11			3	2			2		3		39	ABC
navío*	3	2	1					1		1					8	AB
neolítico			1				1								2	A
neutro			21	9			2								32	A
newcastle															0	—
nicotina															0	—
nidada															0	—
nieto*	8	11	1	2			14	10	5	5	14	3		10	83	ABC
nimio	1	1						1							3	AB
niñez	2	15	3	3	1		1				2	1			28	AC
níveo															0	—
nocivo	1	7	6	6				1							21	AB
nómina*	1	2	3					4							10	AB
nonato															0	—
nuca*	8						1	2							11	AB
nuez*	2	1	1	16	1			3			2				26	ABC
ñapa															0	—
ñato															0	—
ñoño															0	—

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
obeso				1					1						2	AB
oca		1													1	A
ocio	3	1	2	1				1							8	AB
ocioso	1	2	4								1				8	AC
oda*	1		3												4	A
odio	28	2	12	1				8			1		1		53	ABC
odioso	4							3	1						8	AB
odisea		1													1	A
ojear	1			1										1	3	AC
ojiva															0	—
ola*	16	11	4	2				5		15	12				65	ABC
óleo*	3	7	1				2				1				14	AC
olfatorio			17												17	A
oligarquía		11	2		7			1							21	AB
olla*	15	3	4	14				2		1	40	5			84	ABC
ondear	4				1										5	A
onza*	1			3		3		16			1				24	ABC
ópalo															0	—
orca															0	—

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
oriental	4	6	11	2						2					25	AB
orina	1		24	1										3	29	AC
osadía	2		1					2		2					7	AB
osado	1			1				1							3	AB
osar	4							1							5	AB
osario															0	—
oso*	4		2	1		1	1			3	1				13	ABC
otear	3				1										4	A
oveja*	6	2	6	2						1			1		18	ABC
ovino*			1	3	1										5	A
pagaré*			1												1	A
palapa				1											1	A
paleolítico							3								3	A
pálido*	18	3	6					12				1	1		41	ABC
paño*	3		1	12						3	1	1			21	ABC
pápalo*															0	—
papeleo		1						1			7				9	ABC
paperas*								1							1	B
paráfrasis	2														2	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
paridera															0	—
párpado*	7		8	4				4						1	24	ABC
parvada*			1	1											2	A
patata															0	—
pato*	2	1		6				2			6				17	ABC
patriarcado															0	—
peroné*															0	—
piara*															0	—
píldora*	1	1	1				1	1					1		6	ABC
piloncillo				2				3			12				17	ABC
pío	3														3	A
pipa*	1	1						6			2			2	12	ABC
pita										1				1	2	BC
pito	2						1	1			7		1		12	ABC
planificación	1	2	4	1											8	A
plaqueta															0	—
pleonasmo															0	—
polifacético										1					1	B
popa*	2		1	3											6	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
porosidad															0	—
postal*	4	3	2	29	4										42	A
pote															0	—
proclítico															0	—
proletario	1	2	1	2							1		1		8	AC
protectorado															0	—
púa	1														1	A
pubertad*							1								1	A
puerperio			3												3	A
puntiagudo															0	—
puño*	12	8	2	2				6			11				41	ABC
pupilo	10	3	4					7							24	AB
quebradizo	1			2											3	A
queda	7		1					6		2	2	1			19	ABC
quema*	1	4													5	A
quicio	9	1				1					1	1			13	AC
quieto	18	1	2					5				3		1	30	ABC
quimera		1	4												5	A
quinina			2	1									3		6	AC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
quinquenio			2	1											3	A
quinta	6	18	7	5					2	1	3	1	1		44	ABC
quiosco	2	1										1			4	AC
rastreo			1	4											5	A
reata	1			7						11	2	3		1	25	ABC
rebaño	6	4	1	1				1							13	AB
recesivo			1												1	A
regata		4						1			2				7	ABC
regatear	3	2						1							6	AB
relinchar*	1														1	A
remisión			5												5	A
renovable		4		1											5	A
residual			9	3											12	A
reuma							1	1			4				6	ABC
rito	7	2	8	2		8	1				1				29	AC
roto	18	1	6	3			3	2	1		2	3		2	41	ABC
rótula															0	—
rúbrica			1												1	A
ruego	1		1			1	1	2				1			7	ABC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
rugoso*	3			2											5	A
sabroso	5	4		3				13	5	5	50	1	5	2	93	ABC
saciar	1							3							4	AB
salado	2	1	6	3			1	2			8	1			24	ABC
salvajismo													1		1	C
saqueo	2								4						6	AB
sazonar	2			28					2		5				37	ABC
sebo	3	1							1		1				6	ABC
secuaz															0	—
secular	2	2	7		2	1		1			4		1		20	ABC
sedición															0	—
sedoso	1		1	1				1							4	AB
senado		17	1	1	3										22	A
sensitivo	2		2	1			1	2							8	AB
seroso			5	1											6	A
serranía	1	1	5							6	1				14	ABC
servidumbre	5	1	1	1				1	2		4				15	ABC
seso	2	1		5				1			2				11	ABC
seudónimo	3	1	1	1				1							7	AB

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
sien	7														7	A
sigla		2	1		1										4	A
símil				1				1	1						3	AB
similitud	1	4	4	2			1				1		1		14	AC
sinécdoque															0	—
sinóptico		2	1												3	A
sinovial															0	—
sinuoso	1		2					2							5	AB
sisa*				8											8	A
sistemático	7	5	16	3	1	1		1							34	AB
sobrenatural	3	5	2			3	1	2			2				18	ABC
socio	1	12	15	7	1			1	2		2	4		1	46	ABC
soez	3														3	A
soso		1	8	1							1				11	AC
soviético	2	33	5					1							41	AB
subalterno	3	2	1												6	A
subarrendatario															0	—
subsidiar		1	1	1											3	A
suburbano				1											1	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
suburbio	2														2	A
subversión	1	2													3	A
subvertir	1		1												2	A
sucedáneo															0	—
sucumbir	4	2	1	2											9	A
suicidio	11	3	1				1	1							17	AB
supeditar		1	3	2	1		1								8	A
supervivencia	1	3	8	2	1	1							3		19	AC
tácito	2	3	2						1						8	AB
tacón	2		2	2				3	2	27	5	1	1	9	54	ABC
tajo	5	1					1								7	A
tallar	2	2	8	3				2			11				28	ABC
talle	4			2				3		1					10	AB
tañer	2		6					1							9	AB
tapete	2	12	2	6				1	2			3			28	ABC
tapia	1														1	A
tarro	1														1	A
tea	1														1	A
teja*				3				3			3	4			13	ABC

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
telecomunicación*		2		1											3	A
tenate											1				1	C
tenaza*	1			2					4			2		2	11	ABC
tendón*		1	3												4	A
teniente*	2	8	5				2	2		3	2		2	2	28	ABC
tenue	6	1	3	1				2							13	AB
teta				1							2				3	AC
tetilla*			2												2	A
tibia	11	1	6	7				8		2	3	1			39	ABC
tiento*	1														1	A
tina				3					1		2	2			8	ABC
tinaco	1											1			2	AC
tino	1	2	1	1	1			1							7	AB
tirria															0	—
titán				1											1	A
toga															0	—
topar	5		2					2	3	2		2		1	17	ABC
topo*	1	5	1	12	3				1			1	1		25	ABC
toque*	5	8	2	7			2	2					17		43	ABC

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
torcedura	1			1							1				3	AC
tornillo	1		4	43											48	A
torta*	7		3					3	1	2	13	1	6	1	37	ABC
tragar	12	3	1					4				3		1	24	ABC
trago	11	1		1				8		6	5	3			35	ABC
transpirar															0	—
trasplantar			1								4			2	7	AC
trasplante		1	5	5										1	12	AC
trapezoide															0	—
trastorno	4	3	36	4	1			1			2		1		52	ABC
trebejo			1												1	A
treta	1			1											2	A
tribal	1														1	A
trilogía		1													1	A
triturar			1	3				1							5	AB
troje								1							1	B
trompa	4						1	1				1			7	ABC
trotar	1														1	A
trote	6	1													7	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta (B)*			*Lengua no estándar (C)*					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
trova										1			2		3	BC
tuteo	1														1	A
tutor															0	—
ubre				7											7	A
úlcera			12												12	A
unción					1	1									2	A
unísono			1		1										2	A
unívoco	4		4												8	A
uretra			5												5	A
usado		2	12	16				1	2		1			1	35	ABC
vacuo	1														1	A
vacuola			2												2	A
vaho	4		1												5	A
vale															0	—
válvula	1	2	13	43			1				2				62	AC
vano	17	2	3			2		5		4					33	AB
vejar		2													2	A
velarización															0	—
vendaje			3	1											4	A

	Lengua estándar															
	Lengua culta (A)							Lengua subculta (B)			Lengua no estándar (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
venéreo			2									1			3	AC
ventosa															0	—
ventrículo			8												8	A
venus			1	1											2	A
vera	2														2	A
veraz	2	2	1	2	1	1									9	A
veredicto	2	1			1										4	A
verruga*	2		1	1				2							6	AB
vertebral		3	2	1											6	A
verter	10	3	10	28											51	A
vértigo	4	1	2	2				1							10	AB
vesícula	1		3								1				5	AC
vestigio	1	3	2	2	1										9	A
veto		1													1	A
vicioso	4	1	2	1		1			1		1	1	6		18	ABC
vigía	3														3	A
vinagre				23				4			4				31	ABC
vira															0	—
visa		1													1	A

	Lengua estándar															
	Lengua culta (A)							*Lengua subculta* (B)			*Lengua no estándar* (C)					
Vocablos	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*Frecuencias*	*Géneros*
visceral*			2												2	A
víveres		4									1	1		1	7	AC
vocear															0	—
voraz		1				1									2	A
vulva			4												4	A
yate*	1	1						2							4	AB
yegua	7		1	6			5	1		4	4				28	ABC
yugo	1	1						1	1			1			5	ABC
yute*				2											2	A
zahúrda															0	—
zahino															0	—
zootecnia															0	—
zueco															0	—
zumo	1			2				4							7	AB

* Estos 175 vocablos, también fueron sugeridos por el equipo lexicográfico del DEM, en una reunión de trabajo a fines de 1988 para eliminar lagunas del diccionario básico, respecto al léxico de temas (p.e., deportes, cocina, filosofía, etc.). A éstas, se les puede adjudicar relevancia desde el punto de vista de la disponibilidad léxica.

APÉNDICE II

Atribución de peso a la información recuperada a partir de su pertinencia y relevancia: orden descendente en importancia

Es fundamental hacer notar que la información obtenida del sistema de recuperación se refiere a un conjunto cerrado de textos (*corpus*), el cual tiene bases de tipo estadístico y sirve como muestra del español hablado en México; sin embargo, al considerar que la lengua es una estructura abierta y en constante movimiento, puede pensarse que estos resultados están sujetos a limitaciones y es probable que si la muestra hubiera sido más grande, algunas palabras pasarían de un nivel a otro en cuanto a su importancia documental.

En el primer nivel de importancia se tiene a las palabras que satisfacen las necesidades de información para seleccionar entradas en un diccionario al estar contenidas en la intersección de los conjuntos A, B y C, éstas se caracterizan por pertenecer a los tres niveles de lengua hablada por los mexicanos y se les puede denominar del *léxico común*. A continuación se presentan las 166 palabras pertinentes por orden alfabético y con sus frecuencias de aparición:

A: adicto 8; agitar 51; agotar 48; aguardiente 32; alinear 22; alisar 13; aluminio 33; amargo 42; amo 64; anexo 32; anular 26; arqueológico 30; as 16; asado 28; asco 12; asear 21; asumir 57; atar 31; atinar 14; asunto 45; ayate 3; azúcar 175.

B: baba 9; baboso 3; barata 33; barato 61; barca 12; bata 22; beca 14; becerro 42; bigote 22; bola 82; bolillo 18; botar 31; bote , bruja 54; bruto 48; buey 36; buque 33.

C: cacao 21; caída 67; cana 17; canoa 18; canto 88; caos 24; ceja 29; cera 26; cesar 62; cicatriz 18; ciego 80; coco 74; comezón 5; copiar 28; cuate 76; cueva 55; cuna 28; cuota 74.

CH: che 3; chulo 24.

D: dañar 43; decaer 12; declamar 8; denunciar 43; desnudo 66; devorar 33; dominante 38; dona 6.

E: enano 18; encanto 37; ente 21; erizo 7.

F: faena 53; fierro 63.

G: gorra 11; gorro 11; gota 68;

H: haba 17; haz 39; hiel 5; hojear 11; hola 15; horario 30; horror 38; huida 14; hundir 66.

I: ida 17; ido 10; indeciso 20; irradiar 12.

J: jarabe 26; jeringa 13.

K: kiosco 9.

L: lagartija 6; lima 38; lío 28.

M: macizo 30; mamar 18; mandado 22; manía 8; manso 29; materia 360; mesón 9; mira 21.

N: nabo 11; nacido 31; nado 3; nana 34; nave 39; nieto 83; nuez 26.

O: odio 53; ola 65; olla 84; onza 24; oso 13; oveja 18.

P: pálido 41; paño 21; papeleo 9; párpado 24; pato 17; píldora 6; piloncillo 17; pipa 12; pito 12; puño 41.

Q: queda 19; quieto 30; quinta 44.

R: reata 25; regata 7; reuma 6; roto 41; ruego 7.

S: sabroso 93; salado 24; sazonar 37; sebo 6; secular 20; serranía 14; servidumbre 15; seso 11; sobrenatural 18; socio 46.

T: tacón 54; tallar 28; tapete 28; teja 13; tenaza 11; teniente 28; tibia 39; tina 8; topar 17; tope 25; toque 43; torta 37; tragar 24; trago 35; trastorno 52; trompa 7.

U: usado 35.

V: vicioso 18; vinagre 31.

Y: yegua 28; yugo 5.

A partir del segundo nivel, al existir información que funja como respuesta a las preguntas realizadas al sistema, se estará presentando un contacto efectivo en la comunicación y por lo tanto se podrá considerar cumplida la noción de relevancia.

En el segundo nivel se presentarán las palabras recuperadas bajo la intersección de dos tipos de lengua, la lengua culta y la subcultа, que corresponden a los subconjuntos A y B:

A: abanico 12; acaecer 5; acalorado 3; alcohólico 32; amado 34; amplificar 19; anonadado 3; apelar 22; arboleda 11; asomo 6; atai- 3; ateo 6; audacia 19; auditivo 22; aullar 4; aura 8; 14; avaricia 8; avivar 6; azucena 8.

B: basurero 8; bobina 11; bobo 3; bocina 12; bota 16; brócoli.

C: cabina 4; caimán 22; calcio 49; caldera 11; camada 3; can 7; censura 11; centeno 3; cerradura 6; cigüeñal 4; cohibir 3; cotona 3; cuaco 3.

D: denuncia 50; descenlace 11; desintegración 12; deudo 16; diana 15; duque 30.

E: eminencia 4; encía 4; espejismo 7; espinaca 11; esternón 7; ética 43; evocar 20.

F: factura 13; fértil 11; folio 5.

G: gotear 4; graznar 2.

H: hipo 3; hispano 7; homicidio 13; hosco 5; hurto 7.

I: incauto 2; incógnita 14; índice 93; innato 8.

J: jauría 5.

L: lecho 31; liberación 45; libertador 2; localidad 33.

M: menú 8; miedoso 4; mímica 3; miscelánea 8; monada 3; monta 5; mueca 7; musa 12.

N: navío 8; nimio 3; nocivo 21; nómina 10; nuca 11.

O: obeso 2; ocio 8; odioso 8; oligarquía 21; oriental 25; osadía 7; osado 3; osar 5.

P: pupilo 24.

R: rebaño 13; regatear 6.

S: saciar 4; saqueo 6; sedoso 4; sensitivo 8; seudónimo 7; símil 3; sinuoso 5; sistemático 34; soviético 41; suicidio 17.

T: tácito 8; talle 10; tañer 9; tenue 13; tino 7; triturar 5.

V: vano 33; verruga 6; vértigo 10.

Y: yate 4.

Z: zumo 7.

En el tercer nivel de importancia se encuentran las palabras que corresponden a la intersección de los subconjuntos A y C, pertenecientes a la lengua culta y a la lengua no estándar respectivamente:

A: aborto 18; abstracto 34; ahogo 3; alcoholismo 15; alfalfa 24; almíbar 12; anodino 4; arrear 5; arruga 12; artritis 4; aterrar 5; audición 20; auge 24; aula 23; autóctono 14.

B: babero 6; beato 4; boa 3; bujía 5; buzo 5.

C: callo 3; carretilla 8; cedazo 7; colocación 30; conserva 5.

CH: chal 11; chícharo 19.

D: derribar 19; desmedida 9; desmenuzar 17; desunión 3; desuso 4; donar 17; duende 12.

E: engorda 8; escisión 5; establo 14.

F: fractura 24.

G: gancho 44; gestión 20; guarnición 6.

H: hebra 17; heno 11; hilar 13.

I: imparcial 7; inca 3; inevitable 47; infraestructura 40; inicio 26; inmediación 8; invocar 20.

L: lamer 8; lelo 12; lucro 11.

M: malear 4; mediano 86; moco 7; mora 9; muda 3; munición 2; muslo 17; náutico 4; niñez 28.

O: ocioso 8; ojear 3; óleo 14; orina 29.

P: proletario 8.

Q: quicio 13; quinina 6; quiosco 4.

R: rito 29.

S: similitud 14; soso 11; supervivencia 19.

T: teta 3; tinaco 2; torcedura 3; trasplantar 7; trasplante 12.

V: válvula 62; venéreo 3; vesícula 5; víveres 7.

En el cuarto nivel de importancia se encuentran las palabras pertenecientes a los textos de lengua culta o subconjunto A, que son las fuentes a las que mayor peso se les da generalmente en la elaboración de un diccionario, aun no teniendo intersección con otro subconjunto:

A: abrevar 5; abridor 1; abrojo 2; acanaladura 3; acatar 5; acera 19; acotar 5; aculturación 1; acunar 1; adecuación 2; adeudo 8; adicción 1; aditivo 10; afear 2; ágata 2; agio 2; agudeza 10; ahínco 2; ahíto 1; ají 1; alambrar 1; alcalino 27; alergia 4; algebraico 6; aligerar 3; alineado 5; aluvión 1; amainar 1; aminorar 6; amnistía 1; ampliación 50; amplificación 7; anca 3; andado 1; andanada 1; andino 1; animoso 4; anudar 3; apelar 1; apiñar 1; aportación 48; armamentismo 1; arrogar 1; ascua 2; asedio 5; asenso 1; asimétrico 10; asintótico 2; asociativo 3; atañer 6; atrofia 6; atrofiar 3; áureo 1; aurícula 3; auscultación 2; auscultar 4; avaro 3; avieso 1; avío 4; avocar 5; axón 2; azada 1; azaroso 5; azimut 12.

B: balanceado 4; balón 11; barreta 3; basamento 2; batear 4; bibliógrafo 14; bisector 1; bisectriz 1; bólido 2; brama 1; brete 1; breva 2; búho 3.

C: caduco 7; calambre 8; calculadora 1; calentamiento 18; calostro 8; canícula 1; caótico 2; capilar 11; captura 39; carie 4; cauda 5; caudillismo 4; caudillo 24; cauto 3; cebo 3; celulosa 7; cervical 7; cianuro 13; cieno 4; cilíndrico 11; cincel 10; cisne 7; cláusula 11; clavícula 2; cloroformo 5; coacción 1; coagulación 7; coagular 2; coágulo 3; comedido 2; cónico 8; conmutativo 2;

convocar 21; convocatoria 13; coto 2; croar 1; cuartelazo 3; cúspide 13; cutáneo 18.

CH: chauvinismo 2; chequeo 1.

D: dado 21; década 68; deceso 2; declinación 4; declinar 12; déficit 39; degeneración 13; deidad 7; delator 1; deliberación 3; delinear 8; dendrita 1; denigrar 3; denuedo 1; deshidratación 7; desidia 1; desprendimiento 8; dicción 10; didáctico 16; diésel 8; dimanar 1; dimisión 1; dislocación 1; dúo 5.

E: ébano 1; ego 5; emanación 2; émbolo 12; encono 2; enhiesto 3; enjambre 1; enmienda 2; enmudecer 8; enunciación 2; envés 4; epígrafe 2; equidistante 3; equívoco 11; erario 4; errata 2; esbozo 6; esclavismo 1; escroto 4; esguince 1; especulación 23; estribo 28; eunuco 6; exiguo 6; expiración 1; eyaculación 1.

F: fémur 4; fertilizar 4; fideicomiso 12.

G: gaita 1; garrucha 1; gástrico 14; gatear 2; glaciación 3; gravar 2; gregario 2; grenetina 16; grotesco 12; guerrear 1; guetto 1.

H: habano 1; hacendoso 2; hacinado 1; halo 4; hato 7; hegemonía 9; hidratación 2; hipófisis 10; hito 1; homicida 10; homólogo 4; hulla 3; humanismo 4.

I: ideólogo 4; idóneo 9; ilegítimo 2; improductivo 7; impugnar 6; inanimado 2; incidir 19; incisión 16; inciso 19; inconsistente 2; incubar 9; indecisión 3; indigente 1; indulto 4; inflamable 4; inhumano 5; inicuo 2; inminencia 1; inmundicia 6; inmune 4; inicuo 2; innovar 2; insano 2; institucional 22; intriga 4; irracional 20; irrigar 3.

L: lapa 1; laringe 6; lía 1; liderazgo 1; ligamento 4; lítico 1; loa 1; lubricación 4; lubricante 5.

M: mafia 2; manuscrito 26; marginal 25; mazo 4; mecenas 2; médano 2; meneo 1; menopausia 2; mimo 1; mofar 1; mohoso 1; mongol 2; mucama 1; mudo 6; mugir 1; mutación 6.

N: neolítico 2; neutro 32.

O: oca 1; oda 4; odisea 1; olfatorio 17; ondear 5; otear 4; ovino 5.

P: pagaré 1; palapa 1; paleolítico 3; paráfrasis 2; parvada 2; pío 3; planificación 8; papa 6; postal 42; púa 1; pubertad 1; puerperio 3.

Q: quebradizo 3; quema 5; quimera 5; quinquenio 3.

R: rastreo 5; recesivo 1; relinchar 1; remisión 5; renovable 5; residual 12; rúbrica 1; rugoso 5.

S: senado 22; seroso 6; sien 7; sigla 4; sinóptico 3; sisa 8; soez 3; subalterno 6; subsidiar 3; suburbano 1; suburbio 2; subversión 3; subvertir 2; sucumbir 9; supeditar 8.

T: tajo 7; tapia 1; tarro 1; tea 1; telecomunicación 3; tendón 4; tetilla 2; tiento 1; titán 1; tornillo 48; trebejo 1; treta 2; tribal 1; trilogía 1; trotar 1; trote 7; tuteo 1.

U: ubre 7; úlcera 12; unción 2; unísono 2; unívoco 8; uretra 5.

V: vacuo 1; vacuola 2; vaho 5; vejar 2; vendaje 4; ventrículo 8; venus 2; vera 2; veraz 9; veredicto 4; vertebral 6; verter 51; vestigio 9; veto 1; vigía 3; visa 1; visceral 2; voraz 2; vulva 4.

Y: yute 2.

En el quinto nivel de importancia se encuentran las palabras de la intersección de los subconjuntos B y C, que corresponden a la lengua subculta y a la no estándar respectivamente. Con este nivel termina la información considerada relevante:

A: anona 7; apilar 3.

G: garrote 10.

M: mana 7; marrano 6.

P: pita 2.

T: trova 3.

En el sexto nivel de importancia se encuentran las palabras del subconjunto B, que corresponden a la lengua sub culta, la cual no tuvo intersección con otros subconjuntos por lo que no se le consideró relevante; generalmente lleva las etiquetas de *coloquial, familiar*, etc., y a continuación se elistan:

A: apear 1.

B: bizco 3.

C: caca 1.

D: denegar 1; desgano 3.

F: flema 1.

H: higo.

M: mohín 1; mohíno 1; momia 2.

P: paperas 1; polifacético 1.

T: troje 1.

En el séptimo nivel de importancia con el cual terminan las respuestas emitidas por el sistema y con esto la relevancia, se encuentran las palabras del subconjunto C, que pertenecen a la lengua no estándar y que las usan los mexicanos en grupos cerrados y en algunos casos con carácter de lenguajes secretos por lo que tampoco se consideraron relevantes. Aquí se encuentran los dialectalismos y las jergas, principalmente:

A: apañar 25.

B: balín 2; boina 1.

C: cócono 1.

H: hediondo 2.

I: itacate 1.

S: salvajismo 1.

T: tenate 1.

En el octavo nivel de importancia se encuentran las palabras que no fueron recuperadas por el sistema de recuperación, pero que al cotejarse su existencia en diccionarios mexicanos, hispanoamericanos y de la lengua española general utilizados en el DEM, se encontró en su documentación que se usan en el español general y en México, por lo que se pueden considerar palabras relevantes —bajo criterio documental— no recuperadas:

A: acedo; acoso; acrónimo; alabeado; alhelí; aliado; anquilosis; azimutal; azolve.

B: bazuka; berenjena; beriberi; bicoca; bifocal; bípedo; bisección; bisexual; bocio; boludo; botija; bozo.

C: canevá; canuto; cerbatana; cisma; climaterio; clitelo; cocada; cocoa; corifeo; coz; cubicar; cúbito; cursiva.

D: depauperar; diacrítico; diptongo; divisibilidad; divisible.

E: encefalitis; enchiquerar; entuerto; estabular; eufónico; extractivo.

F: fofo.

H: hado; hendido; hipérbaton; hipocorístico; homología; homotecia; hoz; húmero.

I: icono; inanición; inquina; inquisición; insidia; internacionalismo.

J: jeta.

K: kilométrico.

L: librecambio.

M: manido; matriarcado; metonimia; mies; módico; modoso.

N: nicotina; nidada; níveo; nonato.

Ñ: ñapa; ñato; ñoño.

O: ojiva; ópalo; orea; osario.

P: pápalo; paridera; patata; patriarcado; peroné; piara; plaqueta; pleonasmo; porosidad; pote; proclítico; protectorado; puntiagudo.

R: rótula.

S: secuaz; sedición; sinécdoque; sinovial; subarrendatario; sucedáneo.

T: tirria; toga; transpirar; trapezoide; tutor.

V: vale; ventosa; vira; vocear.

Z: zahúrda; zaino; zootecnia; zueco.

En el noveno nivel de importancia y último, se encuentran las palabras que no recuperó el sistema de información y que tampoco tienen una documentación en los diccionarios utilizados en el DEM, que permita considerarlas relevantes. Por otro lado, se puede observar que estas palabras principalmente son de terminología especializada y en este estudio resultaron ser la información no relevante, no recuperada:

A: asociatividad; autoconsumo.

C: cardinalidad.

E: extensionista.

N: newcastle.

V: velarización.

OBRAS CONSULTADAS

ALA world encyclopedia of library and information services, Chicago, American Library Asssociation, 1980, pp. 317-319.

Alcalá, Antonio y Huberto Batis. *La comunicación humana y la literatura*, México, ANUIES, 1973, 47 p. (Temas básicos. Área: lengua y literatura).

Alvar Esquerra, Manuel. *Proyecto de lexicografía española*, Barcelona, Planeta, 1976, 271 p. (Ensayos / Planeta. Lingüística y crítica literaria).

Alvar López, Manuel. *Informática y lingüística*, Málaga, España, Ágora, 1984, 117 p. (Cuadernos de lingüística; 1).

Amat Noguera, Nuria. *Técnicas documentales y fuentes de información*, Barcelona, Bibliográf, 1978, pp. 67-79.

Andreewski, Alexandre. "Indexação automática baseada en métodos lingüísticos e estadísticos e sua aplicabilidade a língua portuguesa", pp. 61-73, en *Ciencia da informação*, vol. 12, núm. 1 (1983).

Araujo, Arastóstenes E. R. "Revocação (recall) e precisão (presición) no SDI/CIN/CN", pp. 47-50, en *Ciencia da informação*, vol. 8, núm. 1 (1979).

Baranow, Ulf Gregor. "Perspectivas na contribuição da lingüística e de áreas afins a ciencia da informação, pp. 23-36, en *Ciencia da informação*, vol. 12, núm. 1 (1983).

Barhydt, G. C. "The efectiveness of non user relevance assessments", pp. 146-149, en *Journal of documentation*, vol. 23, núm. 2 (1967).

Bastos Vieira, Simone. "Indização automática e manual: revisão de literatura", pp. 43-57, en *Ciencia da informação*, vol. 17, núm. 1 (1988).

Bradfor, S. C. "The documentary chaos", pp. 106-121, en su *Documentation*, London, Loogkwood.

———. *Sources of information on specific subjects, pp 85-86, en Engineering*, 137 (1934)

Bresson, François. *Lenguaje, comunicación y decisión*, Buenos Aires, Paidós, 1974, 376 p. (Tratado de psicología experimental; 8).

Camarero García, E. y M. F. Verdejo. "Un sistema pregunta-respuesta en castellano, sobre un corpus literario", pp. 4-12, en *Boletín del Centro de Cálculo de la Universidad Complutense*, núm. 32 (mayo 1978).

Carpenter, Ray L. y Ellen Storey Vasu. *Métodos estadísticos para bibliotecarios*, México, Universidad Nacional Autónoma de México, Dirección General de Bibliotecas, 1980, 153 p.

Casares y Sánchez, Julio. *Cosas del lenguaje, etimología, lexicología, semántica*, Madrid, Espasa Calpe, [1961], 236 p. (Colección Austral; 1305).

El Colegio de México. Diccionario del Español de México. *Corpus del español mexicano contemporáneo (Cemc)*, México, Diccionario del Español de México, 1975, *ca.* 100 h.

———. *Corpus del español mexicano contemporáneo (Cemc): lista bibliográfica del corpus*, México, Diccionario del Español de México, 1975, *ca.* 100 h.

———. *Diccionario básico del español de México*, México, El Colegio de México, 1986, 565 p.

———. *Diccionario del español de México: manual de redacción*, México, Diccionario del Español de México, 1975, *ca.* 100 h.

———. *Diccionario fundamental del español de México*, México, Fondo de Cultura Económica, 1982, 480 p.

———. *Documentos de trabajo del* DEM, México, Diccionario del Español de México, 1984, 2 v.

———. *Nomenclatura del diccionario de primaria y lista estadística de vocablos del español mexicano contemporáneo*, México, Diccionario del Español de México, 1981, *ca.* 100 h.

Coll-Vinet, Roberto. *Teoría y práctica de la docmentación*, 2a ed, Barcelona, ATE, 1978, 402 p.

Coloquio sobre Automatización de Bibliotecas de México (1er.: 1984: Co-

lima). *Memorias*, México, Universidad Autónoma Metropolitana, Unidad Xochimilco, 1985, 326 p.

Coyaud, Maurice. *Linguistique et documentation: les articulations logiques du discours*, Paris, Libraire Larousse, 1972, 173 p. (Langage et langue).

Cuadra, C. A. y Katter, R. V. "Opening the black box of relevance", pp. 291-303, en *Journal of documentation*, vol. 23, núm. 4 (1976).

Currás, Emilia. *La información en sus nuevos aspectos: ciencias de la documentación*, Madrid, Paraninfo, 1988, 307 p.

Chávez García, Beatriz. *Análisis de legibilidad de textos para la postalfabetización*, México, La autora, 1985, 184 p. Tesis (Licenciada en Bibliotecología). Universidad Nacional Autónoma de México, Colegio de Bibliotecología.

Díaz-Guerrero, Rogelio, Miguel Salas. *El diferencial semántico del idioma español*, México, Trillas, 1975, 111 p.

Dym, Eleanor D., *et al. Subject and information analysis*, New York, Marcel Dekker, 1985, 488 p. (Books in library and information sciences; 47).

Enciso, Berta. *La biblioteca: bibliosistemática e información*, México, El Colegio de México, 1983, 142 p.

Evens, Martha. "Computer-readable dictionaries", pp. 85-117, en *Annual reiew of information science and technology*, vol. 24 (1989).

Faithone, R. A. "Empirical hiperbolic distributions (Bradford-Zipf-Mendelbrot) for bibliometric description and prediction", pp. 321-342, en *Journal of documentation*, vol. 25, núm. 4 (1969).

Fernández Gordillo, Luz. *La problemática de las macroestructuras en el diccionario general*, México, La autora, 1982, 207 h. tesis (Licenciada en Letras Hispánicas)— UNAM, Colegio de Letras Hispánicas.

Figueiredo, Laura Maia de. "O conceito de relevância e sus aplicações", pp. 75-78, en *Ciencia da informação*, vol. 6, núm. 2 (1977).

Figueiredo, Regina Célia. "Estudio comparativo de julgamentos de rele-

vância do usuário e não usuário de serviços de D. S. I.", pp. 69-78, en *Ciencia da informação*, vol. 7, núm. 2 (1978).

Foskett, D. J. "A note on the concept of relevance", pp. 77-78, en *Information storage retrieval*, vol. 8, núm. 2 (apr. 1972).

François, Frederic, *el al. El lenguaje y la comunicación*, Buenos Aires, Nueva Visión, 1973, 148 p. (Tratado del lenguaje; 1).

Fomm, Erich. ed., *Marx y su concepto de hombre. Karl Marx: manuscritos económico-filosóficos*, México, FCE, [1962], pp 205-226, (Breviarios del Fondo de Cultura Económica; 116).

García Hidalgo, María Isabel. "La formación del analizador gramatical del DEM", pp. 85-155, en Lara , Luis Fernando *et al. Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, 266 p. (Jornadas; 89).

Garza Ramos, Georgina Madrid. "Panorama de cambios en la bibliografía contemporánea", pp. 57-65, en *Anuario de bibliotecología*, Época 4, núm. 3 (1982).

Gifford, C. y Baumanis, G. J. "On understanding user choices: textual correlates on relevance judgements", pp. 21-26, en *American documentation*, vol. 20, núm. 1 (jan. 1969).

Goffman, W. "A general theory of communicarion"., pp. 726-747, cit. en, Saracvic, Tefko. *Introduction to information science*, New York, 1970.

Haller, Johan. "Indexação automática de textos", pp. 27-32, en *Revista biblioteconomia*, vol. 13, núm. 1 (1985).

Ham Chande, Roberto. "Del 1 al 100 en lexicografía", pp. 41-43, en Lara Luis Fernando *et al. Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, (Jornadas; 89).

Harmon, G. "Information need transformation during inquiry: a reinterpretation of user relevance", pp. 41-43, en *Proc. ASIS*, núm. 7 (1970).

Instituto Nacional para la Educación de los Adultos (México). *Memoria 1982-1988*, México, El Instituto, 1988, pp. 17-46.

Kent, Allen. *Information analysis and retrieval*, New York, Becker and Hayes, 1971, pp. 98-113, (Information sciences series).

Kozachov, L. S. "Relevance in informatics ands scientology", pp. 3-11 en *Nauchno-Teknicheskaya Informatsiya*, Serie 2, núm. 8 (1960).

Kochtaneck, Thomas Richard. *A general method for identifiying relevant document sets from a know relevant document*, Michigan, University Microfilms International, 1984, 97 p.

Lancaster, Frederick Wilfrid, M. J. Joncich. *Evaluación y medición de los servicios bibliotecarios*, México, Universidad Nacional Autónoma de México, Dierección General de Bibliotecas, 1983, 443 p. (Estudios de bibliotecología; 1).

———. Some notes on the distinction between pertinence and relevance", en su *Guidelines for the evaluation of information systems and services* [s.p.i.], 1977, Preparado para la Unesco bajo contrato

———. *Toward paperless information sistems*, New York Academic Press, 1978, 179 p. (Library and information systems).

Lara, Luis Fernando. *Dimensiones de la lexicografía: a propósito del Diccionario del español de México*, México, El Colegio de México, 1990, 249 p. (Jornadas; 116).

———. "Noticia del diccionario del español de México", pp. 63-66, en *Boletín editorial de El Colegio de México*, núm. 33 (sept./ oct. 1990).

——— y Roberto Ham Chande. "Base estadística del Diccionario del Español de México", pp. 7-39, en su *Investigaciones lingüísticas en lexicografía*, México, El Colegio de México, 1979, 266 p. (Jornadas; 89).

Luhn, H. P. "A statistical approach to mechanized encoding and searching of literay information", pp. 309-317, en *IBM journal*, vol. 1. núm. 4 (oct. 1957).

Luna López, Francisco. *Lingüística*, México, Secretaría de Educación Pública, 1982, 91 p. (Biblioteca del maestro; 7).

Manrique, Leonardo. "La evolución humana relacionada con la evolución lingüística", pp. 151-167, en *Hombre, tiempo y conocimiento*, Mé-

xico, Escuela Nacional de Antropología e Historia, 1982.
Marín, Francisco Marcos. *Reforma y modernización del español: ensayo de sociolingüística histórica*, Madrid, Cátedra, 1979, 149 p.
Marquez Cintra, Anna Maria. "Elementos de lingüística para estudos de indexação", pp. 5-22, en *Ciencia da informação*, vol. 12, núm. 1 (1983).
Martinet, André. *Elementos de lingüística general*, 2a ed, Madrid, Gredos, [1972], 274 p. (Biblioteca Románica Hispánica III. Manuales; 13).
Martínez Márquez, Alejandro. *Revisión del estado actual de la automatización de los procedimientos de almacenamiento y recuperación de información documental*, /s.p.i./, *ca.* 100 h.
Meadow, Carles T. *The analysis of information systems: a programmers' introduction to information retrieval*, New York, John Wiley & Sons, 1967, 301 p. (Information sciences series).
Meyriat, Jean, Micheline Beauchet. *Guía para establecer centros de documentación en ciencias sociales en los países en vías de desarrollo*, México, Universidad Nacional Autónoma de México, Instituto de Investigaciones Sociales, 1973, 128 p.
Mikhailov, Aleksandr Ivanovich y R. S. Guiliabesuskü. *Fundamentos de la informática*, La Habana, Academia de Ciencias de Cuba, Instituto de Documentación e Informática Científica y Técnica, 1973, 2 v.
Millán, Antonio. *Lengua hablada y lengua escrita*, México, ANUIES, 1973, 39 p. (Temas básicos. Área: lengua y literatura).
Mooers, C. S."Coding, information retrieval, and the rapid selector";-- pp. 225-229, en *American documentation*, vol. 1, núm. 4 (1950).
Muller, Charles. *Estadística lingüística*, Madrid, Gredos, 1973, 116 p. (Biblioteca románica hispánica II. Estudios y ensayos; 201).

National Academy of Sciences. *Proceedings of the International Conference on Science Information*, Washington D. C., National Academy of Sciences, 1959, 2 v.
Nocetti, Milton A. "Línguas naturais e linguagens documentárias: traços inerentes e ocorrências de interação", pp. 23-37, en *Revista bibliotecономia*, vol. 6, núm. 1 (1978).

O'Connor, J. "Relevance dissagreements and unclear request forms", pp. 165-177, en *American documentation*, vol. 18, núm. 3 (1967).

Perales Ojeda, Alicia. "La intersección lingüística en la ciencia de la información", pp. 184-192, en Congreso Internacional sobre el Español de América (2o : 1986: México). *Actas*, México, Universidad Nacional Autónoma de México, Facultad de Filosofía y Letras, 1986.

Perry, J. W. "Superimposed punching of numerical codes on handsorted, punch cards", pp. 205-212, en *American documentation*, vol. 2, núm. 4.

Pignatari, Decio. *Informação, linguagem, comunicação*, Sao Paulo, Brasil, Perspectiva, 1976, 147 p. (Debates; 2).

Polushkin, V. A. "Relevance and pertinence", pp. 52-54, en *Automatic documentation in mathematics linguistics*, vol. 7, núm. 1 (1973).

Pritchard, A. "Statical bibliography or bibliometrics?", pp. 348-349, *Jorunal of documentation*, 25 (1969).

Ramírez Leyva, Elsa Margarita. "El índice de citas bibliográficas", pp. 151-198, en *Anuario de bibliotecología, archivología e informática*, Época 3, año 7 (1978).

Rath, G. J. *et al.* "Comparisons of four types of lexical indicators of content", pp. 126-130, en *American documentation*, vol. 12, núm. 2 (1961).

Resnick, A. y Savage, T. R. "The consistency of human judgements of relevance", en *American documentation*, vol. 15, núm. 2 (1964).

Ribeiro, L. A. "Aplicação dos métodos estatísticos e da teoría da informação e da comunicação na análise lingüística: estudo da linguagen jornalística", pp. 151-154, en *Ciencia da informação*, vol. 3, núm 3 (1974).

Robertson, S. E. y Sparck Jones, J. "Relevance weighting of reserch terms", pp. 129-146, en *Journal of the American Society for Information Science*, 27 (1976).

Robredo, Jaime, "Documentação de hoje e de amanhña", *Brasilia* ABDF, 8 (1976).

———. *Elaboración de un thesaurus agrícola basado en criterios de eficien-*

cia del lenguaje en el proceso de comunicación, Brasil, Ministerio de Agricultura, 1972, 21 h.

Robredo, Jaime, "A indexação automática de textos: o presente já entrou no futuro", pp. 236-274, en Machado, U. D. *Estudos avançados em biblioteconomia e ciência de informação*, Brasilia, ABDF, 1982, v. 1.

———. "Otimização dos processos de indexação dos documentos e recoperação da informação mediante o uso de instrumentos de control terminológico", pp. 3-18, en *Ciencias da informação*, vol. 11, núm. 1 (1982).

——— y José Alberto de Paula Ferreira. "Conceituação de um programa para indexação automática de textos", pp. 254-263, en *Revista biblioteconomia*, vol. 8, núm. 2 (1980).

Saez Godoy, Leopoldo. "Los inventarios léxicos automatizados y el español: proposiciones terminológicas", pp. 67-82, en *Literatura y lingüística*, núm. 1 (1988).

Salton, G. "Automated languaje processing", pp. 169-199, en *Annual review of information science and technology*, 3 (1968).

———. "Computer evaluation of indexing and text procesing", en *Journal of the Association for computer machinery*, vol. 15, núm. 1 (1968).

———. y Yang, C. S. "On the specification of term values in automatic indexing", pp. 351-372, en *Journal of documentation*, vol. 29, núm. 4 (dec. 1973).

Saracevic, Tefko. *Introduction to information science*, New York, Bowker, 1970, pp. 110-151.

———. *On The concept of relevance in information science*, Cleveland, Ohio, Case Western Reserve University, 1970.

———. *Relevancia: una reseña y una estructura para considerar el concepto en ciencia de la información*, México, Asociación de Bibliotecarios de Instituciones de Enseñanza Superior, 1978, 72 h. (Cuadernos de ABIESI; 7).

———. "Selected results from an inquiry into testing of IR systems", pp. 126-139, en *Journal of American Society of Information Science*, vol. 32, núm. 2 (apr. 1971).

Saussure, Ferdinand. *Curso de lingüística general*, Buenos Aires, Losada, 1945, 378 p. (Filosofía y teoría del lenguaje).

Serrano Limón, Jorge. *Sistema automático para la generación de un diccionario del español de México*, México, El autor, 1975, Tesis (Licenciado en Administración de Empresas) Universidad Latina.

Shannon, Claude. E. y W. Weawer. *The mathematical theory of communication*, Urbana, Illinois, University of Illinois, 1949, 117 p.

Shore, L. "The measure of reference", pp. 297-302, en *Southeastern Librarian*, 11 (1961).

Simposio de la Asociación Mexicana de Lingüística Aplicada (4o: 1987: México). *Presente y perspectivas de la lingüística computacional en México*, México, Asociación Mexicana de Lingüística Aplicada, 1987, 2 v.

Soergel, Dagobert, "Automatic and semi-automatic methods as an aid the construction of indexing languajes and theasauri", pp. 34-39, en *International classification*, vol. 1, núm. 1 (may 1974).

Sparck, Jones, Karen. "Idexing term weighting", pp. 619-633, en *Information storage and retrieval*, vol. 9, núm. 11 (nov. 1973).

———. *Linguistics and information science*, New York, Academic Press, 1973, 244 p. (Library and information science).

Steinacker, Ivo. "Indexing and automatic significance analysis", pp. 237-241, en *Journal of American Society of Information Science*, vol. 25, núm. 4 (jul./ago. 1974).

———. "A statistical interpretation of term specificity and its application in retrieval", pp. 11-21, en *Journal of documentation*, vol. 28, núm. 1 (mar. 1972).

Tague, J. "Matching of question and answer terminology in an educational research file", pp. 26-32, en *American documentation*, vol. 16, núm. 1 (1965).

Taube, M. *et al.* "Storage and retrieval of information by means of the association of ideas", pp. 1-17, en *American documentation*, vol. 6, núm. 1 (1955).

Travis y Fidel, R. "Subjet analysis", pp. 123-157, en *Annual review of in-*

formation science and technology, 17 (1982).

"The Unisist draft on indexing principles: test and coments", pp. 29-34, en *International classification*, vol. 4, núm. 1 (may 1977).

Vickery, B. C. " Structure and funtion in retrieval languages", pp. 69-82, en *Journal documentation*, vol. 27, núm. 2 (1971).

Willson, P. "Situational relevance", pp. 457-471, en *Information storage retrieval*, vol. 9, núm. 8 (1973).

Zipf, G. K. *The psycho-biology of language: an introduction to dynamic philology*, Cambridge, Mass., MIT, 1965, 336 p.