NUEVA REVISTA DE FILOLOGÍA HISPÁNICA
TOMO XXIII NÚM. 2

BASE ESTADÍSTICA DEL DICCIONARIO DEL ESPAÑOL DE MÉXICO

0. El *Diccionario del Español de México*[1] es una obra lexicográfica cuyo objetivo fundamental es reflejar el léxico del español utilizado actualmente en el país, en cuanto "lengua nacional" y en cuanto a sus modalidades escritas y orales, cultas y coloquiales, urbanas y rurales.

Tal pluralidad de necesidades —tanto más obligatorias cuanto que el trabajo tiene una finalidad práctica inmediata: servir a un hablante mexicano medio como obra de consulta y como punto de referencia en su apreciación pre-científica del idioma— se enfrenta con algunos problemas que derivan de la naturaleza misma de la lexicografía: 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 (y, en particular, la hispanoamericana). Este último, en razón de la influencia que pudieran tener sobre nuestro trabajo los diccionarios existentes del español.

[1] El proyecto de investigación para un *Diccionario del Español de México* se inició en el Centro de Estudios Lingüísticos y Literarios de El Colegio de México a partir de enero de 1973, bajo la dirección de Luis Fernando Lara y con la colaboración de un pequeño número de investigadores. Se han hecho presentaciones generales del proyecto en: L. F. LARA, "Sobre la justificación de un diccionario de la lengua española hablada en México", *La Gaceta,* Fondo de Cultura Económica, 19 (julio de 1972), 1-5; "El Diccionario del Español de México, informe sobre su desarrollo", Reunión Continental sobre la Ciencia y el Hombre, Simposio de Sociolingüística y Planeación Lingüística, CONACYT-AAAS, México, junio de 1973; "La elaboración del Diccionario del Español de México", comunicación al XIV Congreso Internacional de Lingüística y Filología Románica", Nápoles (Italia), abril de 1974; L. F. LARA e ISABEL GARCÍA HIDALGO, "El uso de la computadora electrónica en la elaboración del *Diccionario* del Español de México", Reunión de trabajo sobre la aplicación de las computadoras en el área de ciencias sociales, Instituto Nacional de Antropología e Historia — Centro de Investigación en Matemáticas Aplicadas y Sistemas, México, 2-3 mayo, 1974.

En este artículo nos referimos a los problemas de la objetividad en la descripción del léxico mexicano, y al de la cantidad de datos necesaria para la labor lexicográfica. Particularmente nos ocupamos de la aplicación de la estadística lexicológica en la investigación del español de México como el mejor instrumento de documentación y análisis del vocabulario.

1. Los problemas básicos que se plantean al iniciar las labores de documentación para cualquier diccionario de interés lingüístico se pueden definir respecto a: *a)* los documentos lexicológicos de que se dispone antes de iniciar el trabajo; *b)* el valor científico de tales documentos; *c)* el uso que se les pueda dar en el trabajo que se emprende; *d)* la recolección de nuevos datos que complementen o sustituyan a los anteriores.

Una caracterización muy general de la lexicografía española nos permite distinguir dos tipos de obras lexicográficas al alcance de los hablantes: los diccionarios *generales* del español y los diccionarios de *regionalismos* del español. Los primeros dependen, en su totalidad, de los registros que presenta el *Diccionario de la Lengua Española (DRAE)* como reflejo de la labor de las Academias de la Lengua en el mundo hispánico[2]. En todos ellos hay algunos vocablos no registrados por el *DRAE,* pero fundamentalmente se trata de trabajos individuales en los que la inclusión de voces depende de las necesidades lexicológicas que el autor supone que existen. En ninguno de ellos hay una descripción del léxico como la lingüística moderna lo exigiría.

Los diccionarios de regionalismos eliminan, por definición, "el vocablo de estructura y significado estrictamente castizo, es decir, incluido como tal en el Diccionario vulgar de la Academia Española"[3]. No hay en ellos, por lo tanto, una descripción integral del

[2] Cf. MARÍA MOLINER, *Diccionario de uso del español* (Madrid, 1970): "En un principio se pensó tomar las definiciones para este 'diccionario de uso' del *Diccionario de la Lengua Española,* diccionario oficial de la 'Real Academia Española', como lo han hecho hasta ahora absolutamente todos los diccionarios españoles" (pp. xiii-xiv). Si bien en el aspecto de las definiciones Moliner sostiene haber reconstruido totalmente el diccionario *(loc. cit.),* más adelante dice: "Están incluidas en el presente diccionario todas las voces contenidas en el D.R.A.E., con ... excepciones" (p. xxiv); en la sección de "Obras utilizadas" dice que, "dejando aparte las obras de consulta empleadas esporádicamente, cuya relación completa sería difícil de hacer y carecería de interés, se basa fundamentalmente en el *Diccionario de la Lengua Española* de la Real Academia Española ... seguido paso a paso en la redacción de los artículos, si bien refundiendo y reorganizando las acepciones". El *Diccionario general ilustrado de la lengua española* (VOX), de S. GILI Y GAYA no parece haberse separado de esa tradición, aunque en su prólogo no se indica explícitamente (véase p. xxx).

[3] FRANCISCO J. SANTAMARÍA, *Diccionario de Mejicanismos,* 1ª ed. Porrúa, México, 1959. Introducción, p. xii.

español usado en cada región del mundo hispánico; su utilidad para el usuario se define solamente por las singularidades léxicas de la región en cuestión y no por la presentación de la común lengua española.

Esta clara bipartición entre diccionarios generales y de regionalismos permite suponer que, hasta hoy, la lexicografía hispánica ha intentado establecer una coordinación entre ambos tipos de diccionario; pero, como se puede ver, el resultado obtenido deja una amplia laguna respecto a la realidad del léxico español en cualquier parte de sus dominios y no permite que el hablante tenga ante sí la imagen real de una lengua extendida por millones de kilómetros, revitalizada por las diferentes comunidades que la utilizan y, sin embargo, unida de una manera sorprendente si se la compara con otras lenguas de cultura en condiciones semejantes.

La documentación lexicológica de que disponemos es fragmentaria, guiada por oscuros principios de recolección y de un valor relativo para los usuarios, ya que los diccionarios generales no dan cabida suficiente al español de distintas regiones, y los de regionalismos no dan lugar al vocablo general o a diversas particularidades del significado de los vocablos.

Desde un punto de vista exclusivamente lingüístico el valor de la documentación de que se dispone actualmente es aún más discutible, puesto que los registros no son homogéneos, mezclan diferentes estados de lengua y se dejan llevar por criterios como el purismo, reales para una comunidad, pero totalmente ajenos al estudio descriptivo.

En cuanto al uso que se puede dar a los documentos lexicológicos anteriores a nuestro trabajo, se hace necesario definir previamente las características de nuestro diccionario para, de acuerdo con ellas, saber exactamente hasta dónde nos son útiles ya que todo registro lexicológico tiene valor en sí mismo. Por esta razón queremos exponer, aunque sea brevemente, las pautas que han guiado nuestro modelo de diccionario.

1.1. El *DEM* se define como un diccionario sincrónico[4], descriptivo y —por limitaciones de tiempo y dinero— selectivo[5]. Nos

[4] La "sincronía práctica" (como propone llamarla J. Rey-Debove) para el *DEM* ha quedado definida respecto a todo texto (hablado o escrito) que se haya producido entre 1970 y 1973; en el caso de obras literarias (o científicas) cuya publicación sea anterior a esa fecha y aparezcan como textos fundamentales para el corpus, las publicadas después de 1921.

[5] La selectividad del *DEM* se refiere exclusivamente a la cantidad de vocablos que hemos calculado podrá contener y no a la existencia de criterios prescriptivistas. Evidentemente, para reducir el tamaño de la nomenclatura habrá que seleccionar los vocablos más usuales. Para evitar toda confusión con otras acepciones del término "selectivo" vale la pena recordar que, según la

interesa mostrar en él el léxico del español que se utiliza entre las fronteras geográficas de México[6] y, a diferencia de los diccionarios de regionalismos, lo entendemos como un *diccionario regional* de la común lengua española.

Esta breve descripción tipológica de nuestro trabajo nos permite evaluar los registros lexicológicos que existían antes del inicio de nuestra investigación. Para usar los diccionarios del español que se encuentran hasta hoy tendríamos que distinguir entre varias sincronías allí mezcladas, analizar severamente los criterios de inclusión de vocablos en los diccionarios y comprobar la existencia de cada uno de ellos en México con el objeto de mostrar una realidad. Resulta fácil ver que, por un lado, el trabajo de desglose y comprobación de los datos sería inmenso y, por otro, de ninguna manera conduciría a la descripción del español usual en México. Es necesario, por lo tanto, buscar otro método más adecuado a nuestras necesidades.

1.2. Son bien conocidas las aplicaciones que se han hecho de las cuantificaciones estadísticas a los estudios lingüísticos[7]; también se sabe hasta qué punto un análisis estadístico puede llegar a presentar situaciones de solución imposible para la lexicografía[8]. No

tipología que ofrece L. Zgusta, el *DEM* se inscribe entre los "standard-descriptive dictionaries"; cf. LADISLAV ZGUSTA, *Manual of lexicography*, Praga, y La Haya, 1971, p. 210.

[6] Las fronteras geográficas de la República Mexicana difícilmente tienen alguna realidad desde el punto de vista lingüístico. Al norte, el español se extiende por el sur de los Estados Unidos y al sur solamente podría existir alguna frontera dialectal en alguna parte de Centro América. Pero el hecho de que el *DEM* se preocupe por mostrar el español "nacional" de México nos obliga a respetar cuidadosamente nuestras fronteras políticas. Hay que señalar, sin embargo que, indudablemente, el *DEM* será útil para cualquier comunidad hispanohablante.

[7] Cf. ALPHONSE JUILLAND y EMILIO CHANG-RODRÍGUEZ, *Frequency dictionary of Spanish words*, La Haya, 1964, y U. BORTOLINI, C. TAGLIAVINI y A. ZAMPOLLI, *Lessico di frequenza della lingua italiana contemporanea*, Milán, 1972 en donde aparecen reseñas críticas de varios trabajos dedicados a la estadística lingüística; entre otros citamos los de M. A. BUCHANAN, *A graded Spanish word book*, Toronto, 1927; L. RODRÍGUEZ BOU, *Recuento de vocabulario español*, Puerto Rico, 1952; y V. GARCÍA HOZ, *Vocabulario usual, vocabulario común y vocabulario fundamental*, Madrid, 1953.

[8] Análisis críticos del valor de la estadística en lexicografía se pueden encontrar, por ejemplo, en J. REY-DEBOVE, *Étude linguistique et sémiotique des dictionnaires françaises contemporains*, Mouton, La Haya, 1971, especialmente pp. 67-68; CHARLES MULLER, "Un dictionnaire de fréquence de l'espagnol moderne", *ZRPh*, 81 (1965), 476-483 (en adelante, *Muller 65a); id.*, "Fréquence, dispersion et usage: à propòs d'un dictionnaire de fréquence", *CLex*, 7 (1965), 33-42 (en adelante, *Muller 65b*). Véase también K. H. DEUTRICH y G. SCHOENTAL, "Der Stellenwert der Statistik im Freiburger Analyse-Modell gesprochener Sprache", *Linguistique et statistique*, Colloque organisé par le Centre

obstante, en vista de las dificultades aún mayores que representa el uso de procedimientos tradicionales de recolección lexicológica, hemos considerado que el método estadístico es el único capaz de darnos los registros necesarios y la cantidad de datos suficientes para nuestra tarea lexicográfica de un modo objetivo e imparcial. Describiremos ahora los puntos de vista que han quedado en la base de nuestro análisis, con el objeto de demostrar la utilidad del método seleccionado.

2. Del análisis estadístico de un corpus de datos deseamos obtener:

a) Un número elevado de *vocablos*[9] que puedan constituir la mayor parte de las entradas del diccionario.

b) Una base imparcial de selección de vocablos para la primera edición del *DEM*.

c) Un punto de referencia que nos permita detectar los usos diferentes de los vocablos en la sociedad mexicana.

Estas tres necesidades nos colocan frente a frente, por una parte, con lo que significa un corpus para la lexicografía y la lingüística y, por la otra, con la concepción del corpus para la estadística.

2.1. Desde el punto de vista de la lingüística se hace necesario recordar, como lo ha señalado K. Heger[10], que todo corpus de datos produce tanto un número de documentaciones menor que el número de posibles ocurrencias que se pueden obtener del sistema, como un número mayor de ocurrencias que las que puede generar el sistema (como en el caso de las erratas, que son documentables y sin embargo no son realizaciones del sistema). Además, un corpus, por exhaustivo que sea, no será un documento completo del sistema y por ello el lingüista se verá obligado a trabajar continuamente extrapolando entidades del sistema de entre las realizaciones y, en consecuencia, un corpus de datos lingüísticos es una ayuda muy necesaria para el trabajo, pero no constituye una fuente exclusiva de materiales[11].

d'Analyse Syntaxique de l'Université de Metz, J. David y R. Martin (eds.), Klincksieck, Paris, 1974, pp. 95-104.

[9] Utilizamos la distinción propuesta por CH. MULLER ("Le mot, unité de texte et unité de lexique en statistique lexicologique", *TLL*, 1, 1963, 155-173) y K. HEGER ("Die Semantik und die Dichotomie von langue und parole", *ZRPh*, 85, 1969, 144-215), según la cual la unidad de lengua que nos interesa es el *vocablo* al que corresponde la *ocurrencia* en el *habla* y el *tipo* en la Σ *hablas*.

[10] Cf. KLAUS HEGER, "Belegbarkeit, Akzeptabilität und Häufigkeit", *Theorie und Empire der Sprachforschung*, H. Pilch y H. Richter (eds.), Basilea, 1970, 22-33.

[11] K. HEGER, *Monem, Wort und Satz*, Tübingen, 1971, § 1.2, p. 9.

2.2. Para la lexicografía, en virtud de su carácter aplicado, la consideración del corpus puede tomar en cuenta las exigencias que impone la lingüística (y en nuestro caso debe hacerlo), pero también debe basarse en la riqueza del material y en el grado de objetividad de los datos respecto a una realidad léxica determinada. Estas dos condiciones son de una importancia extrema, puesto que son las que verdaderamente califican la utilidad del corpus; así, en cuanto a la riqueza de los materiales, J. Rey-Debove ha señalado que el tamaño reducido de todo corpus en comparación con el volumen real del léxico "entraîne des conséquences genântes: (1) l'absence de très nombreaux mots, (2) l'importance relative accrue des mots employés plusieurs fois par le même auteur (un idiolecte) par rapport à ceux employés une fois par plusieurs auteurs (plusieurs idiolectes), qui ont une valeur d'échange plus grande (d'où la nécessité de corriger la notion de fréquence par celle de répartition), (3) la faible fréquence des mots thématiques (liés à un thème conceptuel à l'exclusion d'un autre), même courants, due au fait que tous les thèmes ne sont pas abordés dans le corpus. D'où la nécessité de corriger la notion de fréquence par celle de disponibilité"[12].

En cuanto al grado de objetividad del muestreo, conviene señalar la diferencia entre juzgar los resultados a partir de una consideración intuitiva de la "realidad" léxica (con lo que la evaluación se torna imposible al quedar sujeta a la experiencia de cada hablante), juzgarlos en comparación con trabajos realizados bajo muy diferentes enfoques (por ejemplo, con diccionarios anteriores elaborados de manera tradicional) y evaluarlos tras un análisis de sus diferencias con obras estadísticas previamente elaboradas. Este último criterio nos parece el único consecuente con el método seleccionado, pero aun aceptando otro tipo de evaluación subjetiva, la forma en que se realizó el muestreo para el *DEM* nos permite esperar un gran acercamiento a las "realidades léxicas" de los hablantes[13].

2.3. Para la estadística, el léxico de una lengua es el resultado de la unión de los léxicos individuales de los hablantes y constituye un conjunto finito. Pero, como Martinet observa[14], una característica esencial de todo léxico es su carácter "abierto", el constante aumento de vocablos dentro de una lengua, con lo que la identificación del conjunto resulta imposible; además, el léxico individual

[12] J. Rey-Debove, *op. cit.*, § 3.4.2.2, p. 68.

[13] Cf. *Muller 65b*, p. 33: "En matière de lexique au contraire, la probabilité ne pourra jamais être confrontée qu'avec les fréquences constatées dans de nouveaux échantillons extérieurs au corpus, mais appartenant au même état de langue".

[14] Cf. André Martinet, *Eléments de linguistique générale*, 6ª ed., Paris, 1966, § 4.19.

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 estas circunstancias, solamente se puede definir el léxico común del español mexicano como una intersección de léxicos individuales.

Al tomar en cuenta a los hablantes para seleccionar sus léxicos particulares, encontramos que, en nuestro caso, el número de mexicanos hablantes del español es elevadísimo y hace imposible conocer todos sus léxicos individuales; por ello nos vemos obligados a establecer, en primer término, una muestra de hablantes o, lo que es lo mismo, una muestra de los textos producidos por los hablantes. En consecuencia, desde un punto de vista estadístico, el universo es única y exclusivamente el conjunto de los textos que forman la muestra, es decir, el conjunto de vocablos con sus frecuencias de uso que podemos encontrar en los textos seleccionados. Ahora bien, en vista de que los objetivos de la muestra no son hacer inferencias sobre este universo restringido, sino sobre uno muchísimas veces mayor y —para los hablantes— de mayor interés como lo es el "español de México", se hace necesario suponer que el conjunto de textos que hemos establecido *representa* al español de los mexicanos.

2.4. Esta última suposición se basa en presupuestos lingüísticos —como señala J. Rey-Debove[15]— puesto que no existe ningún medio estadístico de asegurar previamente las características de representatividad que debe tener la muestra en el momento de seleccionarla. Queda, por lo tanto, la interrogante de cómo seleccionar el corpus de tal manera que no solamente podamos confiar en su acercamiento a la "realidad" del léxico del español mexicano, sino que también resulte un corpus en el que las limitaciones a que se hizo referencia en el § 2.2 sean superadas al máximo.

3. Tras revisar las causas más evidentes de distorsión en una muestra lingüístico-estadística podemos confiar en que nuestro corpus podrá eliminarlas en la medida en que tome en cuenta:

a) Una cantidad de textos lo suficientemente grande como para obtener el número de vocablos que deseamos (aproximadamente 30,000) y lo suficientemente reducida como para que sea económico y rápido su tratamiento con una computadora electrónica.

b) Una gran diversidad de textos que asegure la aparición del

[15] J. REY-DEBOVE, *loc. cit.*

mayor número de vocablos "disponibles"[16], es decir, que permita la entrada de muy diferentes temas.

c) Una gran diversidad de autores, que elimine, tanto como sea posible, los estilos individuales.

d) Una adecuada estratificación de los textos que permita obtener buenos resultados en el campo de la dispersión y el uso estadístico[17].

e) Una longitud suficiente de los textos, que permita la recuperación del significado global en que aparezcan los vocablos, para que la definición lexicográfica cuente con todos los elementos de juicio necesarios para el análisis semántico.

3.1. Respecto al tamaño y costo de la muestra nos hemos orientado por los objetivos centrales del *DEM* (presentación de una "lengua nacional" con todas sus variedades, cf. *supra* § 0) y no por cuestiones de precisión y confiabilidad sobre variables numéricas, como podría ser, por ejemplo, el cálculo de las frecuencias estimadas de uso de cada vocablo[18]. Esto significa que nos ha interesado

[16] El término *disponible* ha tenido su origen en los estudios de G. Gougenheim en torno al "francés fundamental"; se refiere a aquellos vocablos de baja frecuencia cuya utilidad es muy grande para cualquier usuario de un idioma; cf. JEAN DUBOIS *et al., Dictionnaire de linguistique*, Larousse, Paris, 1973, s. v. La disponibilidad de un vocablo puede verse directamente afectada por la mayor o menor variedad de textos que se exploren. Generalmente es más probable encontrar vocablos disponibles en una muestra muy diversificada que en una demasiado homogénea.

[17] Cf. *Muller, 65a*, p. 481: "C'est là surtout que la brièveté relative du corpus devrait être compensée par une stratification intensive à fin d'éliminer les mots propres à une discipline particulière, et de ne retenir que le vocabulaire commun du langage scientifique, et surtout à fin d'éviter des effets du sort comme celui que vient d'être cité" (a propósito del reducido número de "universos" empleados por Juilland y Chang Rodríguez). La frase de R. Moreau "estratifiquen a ultranza" se ha vuelto clásica en la formación de muestras estadísticas; véase su "Au sujet de l'utilisation de la notion de fréquence en linguistique", *CLex*, 3 (1962), 140-159.

[18] La muestra que constituye el corpus proporcionará una serie de estimaciones estadísticas sobre cada vocablo. Estas estimaciones son, por una parte, la frecuencia de uso del vocablo tanto dentro del total de la lengua, como dentro de cada género, y por otra las medidas específicas de la estadística lingüística. Si fijamos nuestra atención sobre alguna de las estimaciones —por ejemplo alguna de las frecuencias del uso del vocablo— estamos seguros de que, debido a las características del corpus como muestra y del léxico total como población teórica, no tendremos el valor exacto de esa frecuencia sino sólo una estimación que esperamos se acerque a ella de manera suficiente. Desde un punto de vista exclusivamente teórico estadístico, se podría diseñar un muestreo que nos garantizara, con una alta probabilidad (que nunca puede ser del 100%), que la estimación obtenida no se aleje en más de cierta cantidad (que nunca puede ser 0) del valor que tratamos de estimar. Sin embargo, para lograr una muestra tal para todos los vocablos identificados, o para la mayoría de ellos, con una precisión y un grado de confiabilidad aceptables, llegaríamos

menos el valor probabilístico de los vocablos en el corpus y que, en cambio, ha sido más importante calcular el tamaño y el costo de nuestro corpus respecto al número absoluto de vocablos que podamos encontrar en la muestra[19]. Para esto último las experiencias anteriores a la nuestra en cuanto al tamaño del corpus han sido muy aleccionadoras, como se ve en el siguiente cuadro:

Obra	Extensión del corpus	Vocablos obtenidos
Frequency Dictionary of Spanish Words [20]	500,000	5,024
Computational Analysis of Present-day American English [21]	1,014,232	50,406
Trésor de la Langue Française (TLF)[22]	70,317,234	71,415

Como se puede deducir de los datos del cuadro anterior, no es necesario contar con una muestra muy grande para obtener el número de vocablos que nos proponemos incluir en el *DEM*. Consideramos que un corpus de 2,000,000 de ocurrencias nos proporciona un vocabulario lo suficientemente amplio para nuestras necesidades, y que darle mayor extensión elevaría el costo desproporcionadamente respecto al número de palabras que se podrían obtener, como demuestran los resultados que obtuvo la exploración del *TLF*. A esto último hay que agregar que, dadas las proporciones que deseamos dar al *DEM*, los vocablos que se obtuvieran más allá de los dos millones de ocurrencias serían los menos usuales y que, más allá de la frontera de los 30,000 vocablos no solamente resultaría más

a tal complejidad en el diseño de la muestra y a un tamaño tan enorme, que automáticamente se invalida esta forma de analizar el problema.

[19] Cada una de las ocurrencias en el corpus tiene prácticamente el mismo costo de recolección y recuento, con la ligera excepción de aquellos textos hablados que deben ser transcritos previamente, pero que en realidad no representan costos adicionales notorios, por lo que la asignación de muestreo no lleva la complicación adicional de costos variables.

[20] El objetivo del trabajo de A. Juilland y E. Chang-Rodríguez no era hacer un diccionario en el sentido estricto del término, sino un estudio estadístico con finalidades estructuralistas. Los 5,024 vocablos que registra son solamente los que obtuvieron una frecuencia lo suficientemente alta como para hacer proyecciones científicas.

[21] Este corpus fue posteriormente parte básica de la nomenclatura del *American heritage dictionary of the English language,* William Morris (ed.), 1969.

[22] La cifra corresponde exclusivamente al corpus literario del *TLF*. Se ha publicado un prolijo estudio estadístico en que se explican los resultados de la investigación. Cf. *Dictionnaire alphabétique de fréquences,* C.N.R.S., Nancy, 1973.

sencillo sino también más aconsejable pasar a un procedimiento de documentación del tipo de diccionario-tesoro[23] en que todo registro es válido (como en el caso de voces usadas por solamente un autor en cierto texto, de vocablos que dependen excesivamente de las modas lingüísticas, etc.). Por otra parte, si se busca una fuerte estratificación de los tipos de texto, como se recomienda en el § 3.*d* y en la nota 17, puede acrecentarse el rendimiento final de la muestra para los fines de la lexicografía.

3.2. La estratificación interna de la muestra se orienta también por los objetivos finales del *DEM* en cuanto implica una idea de lo que entendemos por léxico del español mexicano. Hay que señalar que, al definir nuestro diccionario como *regional*, pensamos que el léxico que presente pertenecerá por lo menos a tres conjuntos: al de los vocablos comunes a todo el mundo hispánico, al de las voces comunes a dos o más comunidades hispanohablantes (una de ellas México), y al de los vocablos usuales solamente en nuestro país. Esto significa que la limitación a las fronteras políticas de la República Mexicana es válida únicamente para *fijar las documentaciones de los vocablos* y no para dar la impresión de que el español mexicano es totalmente distinto del español general, ni para negar la realidad de que, desde el punto de vista de la dialectología, nuestras fronteras no marcan, seguramente, regiones dialectales distintas entre México, Guatemala y el sur de los Estados Unidos.

El punto de referencia en la formación interna del corpus necesitaba, en consecuencia, quedar determinado por un análisis de los tipos de texto que se producen en México. Para ello acudimos al concepto de *lengua culta*, de larga tradición lexicográfica, pero también objeto de estudio por parte de algunas corrientes de la lingüística contemporánea.

Entendemos por *lengua culta* el uso de un idioma en la comunicación intelectual de sus hablantes, uso lo suficientemente fijo como para permitir un amplio entendimiento entre los usuarios, pero también lo suficientemente flexible como para aceptar todas las innovaciones que impone la vida cultural de la comunidad[24]. La

[23] Esta idea, desde luego, corresponde a un pensamiento esencialmente práctico acerca de la factura de diccionarios. Cf. el prólogo de don Ramón Menéndez Pidal a la primera edición del diccionario Vox, citado en nota 2.

[24] Cf. PAUL L. GARVIN, "The standard language problem: concepts and methods", en D. H. Hymes (ed.), *Language in culture and society*, Nueva York, 1964, pp. 521-528: las propiedades de la lengua estándar (cf. *infra*, nota 26) son "flexible stability and intellectualization. Flexible stability here refers to the requirement that a standard language be stabilized by appropriate codification, and that the codification be flexible enough 'to allow for modification in line with culture change'. Intellectualization here refers to the requirement of increasing accuracy along an ascending scale of functional dialects from conversational to scientific' (p. 521).

lengua culta es, en este sentido, el registro sociolingüístico de la lengua española en que *a)* predomina la función referencial sobre las otras funciones del lenguaje (según las definiciones clásicas de R. Jakobson), y *b)* se efectúan las comunicaciones lingüísticas de la mayor parte de los hispanohablantes educados.

La *lengua culta* entendida de esta manera, viene a ser más amplia que el español canonizado por las academias y, al mismo tiempo, consideramos que es la que constituye el punto de referencia en la apreciación de los hablantes a propósito de la "corrección idiomática"[25].

Con la *lengua culta* como punto de partida, elaboramos un modelo de los usos sociales del español mexicano, es decir, un modelo diastrático, que albergara todos los posibles registros en que se realiza el español mexicano. Para ello se propuso una hipótesis a partir de nuestro conocimiento de la comunidad lingüística mexicana y de lo que en otros diccionarios del español y otras lenguas significan etiquetas como *popular, elevado, coloquial,* etc., connotativas de la función sintomática del signo lingüístico.

Una condición básica para elaborar el modelo fue siempre la de prever posibles diferencias en el uso del vocabulario y establecer las cotas suficientes que permitieran realizar, durante el estudio estadístico, agrupaciones que el modelo mismo no supone pero que gracias a él son fácilmente identificables. El modelo del uso del español mexicano, por lo tanto, constituye una hipótesis de valor únicamente operativo.

3.2.1. La *lengua culta* corresponde al registro más alto de los usos del idioma y forma el marco de referencia necesario para el sentido de la corrección lingüística del hablante. Se trata aquí del nivel a partir del cual los diccionarios establecen las calificaciones de uso del léxico y generalmente, en cuanto registro, no aparece marcado de ninguna manera.

Paul L. Garvín ha propuesto en varias ocasiones el concepto de *lengua estándar* como sinónimo de lo que la Escuela de Praga denominaba *lengua literaria* o *lengua escrita*[26] (que nosotros hemos igualado con *lengua culta);* al aplicar estos conceptos a la práctica lexicográfica hemos creído necesario relacionarlos con los otros niveles

[25] GARVIN *(op. cit.,* p. 522) se refiere precisamente a la capacidad de la lengua culta de servir "as a frame of reference for correctness and for the perception and evaluation of poetic speech".

[26] La Escuela de Praga se refería indistintamente a la *langue littéraire* ("Thèses présentées au Premier Congrès des Philologues Slaves", p. 43) o a la *Schriftsprache* (B. HAVRÁNEK, "Zum Problem der Norm in der heutigen Sprachwissenschaft und Sprachkultur"), en J. Vachek (ed.), *A Prague school reader in linguistics,* Indiana University Press, 1964, pp. 33-58 y 413-420. respectivamente.

Cuadro A

Niveles de la lengua en la muestra del *DEM;* función predominante: referencial

Lengua	Nivel		Actualización
STANDARD 1. general (geogr.) 2. urbana (sociol.) 3. irradiadora	culta	a. vocabulario intelectualizado y rico b. sintaxis rica c. modelo de corrección	escrita
	sub-culta	a. vocabulario no intelectualizado b. sintaxis limitada c. desviación del modelo de corrección	
NO-STANDARD 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	hablada

de la lengua de manera tal que nuestras calificaciones de uso queden bien definidas con respecto a una visión global de los usos sociales del español mexicano.

En tales circunstancias preferimos distinguir entre *lengua estándar* y *lengua culta* haciendo más amplia a la primera y más restringida a la última. Y esto por una razón: creemos que en México hay un español uniforme en todo el país, resultado de la poderosa influencia no sólo de la educación, sino también de los medios masivos de información. Este *español estándar* se caracterizaría de la manera siguiente (véase cuadro A): es *general* en todas las regiones de México, es producto de lo que los antropólogos llaman "cultura urbana"[27] y se propaga continuamente a partir de los principales centros de irradiación del país (especialmente la ciudad de México). El nivel elevado del *español estándar* es la *lengua culta*, nivel de la literatura, de los textos científicos, de las conferencias, del periodismo, etc. Hay también un nivel del español mexicano estándar que se desvía de la lengua culta y es más familiar, más del dominio popular: lo llamamos *lengua sub-culta*. El español mexicano estándar cuenta, en consecuencia con dos niveles de uso por lo menos: el de la lengua culta y el de la lengua sub-culta.

Por contraposición con la lengua estándar, creemos que también existen usos del español poco extendidos, limitados a ciertas regiones geográficas (dialectos del español mexicano) o a ciertos grupos sociales cerrados (jergas del hampa, de algunas profesiones, etc.); en el caso de los dialectos geográficos generalmente supondríamos su relación con "culturas rurales". Tanto los dialectos como las jergas resultan generalmente poco capaces de irradiar sus características a grandes zonas del país. Ambos forman lo que denominamos español mexicano *no-estándar*.

3.2.2. Para mostrar con más claridad cómo han surgido las oposiciones entre diferentes niveles correspondientes a la lengua estándar y a la no-estándar, nos pareció conveniente desglosar las características que Garvín y Havránek atribuyen a la lengua culta y, a base de una serie de rasgos opositivos, definir el resto de los niveles; así tendríamos que:

1. La lengua culta se caracteriza por un vocabulario muy vasto y, sobre todo, intelectualizado; por una explotación muy amplia de las posibilidades sintácticas del sistema y por su capacidad de servir como modelo de corrección para los hablantes.

2. La lengua sub-culta, en contraposición con la anterior, no presenta gran intelectualización del vocabulario, tiende hacia la repetición de un número menor de vocablos y de patrones sintácticos

[27] Cf. ROBERT REDFIELD, *The folk culture of Yucatán*, Chicago, 1941, y GARVIN, *op. cit.*

de la lengua y, algo muy importante, se concibe como "desvío del modelo de corrección"[28].

3. En la lengua no-estándar, los dialectos del español mexicano tienen vocabularios amplios y característicos de cada zona, pero menos intelectualizados; posiblemente muestren una explotación sintáctica singular en cada caso y, por lo menos teóricamente (en el caso de México), pueden formar marcos de referencia para el sentido de la corrección lingüística de sus hablantes[29].

4. Con las jergas se trata ante todo de vocabularios reducidos y algunos clichés sintácticos, sujetos totalmente a las modas y, por ello, capaces sólo fugazmente de formar modelos de prestigio.

Debemos señalar que todas estas oposiciones (a excepción de los dialectos, que se tratarán en seguida) no son resultado de una oposición previa entre tipos de hablantes (por ejemplo: hablantes considerados "cultos" frente a hablantes analfabetos), sino que solamente se refieren a niveles de uso de la lengua, a registros que todos los hablantes pueden utilizar en diferentes situaciones. Un profesor universitario, por ejemplo, puede utilizar todos los niveles, aunque posiblemente tenga más dominio del primero (lengua culta) que de los demás. Esto se vuelve evidente en el caso de las jergas puesto que conviven con la lengua culta: un médico, por ejemplo, alterna el vocabulario intelectualizado con vocablos o expresiones propias de su gremio.

Tratándose de los dialectos la situación es más complicada, puesto que generalmente el hablante de un dialecto domina también la lengua estándar, pero en cambio difícilmente es capaz de utilizar un dialecto diferente del suyo. De ser así, estos hablantes se tendrían que considerar como "bidialectales" o "multidialectales".

3.2.3. Al traducir el modelo en clases o "géneros" de textos para la muestra, tenemos la siguiente división[30]:

LENGUA ESTÁNDAR:

LENGUA CULTA. **Literatura**: novela, cuento, ensayo, teatro. **Periodismo**: reportajes de autor mexicano, editoriales, reseñas

[28] Cf. JEAN et CLAUDE DUBOIS, *Introduction à la lexicographie: le dictionnaire*, Larousse, Paris, 1971, § 11.2, pp. 100-101.

[29] Dado el carácter de los dialectos del español mexicano, que no son sino modalidades del castellano trasplantado a América, nos parece difícil que el sentido de corrección de los hablantes de un dialecto se manifieste tan nítidamente como en otras comunidades de la Península ibérica; podría ser así, sin embargo, en los casos de dialectos muy caracterizados como el veracruzano o el del noreste (Monterrey, N. L.).

[30] La clasificación de las ciencias y las técnicas no tiene absolutamente ningún fundamento de orden epistemológico. Ha dependido, fundamentalmente, de nuestras necesidades prácticas de agrupar materias en grupos de tamaño relativamente homogéneo.

(políticas, sociales, culturales, deportivas, policiacas, taurinas).—**C**iencias: *humanas* (bibliotecología, filosofía, historia, culturas indígenas, pedagogía y educación, psicología); *sociales* (antropología, arqueología, derecho, economía, geografía, politología, sociología); *físico-matemáticas* (astronomía, electricidad y electrónica, física, geofísica, matemáticas, computación); *químico-biológicas* (biología, química, farmacología); *administrativas* (administración, contabilidad, comercio); *medicina* (humana, veterinaria); *arte* (arquitectura, danza, artes plásticas, artes gráficas, arte dramático, música, cine).—**T**écnicas: *ingeniería* (civil, industrial, química, automotriz, aeronáutica, naval, de ferrocarriles, de minas); *comunicación* (correo y filatelia, periodismo, publicidad y mercadotecnia, radio y televisión); *oficios* (agricultura, ganadería, pesca, carpintería, electricidad, etc.); *labores domésticas* (costura, cocina, decoración, etc.); *deportes* (charrería, tauromaquia, fútbol, etc.).—**O**tros: discursos políticos, religión, habla de la ciudad de México[31].

LENGUA SUB-CULTA. **L**iteratura popular: novela rosa, fotonovela, historietas.—**O**tros: conversaciones grabadas[32].

LENGUA NO-ESTÁNDAR: **T**extos regionales[33], documentos de estudios antropológicos, jergas, conversaciones grabadas.

3.2.4. Para que los autores de los textos fueran siempre distintos hemos vigilado que no se repitan más de dos veces en toda la muestra (en algunos casos excepcionales hay tres textos de un solo autor), y que en caso de repeticiones no sucedan dentro de un mismo "género".

3.2.5. Un problema especial se nos planteaba en el momento de asignar determinados porcentajes de importancia a cada género de la muestra. Inmediatamente nos dimos cuenta que no era posible tomar siempre el mismo número de textos de cada división, puesto que algunas estaban compuestas por textos cuyo léxico siempre es reducido (por ejemplo en la reseña deportiva o en algún texto cien-

[31] Estos textos pertenecen a las encuestas realizadas por el Centro de Lingüística Hispánica de la UNAM, publicadas bajo el titulo de *El habla de la Ciudad de México. Materiales para su estudio*, UNAM, México, 1971. Corresponden al nivel que, dentro del trabajo dialectológico del Centro, denominan como "informantes cultos".

[32] Se trata de encuestas realizadas por el Seminario de Dialectología de El Colegio de México sobre informantes de cultura "media". Las conversaciones que se citan en el nivel no-estándar corresponden, a su vez, a los informantes "analfabetas" de la Ciudad de México.

[33] Aprovechamos los materiales que ha reunido el Seminario de Dialectología de El Colegio de México bajo la dirección del Prof. Lope Blanch para la delimitación de las zonas dialectales de México. Cf. J. M. LOPE BLANCH, "Las zonas dialectales de México", *NRFH*, 19 (1970), 1-11.

tífico). Optamos mejor por asignar diferentes "pesos" a cada género de acuerdo siempre con los objetivos finales del *DEM*. Nuestra ponderación resultó de la siguiente manera:

Total de la muestra: 100%		*Porcentajes por géneros*	
LENGUA CULTA:	66.80%		100%
		Literatura	22.45
		Periodismo	26.34
		Ciencia	26.94
		Técnica	15.26
		Discurso político	2.69
		Religión	1.79
		Habla de la Ciudad de México	4.49
LENGUA SUB-CULTA:	11.70%		100%
		Literatura popular	53.00
		Conversaciones grabadas	47.00
LENGUA NO-ESTÁNDAR:	21.50%		100%
		Textos regionales	60.46
		Documentos de antropólogos	15.34
		Jergas	13.95
		Conversaciones grabadas	10.25

3.3. Los géneros de textos que han quedado establecidos a partir del modelo del § 3.2.2 se reparten entre el total de 2,000,000 de ocurrencias que fijamos previamente. Explicamos ahora cómo delimitamos cada texto. Como hemos dicho (ver § 2.2 y 2.4*c*), una de las causas principales de distorsión de la muestra estadística es la influencia del estilo individual de los autores; por lo que no es conveniente tomar para nuestro corpus un conjunto de libros completos (el hecho de que no intentemos hacer ni un diccionario literario ni un diccionario exhaustivo nos permite desechar esta tendencia generalizada por la tradición lexicográfica). La cuestión es entonces determinar el tamaño de los textos y la manera de seleccionarlos.

3.3.1. En cuanto al tamaño de los textos nos ha parecido conveniente aplicar la misma fórmula utilizada en el estudio de H. Kučera y W. N. Francis sobre el inglés norteamericano contemporáneo[34], pues consideramos que, desde el punto de vista de las relaciones de frecuencia entre los vocablos, el español presentará características semejantes a las del inglés, ambas lenguas cultas con-

[34] HENRY KUČERA y W. NELSON FRANCIS, *Computational Analysis of Present day American English*, Providence, R. I., 1970.

temporáneas. Un texto en nuestra muestra quedará formado, por lo tanto, por 2,000 ocurrencias extraídas de las obras que constituyen las fuentes para la muestra. Nuestro corpus queda como un conjunto de 1,000 textos con 2,000 palabras gráficas cada uno. Resumismos la lista presentada en § 3.2.3 indicando el número de textos en cada división:

LENGUA CULTA: 668; Literatura: 150; Periodismo: 176; Ciencias: 180; Técnicas: 102; Otros: 30.
LENGUA SUB-CULTA: 117; Literatura popular: 62; Conversaciones grabadas: 55.
LENGUA NO-ESTÁNDAR: 215; Textos regionales: 130; Documentos de estudios antropológicos: 33; Jergas: 30; Conversaciones grabadas: 22.

3.3.2. En cuanto al modo de extraer los textos de las obras fuente es necesario tomar en consideración tanto los resultados que se desea obtener como también la clase de datos que requiere el trabajo lexicográfico.

Como se ha dicho anteriormente no sólo nos interesa obtener del análisis de nuestro corpus la lista alfabética de los vocablos incluidos en él, sino que también deseamos conocer los contextos en que aparece utilizada cada palabra. Si el objetivo fuera solamente el primero, lo más sencillo sería utilizar un muestreo aleatorio simple a todo lo largo de la obra fuente, es decir, se seleccionarían ocurrencias aisladas, y tendrían todas la misma probabilidad de aparecer en nuestro texto. Pero este procedimiento no conduce a la obtención de contextos y no permite obtener las selecciones tan rápidamente como lo desearíamos. Si, en cambio, tomamos las palabras con su contexto, la selección viene a ser continua y se aleja, en cierta medida, del ideal aleatorio. Dentro de un texto las palabras se reúnen en torno a una misma idea y de este modo se condicionan unas a otras. En muchos casos este mutuo condicionamiento es causa de repeticiones de vocablos, lo cual aumenta las frecuencias respecto del ideal estadístico. No obstante hemos preferido este último procedimiento tanto por las necesidades que se asientan antes, como por la facilidad que nos representa la selección de pasajes continuos.

Para definir la extensión de los pasajes que forman el texto convenimos en llamar *palabra* a la unidad tipográfica que aparece entre dos blancos y en llamar *párrafo* al conjunto de unidades tipográficas que aparecen entre dos puntos aparte. La identificación de estos elementos no ofrece ningún problema y, en cambio, nos permite obtener sentidos completos para las palabras que queramos analizar. Nuestra unidad de muestreo viene a ser el *párrafo* y un

texto tendrá tantos párrafos como se necesite para alcanzar la extensión de 2,000 ocurrencias.

3.3.3. La selección de párrafos para cada texto se realiza con un esquema aleatorio: mediante una tabla de números al azar hecha especialmente para nuestro trabajo, se efectúan las siguientes operaciones: *1)* se elige una página de libro o de revista (o un lugar dentro de una cinta magnética)[35]; *2)* se selecciona un párrafo de esa página (o cinta); *3)* se seleccionan tantos párrafos como sea necesario para lograr la cifra de 2,000 palabras gráficas; *4)* el proceso se repite cuantas veces sea necesario.

Hay casos en que no se encuentra un párrafo en la página escogida y se vuelve necesario tomar un nuevo número de página; hay otros en que el párrafo no comienza al principio de la página, por lo que el seleccionador toma el primero que cumpla con esa condición. También hay párrafos que terminan en la página siguiente y entonces el pasaje se tiene que extender hasta más allá de los límites de la página inicial.

4. Una vez realizadas las operaciones de selección de textos y de alimentación de datos a la computadora electrónica, iniciamos el análisis estadístico de los materiales aplicando las siguientes estimaciones y medidas estadísticas (véase cuadro B, en que se presentan en forma simbólica y con el formato numérico de salida de la computadora):

G_j caracteriza al j-ésimo género, en que j varía de 1 a m.

4.1. La clasificación de géneros se hará en dos versiones. En la primera se toma $m = 3$:

$G_1 =$ lengua culta
$G_2 =$ lengua sub-culta
$G_3 =$ lengua no-estándar

En la segunda clasificación tenemos $m = 13$:

$G_1 =$ literatura	$G_6 =$ religión	$G_{11} =$ docs. antropológicos
$G_2 =$ periodismo	$G_7 =$ habla Cd. México	
$G_3 =$ ciencias	$G_8 =$ literatura popular	$G_{12} =$ jergas
$G_4 =$ técnicas	$G_9 =$ conv. subculta	$G_{13} =$ conversacs. no-est.
$G_5 =$ discurso político	$G_{10} =$ textos regionales	

[35] En el caso de la transcripción de textos hablados se ha establecido previamente una tabla de convenciones de transcripción que no solamente permite aprovechar el material para el *DEM*, sino en general para cualquier trabajo de tipo dialectológico que no sea exclusivamente fonético.

4.2. Cada vocablo[36] identificado dentro del corpus aparecerá listado una sola vez junto con todas las frecuencias estimadas y las medidas estadísticas que le correspondan. Estas listas constituirán el núcleo de lo que podríamos llamar nuestro "diccionario estadístico del español de México".

Si se representa como V_i al *i*-ésimo vocablo de la muestra, se tendrá una lista de vocablos, esto es: $i = 1, 2, \ldots, \Omega$, donde Ω es el total de vocablos distintos en el corpus. La ordenación de los V_i no será arbitraria, sino que se ajustará a los siguientes requisitos: *a)* una ordenación que tome en cuenta la frecuencia total de ocurrencias del vocablo, en sentido decreciente. Las palabras con frecuencia idéntica se agruparán alfabéticamente; *b)* una ordenación estrictamente alfabética.

Estas dos clasificaciones y las dos mencionadas por géneros dan lugar a cuatro tabulaciones básicas que producirá la computadora. La ordenación por frecuencias nos permitirá identificar el vocabulario más usual —estadísticamente hablando. La ordenación alfabética nos permitirá conocer las características estadísticas de cualquier vocablo (un vocablo que no aparezca dentro del corpus, es decir, $t = 0$ tendrá precisamente esa característica estadística[37]).

4.3. Las frecuencias absolutas cuentan el número de ocurrencias de cada vocablo dentro del corpus. De esta manera definimos como t_i al número de ocurrencias del vocablo V_i dentro del total; para contar la frecuencia de los vocablos en cada género tomamos como f_{ij} al número de ocurrencias del vocablo V_i dentro del *j*-ésimo género.

Las frecuencias relativas miden el porcentaje de ocurrencias de cada vocablo respecto al total dentro de cada género y entre los géneros, de acuerdo con las siguientes relaciones:

$$h_i = \frac{100\, t_i}{\sum_{i=1}^{\Omega} t_i} = \text{porcentaje de ocurrencia de } V_i.$$

$$e_{ij} = \frac{100\, f_{ij}}{t_i} = \text{porcentaje de ocurrencia entre géneros del vocablo } i \text{ en el género } j.$$

$$d_{ij} = \frac{100\, f_{ij}}{\sum_{i=1}^{\Omega} f_{ij}} = \text{porcentaje de ocurrencia dentro de géneros del vocablo } i \text{ en el género } j.$$

[36] La obtención de *vocablos* según la definición que hemos seguido a lo largo de este trabajo (cf. *supra*, nota 9) supone que los problemas de lematización del corpus han sido resueltos en una etapa anterior.

[37] Una discusión a propósito de la frecuencia cero en estadística lingüística se puede encontrar en K. HEGER, *Belegbarkeit...*, pp. 26-27.

CUADRO B

Núm. de orden	*Palabra*	*Frecuencias absolutas*					*Frecuencias relativas*								*Medidas estadísticas*		
							Respecto al total	*Entre géneros*				*Dentro de géneros*						
		Total	G_1	G_2	...	G_m		G_1	G_2	...	G_m	G_1	G_2	...	G_m	*KF'*	*S*	*C*
1	P_1	t_1	f_{11}	f_{12}	...	f_{1m}	h_1	e_{11}	e_{12}	...	e_{1m}	d_{11}	d_{12}	...	d_{1m}	KF_1	S_1	C_1
2	P_2	t_2	f_{21}	f_{22}	...	f_{2m}	h_2	e_{21}	e_{22}	...	e_{2m}	d_{21}	d_{22}	...	d_{2m}	KF_2	S_2	C_2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
i	P_i	t_i	f_{i1}	f_{i2}	...	f_{im}	h_i	e_{i1}	e_{i2}	...	e_{im}	d_{i1}	d_{i2}	...	d_{im}	KF_i	S_i	C_i
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ω	P_Ω	t_Ω	$f_{\Omega 1}$	$f_{\Omega 2}$	...	$f_{\Omega m}$	h_Ω	$e_{\Omega 1}$	$e_{\Omega 2}$	...	$e_{\Omega m}$	$d_{\Omega 1}$	$d_{\Omega 2}$	...	$d_{\Omega m}$	KF_Ω	S_Ω	C_Ω

4.4. La mera frecuencia total, ya sea absoluta o relativa, es en realidad una medida lingüística poco útil. Dos vocablos con la misma frecuencia, pero con distinta distribución entre géneros, tienen distinto comportamiento dentro del idioma: una distribución irregular entre los géneros denota que el vocablo está sujeto a determinantes de estilo o de tema, mientras que una distribución regular —en todos los géneros— indica que la palabra es independiente de la clasificación por géneros y que por lo tanto es mayor su importancia y su utilidad dentro del idioma. Se busca entonces una medida que combine tanto la frecuencia del vocablo como su *dispersión* entre géneros, de modo que entre vocablos de parecida frecuencia total se favorezcan aquéllos con dispersión más uniforme.

Luego de analizar distintas medidas que se han propuesto y usado en varios diccionarios de frecuencias, para reflejar la situación planteada, adoptamos como la más adecuada la creada por J. Lanke[38], quien, tras sopesar distintas cualidades y defectos de las medidas que hasta ahora se han utilizado, llega a preferir el índice de frecuencia y dispersión entre géneros llamado KF (por *Korrigierte Frequenz)*, que guarda la siguiente definición para el *i*-ésimo vocablo (de acuerdo con la notación adoptada en nuestro trabajo):

$$\mathrm{KF}_i = \frac{1}{100}\left(\sum_{j=1}^{m} \sqrt{\mathrm{d}_{ij}\,\mathrm{f}_{ij}}\right)^2$$

Para propósitos de exploración lingüística, también calcularemos un índice que, independientemente de la frecuencia, nos dé indicaciones de cómo se dispersa un vocablo entre géneros. El mismo J. Lanke propone la medida:

$$S_i = \frac{KF_i}{\mathrm{t}_i}$$

Pero este índice presenta la dificultad de que, si se aplica a una muestra en que los géneros que la componen son desiguales —como en el caso presente—, su variación depende directamente de las desigualdades de la muestra y va desde la menor frecuencia relativa observada en el caso de la distribución más desigual, hasta 1 cuando la distribución es uniforme. Para corregir esta dificultad hemos optado por la siguiente modificación, a la que llamamos índice normalizado de dispersión:

$$\mathrm{C}_i = \frac{100\,\mathrm{S}_i - \min_j \mathrm{d}_{ij}}{\min_j \mathrm{d}_{ij}}$$

[38] Según reporta I. Rosengren, "The quantitative concept of language and its relation to the structure of frequency dictionaries", *ELA*, 1 (1971), 103-127.

El índice normalizado de dispersión entre géneros queda ahora establecido en un rango que varía entre 0 y 1 donde 0 indica la distribución más desigual y 1 la más uniforme.

5. Esperamos haber mostrado a lo largo de este trabajo que el estudio estadístico del léxico del español mexicano representa la mejor posibilidad de obtener una imagen real de la lengua que se utiliza en México y que, a pesar de las limitaciones que nos hemos ido imponiendo en el transcurso de nuestro proyecto, es el procedimiento más útil y adecuado para el problema que enfrentamos. Creemos haber tomado todas las precauciones necesarias para que el rendimiento del corpus sea suficiente y verdaderamente constituya el fundamento para la selección de vocablos que compongan el *DEM*. Sobre esta selección final (posterior al estudio en sí) deseamos agregar que el tema esencial que se propone a todo el equipo lexicográfico es el de saber cuáles vocablos deben integrar el diccionario y cuáles pueden desecharse. Dado que nuestro diccionario no es autoritario, el criterio de "uso de los buenos escritores", o los más conocidos del purismo y del casticismo no pueden constituir el punto de partida para la formación de la macroestructura del *DEM*. Como hemos querido demostrar en los párrafos iniciales de este trabajo, los léxicos de mexicanismos que se pueden encontrar hoy en día no pueden tampoco servir como base para nuestra selección de voces. Mediante el análisis estadístico en frecuencias, dispersión y los dos índices corregidos (KF y C) tendremos una base necesaria para toda consideración del vocabulario por incluir en el *DEM*. No obstante, por las razones que hemos apuntado, será necesario tener un criterio complementario de inclusión para vocablos de baja frecuencia, cuyo análisis estadístico no produzca resultados definitivos, y para vocablos que no hayan aparecido en la muestra y sin embargo sean de importancia para las finalidades del *DEM*.

En estos dos casos, la decisión dependerá del juicio intersubjetivo del cuerpo de redacción del *DEM* formado por el equipo lexicográfico, el consejo de redacción y el consejo consultivo[39]. Con

[39] El equipo lexicográfico, núcleo principal del *DEM*, lo forman: Elisabeth Beniers, Luz Fernández, Lourdes Gavaldón, Paulette Levy, Ángeles Soler y Carmen Delia Valadez. El consejo de redacción, cuya finalidad es revisar el trabajo del equipo, lo forman: Antonio Alatorre, Margit Frenk Alatorre, Raúl Ávila, Jaime García Terrés, Beatriz Garza Cuarón, Andrés Henestrosa, Juan M. Lope Blanch, Carlos Monsiváis, Augusto Monterroso y José Emilio Pacheco. El consejo consultivo está formado por cerca de cincuenta científicos mexicanos de diferentes especialidades. El trabajo de computación electrónica se efectúa sobre una máquina Univac 1106 del Centro de Procesamiento y Evaluación de la Secretaría de Educación Pública, gracias al interés de su director Enrique Calderón; el equipo de matemáticos lo constituyen: Isabel García Hidalgo, Manuel Orona y Jorge Serrano.

objeto de no falsear los resultados obtenidos en el estudio, las voces incluidas por esta otra vía recibirán una marca especial, aunque no hemos decidido si esto se hará en el cuerpo mismo del *DEM* o por separado.

LUIS FERNANDO LARA
ROBERTO HAM CHANDE

El Colegio de México.