LENGUA HABLADA Y ESTRATO SOCIAL: UN ACERCAMIENTO LEXICOESTADÍSTICO*

Introducción

El léxico es un acervo que normalmente se incrementa a lo largo de toda la vida de un individuo. Un vocabulario extenso supone —así sea potencialmente, sin considerar entre otros factores los de orden psicológico y social y la capacidad para usarlo eficientemente— una posibilidad mayor para comprender y expresarse. En otros términos, un vocabulario más rico implica un universo conceptual mayor y, en este sentido, una mayor cultura —sin discriminación de la comunidad que la produzca—, en la medida en que las palabras son portadoras de conocimientos y los generan.

A partir de lo anterior y del interés que pueda tener la evaluación estadística del vocabulario de la lengua hablada, en esta investigación exploro algunas posibilidades para determinar cuantitativamente el léxico del español hablado en México por varios grupos sociales. Me propongo establecer las diferencias entre ellos con base en tres índices a los cuales me referiré *in extenso* más adelante. Para esto me apoyo en un *corpus* léxico que se recogió de grabaciones realizadas en todo el país con informantes de uno y otro sexo, de distinta edad y condición social, y nativos de diversas localidades.

* Este artículo fue leído en versión menos extensa como ponencia en el VIII Congreso Internacional de la Asociación de Lingüística y Filología de la América Latina, San Miguel de Tucumán, Argentina, 7 al 11 de septiembre de 1987.

Textos y corpus léxico

El léxico se obtuvo de 205 textos que fueron seleccionados por el grupo de investigadores del Diccionario del Español de México tomando como fuente otras tantas entrevistas grabadas que se hicieron para el proyecto de delimitación de las zonas dialectales del país[1] y para el estudio del habla de la ciudad de México[2]. Los textos fueron posteriormente procesados por computadora[3] e incorporados al *corpus* del DEM[4]. Más adelante, tras revisar el *corpus* mencionado en los archivos de computadora, seleccioné del

Tabla 1
Número y sexo de los informantes por texto

Número de infs.	*Sexo*			
	Mase	*Fem*	*M y F**	*Total*
uno	83	82		165
dos	11	15	13	39
tres			1	1
Total	94	97	14	205

* Entrevistas con informantes de distinto sexo.

[1] Las entrevistas fueron realizadas por el Seminario de Dialectología del Centro de Estudios Lingüísticos y Literarios de El Colegio de México. Cf., para esto, J. M. Lope Blanch, "Las zonas dialectales de México: proyecto de delimitación", *NRFH*, 19 (1970), 1-11.

[2] Fueron hechas por el Centro de Lingüística Hispánica de la Universidad Nacional Autónoma de México y se publicaron en *El habla de la ciudad de México. Materiales para su estudio,* UNAM, México, 1971.

[3] El trabajo de computación se llevó a cabo en el Centro de Procesamiento de Datos Arturo Rosenblueth de la Secretaría de Educación Pública. Para la delimitación, la revisión y el análisis de mi *corpus* conté con la ayuda de Alejandro Medel y Héctor Vázquez, de la institución mencionada. También recibí el apoyo técnico y la asesoría de la Unidad de Cómputo de El Colegio de México.

[4] Véase L. F. Lara y R. Ham Chande, "Base estadística del Diccionario del Español de México", en L. F. Lara, R. Ham Chande e I. García Hidalgo, *Investigaciones lingüísticas en lexicografía*, El Colegio de México, México, 1979, pp. 27 *ss.* Tomé también los informantes de cultura media y baja del *corpus* del DEM. Como señalan Lara y Ham Chande, su unidad de muestreo seleccionada aleatoriamente es "el párrafo, y un texto tendrá tantos párrafos como se necesite para alcanzar la extensión de aproximadamente 2 000 ocurrencias" (p. 31). El art. cit. apareció originalmente en *NRFH*, 23 (1974), 245-267.

mismo los textos que forman la base y la unidad estadística de mi investigación[5].

Como puede verse en la tabla 1, las entrevistas fueron realizadas en su mayor parte con un solo informante (80.5%); hubo también un buen número de ellas con dos (19%) y una sola con tres personas (0.5%). En cuanto al sexo de los participantes, si se consideran los textos que fueron producidos por una o dos personas del mismo sexo, hubo un porcentaje bastante similar de hombres (45.9%) y de mujeres (47.3%). Fueron menos (6.8%) las entrevistas en las que participaron dos o tres informantes de distinto sexo.

En lo referente al estrato social o nivel cultural[6] de los informantes (tabla 2), los porcentajes fueron, para el nivel alto, 22.9%; para el medio, 31.7%; y para el bajo, 45.4%. Con los datos recogidos en las entrevistas organicé siete grupos, de acuerdo con la edad de los participantes, con la finalidad de advertir cómo esta-

TABLA 2
Edad y nivel cultural de los informantes por texto

	Edad en años							
Nivel	*17-21*	*22-29*	*30-39*	*40-49*	*50-59*	*60-69*	*70-*	*Total*
Alto	0	10	8	14	7	4	4	47
Medio	13	16	14	12	6	4	0	65
Bajo	12	19	21	17	15	6	3	93
Total	25	45	43	43	28	14	7	205

Nota: En los casos de dos o tres informantes se obtuvo la edad promedio.

[5] Tras formar el archivo mediante la selección de textos exclusivamente de lengua hablada, se revisaron los datos procesados por computadora y se reclasificaron los informantes de acuerdo con las características que menciono *infra*. Se recogió el texto completo —alrededor de 2 000 palabras, como indiqué en la nota anterior— de cada una de las entrevistas. Cabe destacar, por otra parte, que los textos del DEM fueron codificados, en este caso, por entrevista y no se delimitaron las intervenciones de cada uno de los informantes. Por ese motivo, cuando me refiero a informantes debe entenderse que los considero unitariamente por cada texto para fines estadísticos.

[6] El nivel cultural fue determinado a partir de características tales como la escolaridad y la ocupación. Utilizo tanto esa expresión —empleada por quienes diseñaron las investigaciones originales— como *estrato* o *nivel social* con el mismo sentido. Personalmente realicé algunas grabaciones de entrevistas para el Atlas Lingüístico del Español de México. Cf. notas 1 y 2 para referencias bibliográficas.

ba constituida la muestra en relación con esta variable[7]. Porcentualmente, los grupos de edad van desde un mínimo de 3.4% para el grupo de 70 o más años hasta el máximo de 21.9% para el grupo de 22 a 29 años. La zona —última variable de la muestra— incluyó, por una parte, la ciudad de México (43.4%) y por otra, las entrevistas hechas en diferentes localidades del país (56.6%)[8].

Del total de los textos procesados por computadora se formó un *corpus* de 428 899 palabras-ocurrencia o palabras gráficas, excluyendo los nombres propios y los números escritos con guarismos, los cuales no fueron registrados en los datos estadísticos que ofrezco más adelante[9]. De ese *corpus* se obtuvo, de nuevo mediante un programa de cómputo, un total de 23 504 palabras diferentes o *tipos* léxicos[10]. Más adelante, después de revisar la lista de tipos, los asociamos a los vocablos correspondientes[11] y, tras ser

[7] Para las entrevistas en las que intervenían dos o más informantes consideré la edad promedio de los mismos. En ningún caso la diferencia de edad de los participantes fue de más de diez años.

[8] Las localidades se distribuyeron prácticamente por todos los estados de la República, de acuerdo con una delimitación provisional de las zonas dialectales de México que propuso J. M. LOPE BLANCH en su art. "El léxico de la zona maya en el marco de la dialectología mexicana", *NRFH*, 20 (1971), mapa 27 y pp. 55 *ss*. Esta delimitación fue tomada en cuenta por quienes determinaron las fuentes —en este caso de lengua hablada— del DEM. Las zonas y las localidades aparecen descritas en el documento de uso interno *Manual de información para los miembros del Consejo Consultivo del DEM*. Decidí, de acuerdo con mis fines, oponer únicamente dos zonas —la ciudad de México y la provincia— porque las entrevistas que se hicieron en el interior del país y con las cuales se formó el archivo computado iban desde un mínimo de 4 hasta un máximo de 9 por localidad. Ese número de datos resulta normalmente insuficiente para obtener resultados estadísticos confiables.

[9] La extensión de mi *corpus* resulta bastante adecuada para los fines que persigo, si se compara con el que emplearon A. JUILLAND y E. CHANG-RODRÍGUEZ. Su *Frequency dictionary of Spanish words*, de acuerdo con la descripción de WILLIAM TAYLOR PATTERSON, *The lexical structure of Spanish, with special consideration for the genealogical and chronological properties*, tesis doctoral, Stanford University, 1967, p. 3, "consists of the 5 000 most frequently used words in a scientifically selected corpus of 500 000 words. These 5 000 basic words account for 97% of the occurrences of any representative text of the Spanish language". Véase *infra* (tabla 6) el número de vocablos que obtuve para los últimos dos deciles (90 y 100%) de mi *corpus*.

[10] La delimitación del conjunto de textos de lengua hablada y la determinación de los tipos fueron hechas también en el Centro de Procesamiento de Datos Arturo Rosenblueth. Los programas correspondientes fueron diseñados asimismo por A. Medel y H. Vázquez.

[11] La asociación se hizo en forma manual, de acuerdo con los criterios que se deducen del *Diccionario de la lengua española* de la Real Academia Española,

procesada esta información electrónicamente, se obtuvo un total de 9 309 vocablos.

Índices de riqueza léxica

Es evidente que la cantidad de vocablos o de tipos que se obtengan de un texto estará en relación con la extensión del mismo[12]. De acuerdo con lo anterior, se puede inferir que si un texto de extensión menor produce un mayor número de unidades léxicas que otro de extensión mayor, el primero tiene mayor riqueza que el segundo[13]. Esto sucede si se comparan los vocablos que obtuve para los niveles medio y bajo, donde el primero resulta más rico pues tuvo un mayor número de vocablos con un conjunto de textos de extensión menor que el segundo. En cambio, no se puede llegar a una conclusión segura cuando se compara un conjunto de textos de menor longitud con otro de mayor tamaño y el primero tiene un menor número de vocablos que el segundo, como ocurre con los que se recogieron para los niveles alto y medio (véase tabla 6).

Los casos como el anterior han hecho necesario emplear otros métodos para poder comparar textos de diferente longitud y evaluar su riqueza léxica[14]. En esta investigación —como he dicho— utilizo tres índices para determinar las diferencias léxicas de las variables de la muestra. El primero, la densidad léxica, se

---

20ª ed., Espasa-Calpe, Madrid, 1984. Dado que no nos fue posible recurrir a contextos, en los casos de homonimias que pudimos detectar optamos por asignar vocablos diferentes al mismo tipo léxico. Para toda esta labor conté con la ayuda de Sara Giambruno del Centro de Estudios Lingüísticos y Literarios de El Colegio de México.

[12] Cf. Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique,* Presses Universitaires, Paris, 1960, p. 84: "plus un texte est long plus il comporte de mots différents". Véase también Charles Muller, *Estadística lingüística*, trad. A. Quilis, Credos, Madrid, 1973, pp. 267 *ss.*

[13] Para decirlo en términos de Muller (*op. cit.*, p. 270), "se puede sistematizar este método de comparación, considerando dos textos A y B, llamando Na, Nb sus extensiones respectivas, Va, Vb la extensión de sus vocabularios. Se puede decir que el vocabulario de A es más rico que el de B si se tiene:

bien $Na \leq Nb$ y $Va > Vb$

bien $Na < Nb$ y $Va = Vb$".

[14] Véanse por ej., Guiraud, *op. cit.*, pp. 85 *ss.*, Muller, *op. cit.*, pp. 269 *ss.* y Gustav Herdan, *The advanced theory of language choice and chance,* Springer, Berlin-Heidelberg-New York, 1966, pp. 72 *ss.*

basa en la evaluación individual de cada uno de los textos. El segundo, las frecuencias acumuladas por deciles, se aplica a un conjunto de textos de un estrato social determinado. Mediante el tercero comparo el número de vocablos que se obtienen en segmentos extensos de igual longitud, los cuales se forman de nuevo a partir de un conjunto de textos.

Densidad léxica

Este índice resulta de la división del número de tipos léxicos T que se obtienen de un segmento de texto de una longitud determinada entre el número N de palabras del segmento[15]. Expresa-

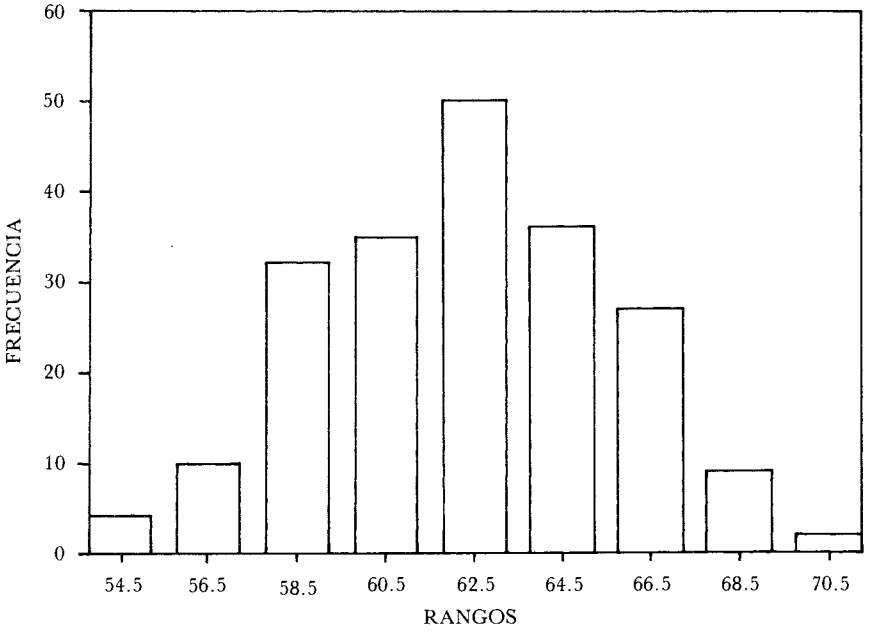
Tabla 3
Densidad léxica: ordenación por rangos

Rango	*Frecuencia*	*Porcentaje*
54.5	4	2.0
56.5	10	4.9
58.5	32	15.6
60.5	35	17.1
62.5	50	24.4
64.5	36	17.6
66.5	27	13.2
68.5	9	4.4
70.5	2	1.0
Total	205	100.0

Prom: 62.3 - Mediana y Moda: 62.5 - Desv est: 3.35 - Var: 11.2

[15] El procedimiento ha sido utilizado, entre otros, por P. L. Baldi, "Fattori sociali dell'abilitá linguistica nella produzione scritta di bambini di novedieci anni", *SILTA*, 1(1974), pp. 335-471. Su índice, sin embargo, difiere del mío en la medida en que él considera sólo las palabras de contenido y no las de función, como yo hago. Véase también la investigación de J. Ure, quien utiliza el mismo índice (sólo con palabras de contenido) para el inglés (*apud* M. A. K. Halliday, *Language as social semiotics,* E. Arnold, London, 1978, p. 32). De acuerdo con los comentarios de Halliday (*ibid.*) sobre el trabajo de Ure, parece evidente que la densidad está en relación con el medio (la lengua escrita tiene una densidad más alta que la hablada) y, dentro de éste, con la función social del lenguaje. En este sentido, parto de la suposición de que los textos que analizo fueron grabados en una misma situación comunicativa: la entrevista. Se refiere también a la densidad léxica Humberto López Morales, *La enseñanza de la lengua materna,* Madrid, 1984, pp. 56 y 91. Yo mismo he utilizado anteriormente el procedimiento para evaluar textos escritos por

do en otros términos, D = T ÷ N. Para una longitud en palabras de N = 100, que es la utilizada en esta investigación, la densidad de un texto es el promedio de las densidades de las unidades de 100 palabras que contiene el texto. Para el total de la muestra se obtuvo una densidad promedio mínima de 54 y una máxima de 71 (18 valores).



Gráfica 1. Densidad léxica: ordenación por rangos

Como he indicado *supra*, la densidad es un valor que se obtiene de cada texto considerado individualmente. Por lo mismo, es posible observar el comportamiento de la muestra en su conjunto con apoyo en esa variable. Para ese propósito, y con el fin de poder hacer un mayor número de observaciones, hice una reagrupación de acuerdo con un primer nivel de rangos, asignando a cada uno de ellos el promedio de densidad de cada dos valores. De esta manera obtuve las frecuencias que aparecen en la tabla 3 y que se ilustran en la gráfica 1. Como se desprende de la tabla mencionada, el comportamiento de la muestra para la característica que investigo se acerca bastante a una curva normal, lo que

niños: véase mi art. "Léxico infantil de México: palabras, tipos, vocablos", *ACIEA* (2), pp. 512 *ss*.

permite suponer *a posteriori* que los informantes fueron bien seleccionados[16].

Para decidir cuáles de las variables se correlacionaban con la densidad, reagrupé los datos en un segundo nivel de rangos: el inferior, el central y el superior[17]. Estos tres valores se tomaron como variable lingüística independiente para, a partir de ella, considerar cuál o cuáles de las demás variables eran significativas en cuanto a su posibilidad de explicar la mayor o menor densidad de los textos[18]. En otros términos, se trató de ver qué variables dependían de la densidad.

De acuerdo con los resultados, las variables que menos explican la mayor o menor densidad son la edad, el sexo y el número de informantes; y las que más, la zona y el nivel cultural[19]. En cuanto al primer grupo de variables, no obstante las pocas diferencias observadas, se pueden comentar algunos aspectos, así sea en un plan especulativo a partir de los indicios que ofrecen los datos. Si se observan los grupos de edad[20], puede pensarse que se sigue aprendiendo léxico, aunque relativamente poco, a lo largo de toda la vida adulta, pues la densidad aumenta conforme aumenta la edad[21]. En lo que toca a los grupos por sexo y por

[16] Para obtener los datos estadísticos por computadora se utilizó el *Statistical Package for Social Sciences,* releáse 1.1, versión para PC. El programa fue aplicado por Javier Rodríguez de la Unidad de Cómputo de El Colegio de México, quien también me ayudó a interpretar los resultados.

[17] Abarcaron, respectivamente, las densidades 54 a 59, 60 a 65 y 66 a 71. En otros términos, cada uno de los rangos de este segundo nivel corresponde al promedio de cada tres rangos del primero y a 6 promedios de densidad.

[18] Para esto se utilizó la prueba de x cuadrada. De acuerdo con ella, se parte de la suposición, por ejemplo, de que el nivel cultural y la densidad son independientes. Si los resultados contradicen esta llamada "hipótesis nula", se comprueba que las variables están correlacionadas en mayor o menor grado de acuerdo con el mayor o menor valor de x^2. Véanse estos valores para las variables zona y nivel en las tablas 4 y 5 respectivamente. La significación (*sign* en las tablas) se refiere a la hipótesis nula y a la probabilidad de que se confirme.

[19] Se confirmó que la zona y el nivel cultural eran las variables de mayor peso en relación con la densidad incluso mediante desagregaciones de datos. Por ejemplo, se consideró separadamente cada grupo de la variable sexo y su densidad promedio por nivel cultural y se vio que las diferencias significativas aparecían en los niveles.

[20] Baso mis observaciones en los datos que obtuve mediante la reagrupación de los grupos de edad en tres rangos: de 17 a 29, de 30 a 49, y de 50 o más años.

[21] En cambio, si se comparan estos resultados con los que obtuve para niños (art. cit., p. 514), las diferencias son muy significativas: de una densidad

número de informantes, tal parece que se emplea más léxico cuando en un diálogo intervienen hombres y mujeres.

En relación con las variables que sí resultaron tener una alta correlación con la densidad, puede destacarse lo siguiente. En cuanto a las zonas (tabla 4), el porcentaje de textos del rango superior (RS) para la ciudad de México (68.4%) es 36.8% más alto que el de las otras localidades (31.6%). En cambio, en los rangos central (RC) e inferior (RI), los porcentajes correspondientes son más altos en la provincia (RC = 54.5%, RI = 82.6%) que en la ciudad de México (RC = 45.5%, RI = 17.4%): hay 9.0% más textos en RC y 65.2% más en RI en el interior del país que en la capital. Por otra parte, si se analiza cada zona independientemente, se advierte que en la ciudad de México es mayor el porcentaje de textos de RS (29.2%) que de RI (9.0%), mientras que en las otras localidades sucede lo contrario: son más los de RI (32.8%) que los de RS (10.3%).

TABLA 4
Densidad léxica: zona y rangos superior, central e inferior

		Zona			
		Otras l.	*Cd. Méx.*	*Total*	*Gran tot.*
Rango	Superior	12	26	38	38
		31.6	68.4	100%	18.5%
		10.3	29.2		
	Central	66	55	121	121
		54.5	45.5	100%	59.0%
		56.9	61.8		
	Inferior	38	8	46	46
		82.6	17.4	100%	22.4%
		32.8	9.0		
	Total	116	89		
	Porcnt.	100%	100%		
	Gran tot.	116	89		205
	Porcnt.	56.6%	43.4%		100.0%

$X^2 = 22.56$ · Sign = 0.0000

promedio de 47 (edad promedio de 9 años) se pasa a otra de 62.3 (ambas para los respectivos *corpus*). Esto permitiría confirmar la intuición de que el aprendizaje —en este caso del léxico— es muy alto entre la niñez y la edad adulta.

La comparación de los niveles culturales (tabla 5) muestra asimismo diferencias importantes y claramente significativas. Si se consideran de nuevo cada uno de los rangos, del 100% de textos del RS, 47.4% provienen del nivel alto, 36.8% del medio y 15.8% del bajo. Entre los niveles extremos alto y bajo hay una diferencia de 31.6% más textos en el primero que en el segundo. Por otra parte, los textos del nivel alto se reparten en 38.3% para RS, 53.2% para RC y 8.5% para RI. En el nivel bajo, en cambio, se invierten las proporciones para las categorías extremas: 6.5% de los textos aparecen en RS, 54.8% en RC y 38.7% en RI. La mayor diferencia se presenta en RI entre los niveles alto (8.7%) y bajo (78.3%): el segundo grupo produjo 69.6% más textos de este rango que el primero.

TABLA 5

Densidad léxica: nivel cultural y rangos superior, central e inferior

		Nivel cultural				
		Alto	*Medio*	*Bajo*	*Total*	*Gran tot.*
Rango	Superior	18	14	6	38	38
		47.4	36.8	15.8	100%	18.5%
		38.3	21.5	6.5		
	Central	25	45	51	121	121
		20.7	37.2	42.1	100%	59.0%
		53.2	69.2	54.8		
	Inferior	4	6	36	46	46
		8.7	13.0	78.3	100%	22.4%
		8.5	9.2	38.7		
	Total	47	65	93		
	Porcnt.	100%	100%	100%		
	Gran tot.	47	65	93		205
	Porcnt.	22.9%	31.7%	45.4%		100.0%

$x^2 = 39.33$ - Sign = 0.0000

FRECUENCIAS ACUMULADAS POR DECILES

Como he comentado *supra*, el índice de densidad proviene de cada texto considerado individualmente. En ellos el léxico se presenta en relación proporcional a su mayor o menor frecuencia. En otras palabras, en un texto se reflejan tanto las palabras de

frecuencia alta como las de media y las de baja, según sus respectivas probabilidades. En este sentido, la densidad evalúa el léxico que llamaré *normal*. Frente a lo anterior, para el análisis de frecuencias acumuladas[22] me he basado en conjuntos de textos de diferente longitud. No obstante esto, el procedimiento permite evaluar comparativamente el vocabulario de los conjuntos, ya que la extensión de cada uno de ellos —sus respectivas frecuencias totales— no condiciona el número de vocablos que se obtienen en determinados deciles, como el 7° (70% de frecuencias) y el 8° (80%). Por otra parte, dado que se requiere ordenar los vocablos en orden descendente de frecuencias[23], las diferencias corresponderían, en los deciles antes mencionados, a los vocablos de frecuencias altas y medias.

En el análisis por deciles y de aquí en adelante presento únicamente los datos relacionados con los niveles culturales ya que es la variable que mejor explica las diferencias léxicas[24]. Como puede observarse en la tabla 6 y en la gráfica 2, las diferencias se empiezan a notar a partir del 4° decil y se van ampliando hasta llegar al máximo en el 7°. En éste se recogieron para el nivel alto, cuyos textos sumaron 110 565 palabras, 250 vocablos; para el medio, 219 de un total de 130 735 palabras; y para el bajo, 183 vocablos de una frecuencia de 187 599. Estos resultados muestran, como dije antes, que los vocablos que se obtienen en ese decil —en este caso para los niveles culturales— no están condicionados por la extensión de los textos: el que tuvo la frecuencia más alta produjo un número menor de vocablos, y el de la más baja, un número mayor.

[22] Un procedimiento semejante al que utilizo para los deciles de una distribución fue utilizado por R. HAM CHANDE, quien hizo su segmentación por cuartiles: véase su art. "Del 1 al 100 en lexicografía" (en L. F. LARA, R. HAM CHANDE e I. GARCÍA HIDALGO, *op. cit.*, pp. 78 *ss.*). Creo, sin embargo, a partir de sus datos (cf. su cuadro, p. 76 y su cuadro 6, p. 81), que sus resultados se basan en tipos léxicos y no en vocablos. No obstante, se acercan bastante a los míos.

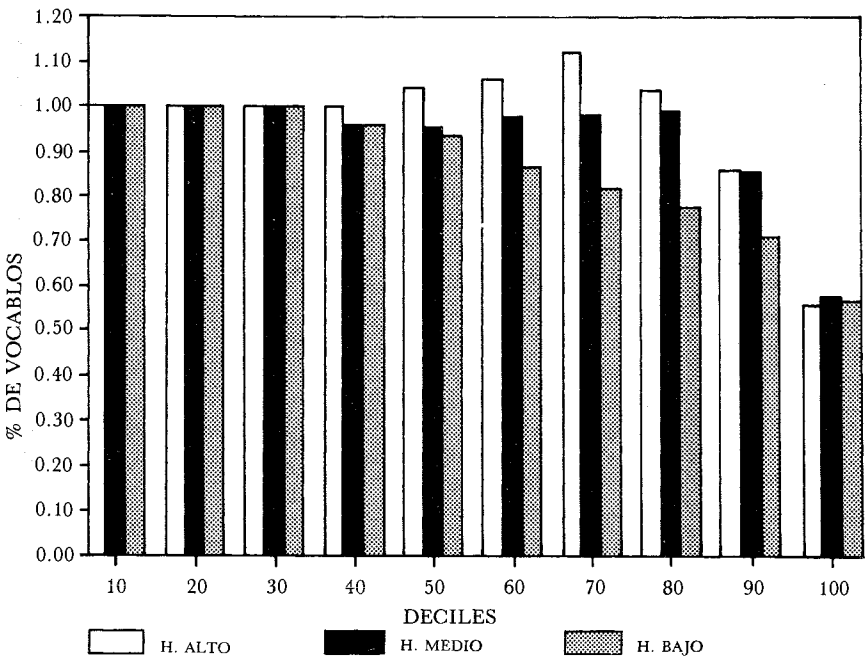
[23] Es decir, la lista de vocablos se inicia con el de frecuencia más alta (V_n) y termina con los de más baja. Como puede verse en la tabla 6, para cubrir el 10% (42 889) de frecuencias del *corpus* son suficientes los tres vocablos de más alta frecuencia ($V_n + V_{n-1} + V_{n-2}$).

[24] Me apoyo en los resultados que se obtuvieron para las zonas y los niveles culturales considerados como variables métricas. En cuanto a la primera, el promedio de densidad para la ciudad de México fue de 63.7 (= 100%), y para las otras localidades de 61.2 (−3.9%). En los niveles culturales, los promedios fueron de 64.3 (= 100%) para el alto, 63.2 (−1.7%) para el medio, y 60.6 (−5.8%) para el bajo.

Si se toma como base de comparación el nivel alto, para el cual se obtuvieron 250 vocablos (= 100%) en el 7° decil, las diferencias porcentuales muestran que tiene 12.4% más vocablos

TABLA 6
Frecuencias acumuladas por deciles: número de vocablos según nivel cultural

	Deciles (%)										
Nivel	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*Frecs.*
Alto	3	7	13	25	48	103	250	708	2 407	5 195	110 565
Medio	3	7	13	24	44	95	219	674	2 387	5 388	130 735
Bajo	3	7	13	24	43	84	183	530	1 992	5 322	187 599
Corpus	3	7	13	25	46	97	223	681	2 793	9 309	428 899



Gráfica 2. Frecuencias ordenadas por deciles (corpus = 100%)

que el nivel medio, y 26.8% más que el nivel bajo. En el 8° decil el nivel alto produjo 708 vocablos (= 100%), lo que representa 4.8% más vocablos que el medio y 25.1% más que el bajo. Estas

diferencias, como he dicho, corresponderían a vocablos de frecuencias altas y medias. En el 10° decil, en cambio, podría considerarse que las diferencias, por una parte, se deben al peso de los vocablos de frecuencias bajas; y por otra, están condicionadas por la longitud de los textos. No obstante esto último, el nivel medio, con una extensión menor que el bajo, produjo más vocablos que éste.

Vocablos en segmentos extensos de igual longitud

Para evaluar la riqueza léxica de textos o conjuntos de textos de distinta longitud se puede emplear un procedimiento obvio: comparar los conjuntos de textos hasta el límite de la máxima extensión común. En otros términos, si tres conjuntos A, B, C, tuvieron respectivamente las frecuencias n, n + 1 y n + 2, el límite de la comparación será n.

De acuerdo con lo anterior, si se agrupan los textos de la muestra considerando únicamente sus características de densidad en los rangos superior, central e inferior, y se toma como límite el de 80 000 palabras gráficas —el máximo del rango superior, donde hubo menos textos—, se obtienen los resultados que aparecen en la tabla 7 y que se ilustran en la gráfica 3. La tabla constata que el número de vocablos está en función de la extensión del texto[25] y que conforme aumenta la longitud decrece el número de vocablos nuevos —de baja frecuencia— que se obtienen[26]. De allí la forma asintótica de las curvas que aparecen en la gráfica[27].

[25] Como señala Muller (*op. cit.*, p. 267), "Está claro que V [el total de vocablos] es función de N [el número de palabras o la frecuencia], es decir, que para un texto dado, V crece con N [...]. Ciertamente, es evidente que V crece menos de prisa que N, puesto que cada palabra que representa un vocablo ya utilizado en el texto infiere una unidad de retraso a V con relación a N". Véase también *supra*, nota 12 y texto.

[26] El vocabulario real presenta en la tabla, para la extensión que considero, algunos casos en los cuales un aumento de frecuencias produce más vocablos —y no menos— que el aumento anterior. Este tipo de desviaciones, si se considera su similitud con las que muestra Muller (*op. cit.*, pp. 296-297), no invalida el planteamiento general de que conforme aumenta la extensión disminuye el número de nuevos vocablos. Véase además la curva que presenta el autor citado (*ibid.*) y en la cual aparecen las extensiones teórica y real del vocabulario.

[27] Para decirlo de nuevo con Muller (*op. cit.*, p. 268), "V no cesa de crecer, pues ningún texto agota el léxico de su autor; habrá que aceptar este pos-

Si se compara el número de vocablos que se obtuvieron en la frecuencia 70 000 (4 085 para el *corpus*) con los que se recogieron en la frecuencia 80 000 (*corpus*: 4 401), la diferencia (316) correspondería al incremento de nuevos vocablos entre los dos segmentos. Esos vocablos son —si no de manera absoluta, al menos en forma relativa— de baja frecuencia en comparación con las frecuencias del vocabulario total, dado que se recogieron en el límite superior de la extensión del texto.

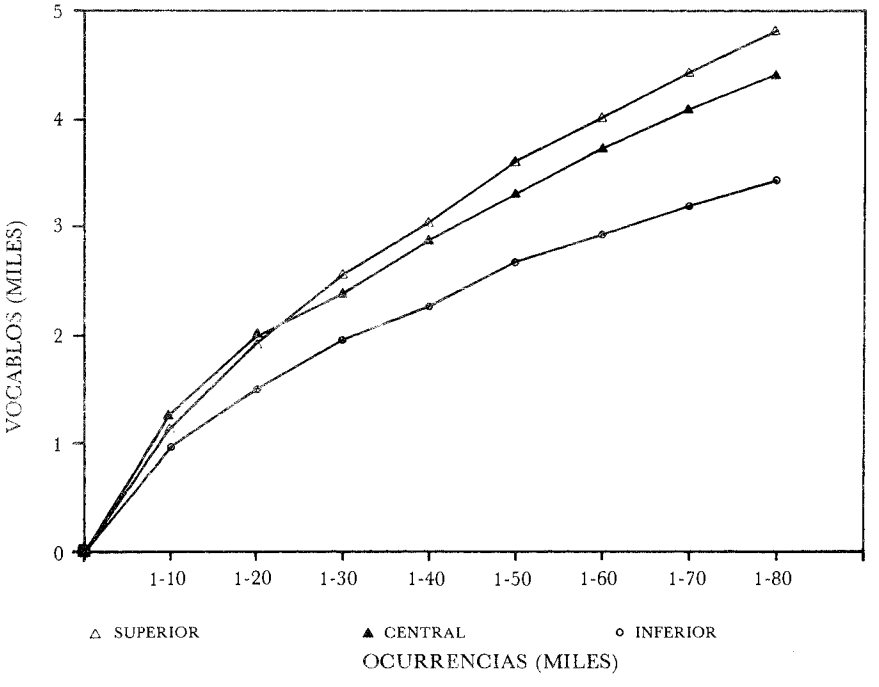
TABLA 7
Densidad léxica: rangos superior, central e inferior

Número de vocablos según frecuencias (en miles)				
	Rangos			
Frec.	*RS*	*RC*	*RI*	*Corpus*
1-10	1 127	1 265	965	1 257
1-20	1 918	1 965	1 492	1 915
1-30	2 553	2 387	1 951	2 369
1-40	3 039	2 868	2 266	2 896
1-50	3 597	3 307	2 667	3 187
1-60	4 014	3 719	2 936	3 645
1-70	4 438	4 096	3 198	4 085
1-80	4 825	4 417	3 428	4 401
Dens. pro. (UT = 100)	68.5	62.5	56.5	62.3

En cuanto a los grupos de textos por nivel cultural, la frecuencia límite fue de 110 000, de acuerdo con la extensión máxima de los correspondientes al conjunto de nivel alto. Para fines comparativos, en la tabla 8 aparecen además las frecuencias 1 a 90 000 y 1 a 100 000, junto con los vocablos que se obtuvieron en cada segmento para los tres estratos. De acuerdo con esos datos, en los

tulado, que es una verdad de experiencia tanto como una evidencia lingüística, y que sólo podrá ser puesta en discusión para un corpus de dimensiones inconcebibles: todavía podemos dudar de ello. El resultado es que nuestra curva tendrá partes planas cada vez más largas [. . .] se obtendrá una línea que tiende a llegar a ser paralela al eje de abscisas, pero sin que jamás llegue a serlo completamente". Véase un planteamiento similar, así como una curva sensiblemente semejante a las que presento en R. M. FRUMKINA, "The application of statistical methods in linguistic research", en O. S. AKHMANOVA, I. A. MEL'CHUK, R. M. FRUMKINA & E. V. PADUCHEVA, *Exact methods in linguistic research*, University of California Press, Berkeley-Los Angeles, 1963, pp. 102-103.

tres rangos de frecuencias se obtuvieron más vocablos en el nivel alto que en el medio, y más en éste que en el bajo. Si se toma el número de vocablos de nivel alto como base de comparación, se advierte que las diferencias entre éste y los otros niveles aumentan conforme se incrementa la frecuencia: en el primer rango el nivel alto (4 630 = 100%) produjo 0.8% más vocablos que el medio, y 16.8% más que el bajo; en el segundo rango (4 928 = 100%), 1.8% más que el medio, y 18.4% más que el bajo; y en el tercer rango (5 209 = 100%), 3.5% más que el medio, y 20.1% más vocablos que el nivel cultural bajo. Esto permite considerar que conforme crezca la longitud crecerán también las diferencias —explicables por el incremento de los vocablos de baja frecuencia— entre los niveles culturales[28].



Gráfica 3. Vocablos por rangos

[28] Es interesante comparar los resultados que obtuve con los adultos y con los niños, pues permite de nuevo (véase lo que mencioné respecto a la densidad en mi nota 21) advertir el alto porcentaje de vocablos (48%) que se adquieren entre una y otra edad. Los vocablos correspondientes al *corpus* de ambos grupos en la extensión de 1-110 000 ocurrencias fueron, para los niños, 3 563 (= 100%); y para los adultos, 5 272 (148%). Cf., en relación con los datos del vocabulario infantil, mi art. cit., p. 511.

Consideraciones finales

Los resultados que he obtenido me permiten hacer tres tipos de consideraciones. La primera se relaciona con los datos estadísticos. De acuerdo con ellos, las diferencias entre los niveles sociales

Tabla 8
Niveles culturales alto, medio y bajo

	Número y porcentaje de vocablos según frecuencias (en miles)					
	Niveles					
	alto		*medio*		*bajo*	
frec.	*voc.*	*%*	*voc.*	*%*	*voc.*	*%*
1-90	4 630	100.0	4 591	99.2	3 850	83.2
1-100	4 928	100.0	4 838	98.2	4 021	81.6
1-110	5 209	100.0	5 027	96.5	4 160	79.9

se presentan, en relación con las medidas de densidad, frecuencias acumuladas por deciles y número de vocablos por frecuencias, en el léxico normal, en los vocablos de frecuencias altas y medias y en los de bajas respectivamente[29]. En cuanto a estos últimos, se puede estimar que conforme se extiende la longitud del texto, las diferencias se acentúan. Por otra parte, el tamaño y el comportamiento de la muestra junto con el tipo de evaluaciones que se utilizaron dan confiabilidad a los resultados obtenidos.

La segunda consideración se refiere a las características lingüísticas que he estudiado. A diferencia de otro tipo de investigaciones o encuestas, el análisis del léxico y de otros componentes del lenguaje mediante grabaciones tiene la ventaja de que el informante, aunque quisiera, difícilmente podría reaccionar y cambiar su conducta lingüística frente al investigador[30]. No se da esa reacción precisamente por el nivel de inconsciencia que tienen los hablantes respecto al sistema de la lengua. Consecuentemente, este

[29] Es importante recordar que, como he mostrado *supra* (véase p. ej. la tabla 5), el mayor o menor acervo léxico —aunque es más frecuente en el nivel alto que en el bajo— no es exclusivo de un estrato social, ya que en cualquiera de ellos pueden encontrarse individuos de una u otra características.

[30] Esto se refuerza por el hecho de que yo mismo no suponía que iba a hacer este tipo de estudio cuando realicé algunas de las entrevistas y, obviamente, los informantes tampoco. Se podría argumentar que es posible que el entrevistado cambie de registro ante el investigador. Sin embargo, como señalé antes (nota 15), las grabaciones se realizaron en una misma situación comu-

tipo de datos lingüísticos resulta altamente confiable para la caracterización de los sujetos investigados. Por otra parte, es necesario destacar que mis resultados no están condicionados por aspectos connotativos del lenguaje. El proceso de computación no distingue, por ejemplo, *fuiste, fuistes* o *juites*: las tres formas se consideran palabras gráficas y son elementos igualmente válidos para la estadística.

Por último, quisiera mencionar algunos aspectos sociales del lenguaje. Se ha discutido extensamente sobre la diferencia o la deficiencia de los códigos lingüísticos que utilizan los hablantes de diferentes estratos[31]. Además de los resultados que ahora he ofrecido, he encontrado una situación similar en los niños[32]. Por eso no parece pertinente volver a argumentar sobre el hecho de que en un estrato social se utilice más léxico que en otro. En cambio, habría que buscar las causas de esas diferencias: muy probablemente tienen que ver con la escolaridad, pero también con el tipo de actividad o de trabajo de las personas. Las funciones del lenguaje en relación con la actividad son, necesariamente, distintas y esto podría explicar las diferencias. El lenguaje para la acción —frente al especulativo que privilegia la función heurística— es precisamente el que tiene menor densidad[33]. Esto permite rechazar la hipótesis del déficit: el lenguaje es adecuado para los

---

nicativa. Esa situación podría, precisamente, condicionar el registro, y no al contrario. De acuerdo con los planteamientos de HALLIDAY (*op. cit.*, pp. 31 *ss.*), el registro es una forma de predicción. Si se conocen los factores que intervienen en la comunicación y el escenario en que éste ocurre, "we can predict a great deal about the language that will occur, with reasonable probability of being right". Para los conceptos de factores y escenarios en los actos de habla, véase D. HYMES, "The ethnography of speaking", en J. A. Fishman (ed.), *Readings in the sociology of language*, Mouton, The Hague, 1970, pp. 110 *ss.*

[31] Me refiero a las conocidas tesis de Bernstein y a quienes las apoyan y las discuten. He comentado esto en mi art. "La langue espagnole et son enseignement: oppresseurs et opprimés", en Jacques Maurais (ed.), *La crise des langues,* Conseil de la langue française-Le Robert, Québec-Paris, 1985, pp. 342 *ss.* Véase además una muy buena condensación de esta discusión en F. WILLIAMS, "Some preliminaries and prospects", en Fredrick Williams (ed.), *Language and poverty,* 5th ed., Rand McNally, Chicago, 1973, pp. 1-10.

[32] Cf. mi art. "Léxico infantil de México...", p. 513. Véase también, para conclusiones que apoyan las tesis de Bernstein a partir del estudio de niños italianos, BALDI, art. cit., pp. 376 y 377.

[33] Cf. HALLIDAY, *op. cit.*, p. 32: "pragmatic language, or «language of action», has the lowest density of all. This is probably true of all languages [...]". Cf. además mi nota 15.

fines del usuario y es diferente justamente por eso. Además, el poseer un léxico extenso no es una condición suficiente para usarlo adecuadamente, con eficiencia comunicativa. Para todos es evidente que en ciertos grupos sociales de alta escolaridad se abusa de esa característica no precisamente para comunicarse sino para buscar *status* mediante el procedimiento de impresionar a los demás a través de redundancias y verborrea[34]. En cambio, las personas con menores recursos léxicos pueden ser más eficientes en su expresión: basta recordar a los excelentes narradores que aparecen por todos los pueblos[35] y todos los barrios perdidos de las ciudades.

Las personas que tenemos educación universitaria compartimos y utilizamos recursos lingüísticos similares e incluso normas muy semejantes aunque seamos de distintos países. Nos parecemos porque nos comunicamos. Esta idea puede explicar el que haya pocas diferencias entre personas de sexo y de edades diferentes: hay comunicación entre ellas. Los que parecen no hablarse son los grupos de distinto nivel social —y no hace falta decir ahora en qué pocos casos sí se dirigen la palabra. Si tiene sentido acortar las diferencias entre ellos y nosotros, el camino sería volvernos sus interlocutores y —entre otros aspectos— devolverles la información que de ellos obtuvimos cuando fueron nuestros informantes.

RAÚL ÁVILA
El Colegio de México

[34] Cf., a propósito de esto, los comentarios de W. LABOV, "The logic of nonstandard English" en F. WILLIAMS, *op. cit.*, p. 164: "in many ways working-class speakers are more effective narrators, reasoners, and debaters than middle-class speakers who temporize, qualify, and lose their arguments in a mass of irrelevant detail".

[35] Vale la pena recordar lo que dijo el reconocido escritor Agustín Yáñez en un coloquio: "Mis principales maestros del idioma fueron mi madre, que conservó siempre el idioma campesino de sus primeros años, y la sagacidad de los arrieros que durante mi niñez nos transportaban en largas jornadas por los campos de Jalisco".