Análisis de consultas al *Diccionario del español de México* en línea

Alfonso Medina Urrea*

Resumen. En este trabajo se examinan las frecuencias y dispersiones geográfica y temporal de las consultas hechas al *Diccionario del español de México* entre agosto de 2012 y septiembre de 2013. Esta información se usa para calcular un índice que permite seleccionar aquellas que son más usuales en más lugares y por más tiempo. Con este índice se puede, por un lado, evaluar la base estadística que dio lugar a la nomenclatura actual del diccionario y, por el otro, seleccionar aquellas consultas sin respuesta en el sistema que sean los mejores candidatos para formar parte de futuras nomenclaturas del diccionario.

Palabras clave. *Diccionario del español de México*, adquisición léxica, frecuencia de consulta, dispersión geográfica de consultas, dispersión temporal de consultas.

Introducción

no de los temas más visibles del proyecto del *Diccionario del español de México* (*DEM*) ha sido el de cómo determinar la nomenclatura de un diccionario, esto es, la lista de lemas que se definen en él. Al tratarse de un diccionario integral, el *DEM* busca documentar las voces más usuales del español en México y no solamente las más coloridas y peculiares de nuestra región. En otras palabras, lo importante no ha sido compilar la lista de los vocablos que nos parecen más característicos del español mexicano, como si se tratara de un diccionario de mexicanismos, sino determi-

* Profesor-investigador del Centro de Estudios Lingüísticos y Literarios (CELL) del COLMEX. Miembro del equipo de investigación responsable del Diccionario del español de México (*DEM*), del CELL-COLMEX. Dirección electrónica: amedinau@colmex.mx

nar cuáles son los que más usamos día a día, que naturalmente tienden a ser los que compartimos con otros hablantes de la lengua española.

Un diccionario integral de nuestro español satisface varias necesidades. Dos de las más importantes son: *1)* documentar el léxico que usamos hoy en día, para contribuir a la educación y comunicación en nuestro país y *2)* afirmar nuestra herencia lingüística, especialmente nuestro patrimonio léxico, que no es prestado y que en mucho compartimos con el resto de los hispanoparlantes. Así, un diccionario integral, además de permitirnos conocer lo particular y peculiar de nuestro dialecto, nos posibilita ponderar lo frecuente, común y tradicional y, por lo tanto, más difundido con el mundo de habla hispana.

En este contexto, el equipo lexicográfico del DEM se apoyó desde un primer momento en criterios estadísticos, los que se le aplicaron al *Corpus del español mexicano contemporáneo* (CEMC)[1] para conformar una base estadística, que nos permite asegurar que los vocablos que forman parte de la nomenclatura del diccionario son los más representativos del español usado en México, en diversos géneros textuales y registros dialectales (Lara *et al*. 1980: 29-39). En su última versión, el diccionario incluye los vocablos con frecuencia de tres o más[2] en el CEMC. Específicamente, tiene una nomenclatura de aproximadamente 25 mil vocablos, que engloban cerca de 50 mil acepciones; y se puede confiar, gracias a esta base estadística, en que las palabras del diccionario han sido usadas en México en los siglos XX y XXI.

Ciertamente, el tamaño de la nomenclatura del diccionario deberá crecer todavía más para satisfacer las necesidades educativas y comunicativas de los hispanoparlantes mexicanos. Por eso, el equipo del DEM trabaja nuevos artículos lexicográficos, correspondientes a las palabras con frecuencia dos o menos en el CEMC, y se ha embarcado en la compilación de un nuevo *corpus*, el $CEMC_2$, que servirá para constituir una nueva base estadística, que permitirá determinar qué vocablos adicionales, muchos muy nuevos, podrán incluirse en la futura nomenclatura.

[1] Se trata de un *corpus* electrónico de cerca de dos millones de palabras que se puede consultar en http://cemc.colmex.mx.

[2] No se trata de la frecuencia absoluta de las palabras en el *corpus*. Más bien es una frecuencia ajustada para tomar en cuenta su dispersión en los diversos géneros del *corpus* (periodismo, literatura, etcétera). De esta manera, aquellos que ocurren en más géneros tienen una mayor frecuencia y aquellos que ocurren en menos géneros tienen una menor.

El ejercicio de encontrar piezas léxicas para compilar diccionarios en papel o electrónicos o lexicones para otras tecnologías es conocido como adquisición léxica (*lexical acquisition*). Existen técnicas para llenar lagunas en la nomenclatura de diccionarios que se basan en los patrones de ocurrencia de las palabras en *corpus* electrónicos. De hecho, la adquisición léxica automática ahorra recursos y es cada vez más viable, dados los avances en el procesamiento del lenguaje natural (PLN), aprendizaje de máquinas y compilación de *corpus* electrónicos.[3]

La primera versión en línea del *DEM* fue la del *Diccionario del español usual en México* (*DEUM*) y apareció en el servidor http://mezcal.colmex.mx, que ya dejó de existir. La última versión se encuentra en http://dem.colmex.mx. El hecho de tener el diccionario en la Web permite mucho más que difundir su contenido. Por un lado, es posible conocer qué información consultan y obtienen los usuarios del *DEM* en México, en los países de habla hispana y en el resto del mundo. Por otro lado, se puede analizar lo que los usuarios buscan en el diccionario que todavía no forma parte de su nomenclatura. Además, como se apuntó arriba, el monitoreo de estas consultas puede ayudar a determinar qué vocablos faltan y podrían considerarse prioritarios para el equipo lexicográfico.

En este trabajo, se analizan las consultas que se han hecho al *DEM* en línea. En esencia, se examinan sus frecuencias y sus dispersiones geográfica y temporal para asignarle a cada una un índice de frecuencia y dispersión que permite seleccionar aquellas que son más usuales en más lugares y por más tiempo. La idea es utilizar este índice para ordenar las consultas de más frecuentes y dispersas a menos frecuentes y menos dispersas. Con esta información se puede, por un lado, evaluar el análisis estadístico que dio lugar a la nomenclatura actual y, por el otro, seleccionar las consultas que constituyen los mejores candidatos para formar parte de la futura nomenclatura del diccionario.

[3] En otras palabras, es posible aplicar estos métodos para orientar al lexicógrafo en cuanto a qué vocablos deben ser considerados para futuras ediciones de un diccionario. Manning y Schütze caracterizan estos métodos como "algorithms and statistical techniques for filling holes in existing machine-readable dictionaries by looking at the occurrence patterns of words" (1999: 265). Así, el monitoreo de estas consultas puede ayudar a determinar qué vocablos faltan y deben ser prioridad del equipo lexicográfico.

Las consultas al *DEM*

La última versión del *DEM* se puso en línea en agosto de 2012. Desde esa fecha, cada consulta queda registrada en el sistema, con información sobre el país de origen, la dirección IP (*Internet Protocol*) del equipo desde donde se hizo,[4] la fecha, hora y el resultado de dicha consulta; esto es, si la petición de información fue o no encontrada en la nomenclatura o en alguna definición o ejemplo del diccionario.

En este trabajo se examinan 13 meses de consultas, de agosto de 2012 a septiembre de 2013. Después de eliminar consultas repetidas y consecutivas, originadas desde una misma dirección IP, se llevaron a cabo en este intervalo de tiempo un total de 180,164 búsquedas, las que corresponden a 16,505 búsquedas distintas (por ejemplo, el vocablo *casa* se buscó 101 veces pero cuenta como una búsqueda distinta o tipo de consulta).

Lo interesante es que, detrás de cada consulta, hay una persona interesada en alguna pieza léxica del español mexicano contemporáneo. Si bien no contamos con datos de corte sociolingüístico que nos permitan caracterizar a los usuarios, es posible analizar dichas consultas mediante la observación de sus frecuencias (qué tanto se repiten) y sus dispersiones geográfica (desde cuántos países) y temporal (cómo se repartieron en el tiempo durante los 13 meses de la muestra).

Frecuencia de las consultas

Como se sugirió arriba, muchas peticiones de búsqueda se repiten; pero muchas ocurren sólo una vez. Así que, de las 180,164 búsquedas mencionadas antes, se identificaron 62,343 tipos de consultas, con frecuencias absolutas que van de la mayor, con 244 ocurrencias, a las consultas únicas que suman más de la mitad del total (37,399). Para este análisis, la información recabada se presenta en tablas cuyos renglones

[4] Una dirección del protocolo de Internet o IP es un número binario de 32 bits que identifica de manera precisa un dispositivo electrónico particular conectado a Internet. Se representa en cuatro números decimales separados mediante puntos, como por ejemplo 172.16.35.22, que corresponde a una computadora de El Colegio de México (COLMEX), ya que 172.16 es el prefijo asignado a esa institución.

contienen consultas ordenadas de mayor a menor frecuencia, dispersión o índice de éstos. Por ejemplo, en la Tabla 1, se muestran en la segunda columna las veinte consultas o peticiones de búsqueda más frecuentes,[5] en la tercera columna sus frecuencias durante el periodo examinado y, en la primera, el rango de ordenamiento. Específicamente, las consultas están ordenadas de mayor a menor frecuencia (esto es, a menor rango en la tabla, mayor frecuencia y a mayor rango, menor frecuencia):

Tabla 1. Las consultas más frecuentes

Rango	Petición de búsqueda	Frecuencia
1	chido	244
2	chale	199
3	pinche	197
4	chingar	191
5	naco	183
6	pendejo	152
7	chilango	145
8	güey	135
9	guey	121
10	gacho	117
11	coger	116
12	madre	115
13	verga	114
14	puto	112
15	bato	109
16	gato	109
17	chingada	105
18	perro	105
19	casa	101
20	cuate	98

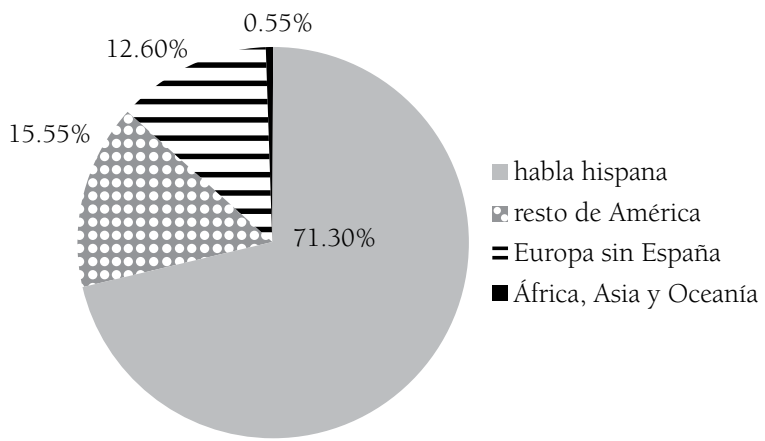
[5] Nótese que en ésta y todas las tablas las consultas se transcriben tal y como fueron hechas, por lo que podrán aparecer con y sin diacríticos (*guey* y *güey*, *cabron* y *cabrón*, etcétera). Lo importante es que en este trabajo se consideran consultas distintas.

Vale la pena examinar el carácter de estas 20 consultas más frecuentes. Muchas pertenecen a los registros de uso coloquial y popular del *DEM* (*chido*, *chale*, *pinche*, etcétera), otras son de uso muy general (*madre*, *gato*, *perro*, *casa*). Es interesante que la petición de búsqueda *guey* (9) pueda corresponder tanto a *güey* (8) como a *gay* (que no aparece entre las 20 más frecuentes pero tiene frecuencia 19), la primera por omisión de la diéresis, la segunda por similitud de pronunciación.

Dispersión geográfica

En cuanto a la procedencia de las consultas, en la Gráfica 1 se observan las proporciones de las búsquedas según la región del mundo desde donde se originaron. Más del 70% se hizo en algún país de habla hispana. Menos del 1 por ciento (0.55%) se originó en África, Asia o alguna de las islas de los océanos Pacífico e Índico. La segunda región desde donde se hicieron más consultas es la de la América no hispanoparlante (incluido Estados Unidos, que tiene una población importante de mexicanos) y la tercera fue la de Europa sin España:

Gráfica 1. Procedencia geográfica de las consultas



En la Tabla 2, se aprecian los porcentajes de consultas por país. De un total de 91 países, en la tabla aparecen 10. La proporción de Estados Unidos se explica en gran medida por la población de origen mexicano, pero seguramente también por el interés comercial entre México y ese país. Por otra parte, resulta interesante que Alemania tenga una proporción mayor a la de España.

Tabla 2. Países donde se originaron las consultas

País	Cantidad de consultas	Porcentaje
México	117116	65.01%
Estados Unidos	23973	13.31%
Alemania	10317	5.73%
España	6988	3.88%
Francia	6625	3.68%
Brasil	3006	1.67%
Italia	2422	1.34%
Argentina	1617	0.90%
Canadá	961	0.53%
Colombia	689	0.38%
… otros 81 países	6432	3.57%

Aunque la afinidad cultural entre México y estos países explique que se hagan peticiones de búsqueda desde ellos, en realidad los porcentajes son más bien bajos (con excepción de las consultas hechas desde Estados Unidos). Sin embargo, es muy significativo que las consultas que más se repiten se hagan desde varios países, lo cual se puede apreciar en la Tabla 3, que muestra el número de países desde donde se hizo cada consulta:

Tabla 3. Las consultas hechas desde más países

Petición de búsqueda	Número de países
pinche	27
chingar	27
chingada	20
güey	18
chido	17
coger	16
pendejo	15
madre	15
bato	14
perro	14
chale	13
naco	13
guey	13
verga	13
gacho	12
chilango	11
gato	10
casa	10
puto	6

Es evidente que con la información que contamos sobre las peticiones de búsqueda (ip, país, fecha y hora) es difícil, si no imposible, hacer una caracterización de los usuarios del *dem*. Al ser Estados Unidos el segundo país desde donde se hacen más consultas, cabe imaginar que muchas son de emigrantes mexicanos y de las personas que se relacionan con ellos. También es posible que se trate de estudiantes mexicanos, turistas, aprendices y estudiosos de la lengua y cultura de México en ese país o en Alemania, que tiene el tercer puesto. Además, hay una gran necesidad e interés de intérpretes y traductores en estos y otros países; así como de lingüistas y lexicógrafos, de obtener información sobre nuestro léxico.

Dispersión temporal

Otra manera de observar la variabilidad de las consultas es considerar su profusión a lo largo del tiempo. En esencia, se trata de determinar cuáles peticiones de información son las más constantes, esto es, cuáles no pasan de moda. Cabe mencionar que las fechas y horas registradas para cada petición de información están representadas mediante números no negativos. Es decir, ambas, la fecha y la hora de una consulta, están codificadas en un número. Por ejemplo, la primera consulta al *DEM* en línea se llevó a cabo el 1 de agosto de 2012 a las 11:12:55 horas, lo cual quedó representado con el número 41122.4673076042. La distancia temporal entre dos consultas será el valor absoluto de la resta entre los números que las representan; esto es, la diferencia entre el número mayor (la fecha y hora más tardías) y el número menor (la fecha y hora más tempranas). El número que resulta representa el intervalo de tiempo entre estas dos peticiones de información. Por ejemplo, la última consulta del periodo examinado se hizo el 30 de septiembre de 2012 a las 23:52:20 horas (lo que se representa con el número 41547.9946728009). Así que el intervalo entre la primera y última de las consultas se obtiene restando el primer número del segundo (41547.9946728009 — 41122.4673076042): 425.5273651967, que se redondea a 426. En la tabla 4 se muestran las 20 consultas más repartidas en el tiempo. La segunda columna contiene la distancia temporal entre la primera y la última vez que se buscaron las expresiones de la primera columna. De nuevo, aparecen ordenadas del mayor al menor tamaño de intervalo:

Tabla 4. Las consultas más dispersas en el tiempo

Petición de búsqueda	Intervalo
casa	421
coger	411
chale	408
chido	408
naco	406
chingada	403
guey	401
perro	400
chingar	400
verga	400
chilango	399
madre	398
pinche	398
güey	397
gacho	397
bato	396
gato	388
puto	386
pendejo	378

Una vez más, se observa que las unidades buscadas son sobre todo de carácter coloquial, popular o son de uso muy general (*casa*, *perro*, *gato*). Es interesante que las listas por frecuencia y dispersión geográfica y temporal muestren la misma tendencia. Por otra parte, vale la pena identificar qué consultas exhiben los valores mayores de los tres criterios de frecuencia y dispersión geográfica y temporal, lo que examinaremos en la próxima sección.

Índice de frecuencia y dispersión

Hay varias maneras de combinar los tres valores examinados en las secciones anteriores. Por ejemplo, para cada consulta se pueden promediar o multiplicar y, dependiendo de la importancia que se le quiera dar a cada valor, se les puede asignar pesos diferentes. En este experimento se consideró un mismo peso a cada uno de los tres criterios y se calculó un índice que los combina multiplicándolos. Los valores de este índice se pueden ver en la última columna de la Tabla 5; donde, por ejemplo, el índice de *pinche* es 197 x 369.12 x 27 = 1963332:

Tabla 5. Índice para combinar frecuencia, intervalo temporal y número de países

Petición de búsqueda	Frecuencia	Intervalo	Número de países	Índice
pinche	197	369.12	27	1963332
chingar	191	371.37	27	1915153
chido	244	379.06	17	1572352
chale	199	379.75	13	982405
naco	183	377.98	13	899215
güey	135	368.09	18	894471
pendejo	152	353.43	15	805812
chingada	105	374.30	20	786040
coger	116	382.53	16	709970
madre	115	369.77	15	637846
chilango	145	370.04	11	590206
guey	121	372.52	13	585980
bato	109	367.14	14	560260
verga	114	371.19	13	550108
perro	105	371.59	14	546234
gacho	117	368.06	12	516760
tomar	96	364.26	14	489563
cabron	85	359.93	16	489503
cuate	98	377.54	13	480992
padre	78	374.31	16	467142

Otra vez, las consultas aparecen ordenadas de mayor a menor índice. De esta manera, tenemos que las cinco consultas hechas desde más países, más profusas en el tiempo y más frecuentes son *pinche*, *chingar*, *chido*, *chale* y *naco*. Es interesante que las primeras dos se hayan hecho desde más países (desde 27), mientras que *chido* sea la más frecuente y *chale* y *naco* sean las que tienen búsquedas más tempranas y más tardías en el periodo examinado (las que más persisten).

Es de notarse que, con excepción de las consultas hechas sin diacríticos (*guey*, *cabron*) todas las peticiones de búsqueda que se muestran en esta tabla ya forman parte de la nomenclatura del DEM. De hecho, se puede ver que mientras mayor sea el índice, mayor será la probabilidad de que la consulta ya se encuentre en la nomenclatura. Las entradas que no están allí tienen rangos mayores (si se mostrara más de la tabla, aparecerían más abajo). La primera consulta que no pertenece a la nomenclatura es *wey*, que no está en la tabla por tener rango de 27; esta forma, que suele usarse para chatear en dispositivos electrónicos, es una variante del vocablo *güey*. Luego, con rango 93 estaría *chela* y con rango 154, *chafirete*. Sobra decir que estas unidades también son de carácter coloquial y popular.

Resultados posibles de las consultas al *DEM*

En el nivel más básico, las consultas están o no están en la nomenclatura. Cuando las consultas no se encuentran en ella, de todas maneras pueden estar presentes al interior de algún artículo lexicográfico, lo cual no deja de ser un acierto respecto al contenido del diccionario. Por ejemplo, en los artículos transcritos al final de este párrafo, después de la entrada y las categorías gramaticales, aparece entre paréntesis información sobre formas de escritura alternativa y pronunciación de la entrada. Llamamos a esta información entre paréntesis "Modelo" porque allí también se especifican, formas especiales del plural "(Su plural es *álbumes* o *álbums*)", del género opuesto "(Su femenino es *alcaldesa*)" y modelos de conjugación verbal "(Modelo de conjugación 1c)".

> **violoncellista** s m y f (Se pronuncia *violonchelista*) Persona que toca el violoncello, en especial la que lo hace profesionalmente: "Pau Casals ha sido considerado el mejor *violoncellista* del siglo", *el famoso violoncellista Duport*.
>
> **xoconoxtle** s m (También *xoconochtle*, *xoconostle*, *xoconoscle*, *xoconoxtli*, *soconoxtle* o *joconostle*. Se pronuncia *soconostle*, *joconostle* o *shoconoshtle*) 1 Tuna jugosa, de sabor agridulce, color rojo y de aproximadamente 3 cm de diámetro, que se emplea en dulcería y como condimento de algunos platillos regionales, particularmente en salsas o moles aguados, como en el mole de olla 2 (*Lemaireocereus stellatus*) Especie de nopal que da esta tuna; [...].
>
> **yoghurt** s m (También *yogurt*, *yogourt*, *yoghourt*, *yogour* o *yogur*. Se pronuncia *yogur*) Leche fermentada con bacilos búlgaros.

Lo importante es que los usuarios pueden buscar vocablos con formas de escritura alternativa (como *xoconochtle*, *xoconostle*, *xoconoscle*, *xoconoxtli*, *soconoxtle* o *joconostle*; y *yogurt*, *yogourt*, *yoghourt*, *yogour* o *yogur*) o, si no saben cómo se escriben, con indicaciones de su pronunciación (como *violonchelista*; *soconostle*, *joconostle* o *shoconoshtle*; y *yogur*). Así, si la consulta no está entre las entradas, se puede detectar que coincide con palabras, en algunos de los campos del artículo lexicográfico como el del "Modelo", que podrán considerarse aciertos en cuanto a la obtención de información pedida.

En otras palabras, los resultados de las consultas al DEM son más complejos que la simple determinación de su ausencia o presencia en la nomenclatura del diccionario. En la Tabla 6 aparecen los posibles resultados de las peticiones de búsquedas hechas al diccionario en línea, también ordenados de más a menos frecuentes. Nótese que algunos renglones de la tabla aparecen sombreados, lo cual se explicará más adelante. Cuando la consulta tiene eco en la nomenclatura, se le aplica automáticamente la etiqueta "Entrada". Cuando no aparece allí, pero aparece en la definición (como la palabra flexionada *aguados* en

la definición de *xoconoxtle*, citada arriba) o en algún ejemplo (como *Casals* y *Duport* en los ejemplos de *violoncellista*), la consulta se marca con la etiqueta "Definición" y, si aparece dentro de alguna locución que encabece alguna acepción, se marca con "Locución". Si resulta que coincide con alguna de las especificaciones de pronunciación o con alguna manera alternativa en que se deletrea una entrada, se aplica la etiqueta "Modelo". Si la consulta no coincide ni con una entrada, ni locución ni modelo por la ausencia o presencia de un diacrítico, a la etiqueta se le agrega la leyenda "sin acento".

Tabla 6. Resultados de las consultas

Resultado de la consulta	Frecuencia	Porcentaje
Entrada	106741	59.2%
No encontrado	42959	23.8%
Definición	11598	6.4%
Locución	7267	4.0%
Entrada sin acento	6541	3.6%
Entrada separada	2217	1.2%
Definición sin acento	1185	0.7%
Modelo	849	0.5%
Palabras similares	482	0.3%
Modelo sin acento	181	0.1%
Peculiar	144	0.1%

Por otra parte, a veces las consultas no están en la nomenclatura, por estar flexionadas (por ejemplo, *ocupaba*), ser nombres propios (*Amazonas*) o ser multipalabra (*Capsicum frutescens*),[6] pero ocurren en la definición, en algún ejemplo (lo que resulta, como se dijo, en la marca "*Definición*" o "*Definición sin acento*") o en la información entre paréntesis en la que se suele registrar datos sobre la pronunciación o maneras alternativas de escritura (lo que, como se dijo, se marca con "*Modelo*" o "*Modelo sin*

[6] Se contaron 2,878 consultas de dos o más palabras gráficas.

acento"). También se detecta si a la consulta le hace falta el acento gráfico o diéresis (por ejemplo, *guey*) para encontrar una coincidencia en la nomenclatura (la que se marca, como se dijo, con "*Entrada sin acento/diacrítico"*) o si se parece a alguna entrada de la misma (*diksionario* se parece a *diccionario*) según la distancia de Levenshtein[7] (estas consultas se marcan con la etiqueta "*Palabras similares"*). Por otra parte, cuando se determina por inspección que la consulta guarda algún parecido en significado o en forma con alguna entrada, se marca manualmente con "*Peculiar"*, como por ejemplo *wey*. Finalmente, si coincide con alguna de las expresiones que introducen alguna acepción (por ejemplo, *domingo siete* aparece en la expresión *salir con un domingo siete* que introduce la acepción tres del artículo lexicográfico dedicado a *domingo*) se marcan con la etiqueta "*Locución"*.

Estrictamente, sólo las que vienen marcadas con la etiqueta "*Entrada"* están en la nomenclatura (59.2%), pero, si tienen las etiquetas "*Entrada sin acento"* o "*Locución"*, es cuestionable que deban considerarse como no encontradas. Tampoco queda claro que una consulta multipalabra, que se marca con "*Entrada separada"*, deba tratarse como encontrada aunque cada una de sus palabras sí esté en la nomenclatura. En un esfuerzo de simplificación, en este experimento las consultas que resultaron con las etiquetas sombreadas en la Tabla 6, "*No encontrado"*, "*Definición"*, "*Definición sin acento"*, "*Entrada separada"* y "*Peculiar"* (similitud por inspección), se consideraron desaciertos, 32.35%; mientras que las demás se consideraron aciertos, 67.75% (respecto a la nomenclatura).

Consultas acertadas

En la Tabla 7 se muestran las veinte peticiones de búsqueda con mayor índice de frecuencia y dispersión que encontraron una respuesta positiva en el sistema; están ordenadas de mayor a menor valor de este

[7] El algoritmo de Levenshtein es un método estándar, para medir similitud entre palabras, que no se explicará aquí por falta de espacio. Este método se describe con detalle en varios capítulos de Jurafsky y Martin (2009).

índice. De nuevo se aprecia que se trata de unidades de uso popular o coloquial en México y algunas de uso muy general.

Tabla 7. Consultas con resultados afortunados en el *dem*

Petición de búsqueda	Estado en el *dem*	Índice
pinche	Entrada	1963332
chingar	Entrada	1915153
chido	Entrada	1572352
chale	Entrada	982405
naco	Entrada	899215
güey	Entrada	894471
pendejo	Entrada	805812
chingada	Entrada	786040
coger	Entrada	709970
madre	Entrada	637846
chilango	Entrada	590206
bato	Entrada	560260
verga	Entrada	550108
perro	Entrada	546234
gacho	Entrada	516760
tomar	Entrada	489563
cabron	Entrada sin acento	489503
cuate	Entrada	480992
padre	Entrada	467142
chingon	Entrada sin acento	444907

Si bien es cierto que se trata de vocablos comunes en el español de México y que sin duda son de interés para intérpretes y traductores, así como para interesados en la cultura mexicana en general, al considerar que la mayoría de los usuarios en todo el mundo puedan ser mexicanos, que las conocen y usan, se podría formular la hipótesis de que están

manifestando su necesidad de corroborar qué tan mexicano es el diccionario, como si lo mexicano fuera lo popular o lo coloquial; como si el diccionario fuera de mexicanismos y no integral.

Consultas fallidas

Con respecto a las consultas que no tuvieron resultados en la nomenclatura, en la Tabla 8 se observan las 40 con mayor índice de frecuencia y dispersión, ordenadas otra vez de mayor a menor índice. Lo más evidente es que de nuevo predominan vocablos de uso popular o coloquial, como *wey*, *chela*, *chafirete* y *achichincle*, algunos de los cuales ya se habían mencionado arriba.

Tabla 8. Consultas sin resultados en el *DEM*

Rango	Petición de búsqueda	Estado en el *DEM*	Frecuencia	Intervalo	Número de países	Índice
1	wey	Peculiar	78	372.52	13	377739
2	chela	Definición	38	341.01	11	142542
3	chafirete	No encontrado	44	400.02	6	105605
4	achichincle	Definición	25	365.63	10	91408
5	monitorear	No encontrado	32	372.05	7	83340
6	conciente	No encontrado	31	403.91	6	75127
7	chundo	No encontrado	44	415.05	4	73049
8	accesar	No encontrado	26	401.13	6	62577
9	chole	Definición	43	360.88	4	62072
10	cantinflear	No encontrado	36	410.36	4	59091
11	internet	Definición	35	329.51	5	57664
12	morra	Definición	25	328.42	7	57474
13	pacheco	Definición	22	315.08	8	55455
14	choro	No encontrado	35	393.83	4	55137
15	empatía	No encontrado	35	392.51	4	54951
16	changarro	No encontrado	25	366.20	6	54930
17	sustentable	No encontrado	34	396.13	4	53874
18	haiga	Definición	38	345.63	4	52536

19	mexico	Def. sin acento	24	357.20	6	51437
20	zape	No encontrado	31	403.05	4	49978
21	madrear	Definición	18	346.32	8	49870
22	agendar	No encontrado	23	407.67	5	46882
23	chompiras	No encontrado	30	389.60	4	46752
24	pollera	No encontrado	16	322.98	9	46509
25	madriza	No encontrado	17	390.37	7	46454
26	guarro	No encontrado	23	390.90	5	44953
27	sale	Definición	17	364.26	7	43347
28	chucho	Definición	25	344.69	5	43086
29	choya	No encontrado	31	345.34	4	42822
30	sustentabilidad	No encontrado	21	400.87	5	42091
31	tacuche	No encontrado	27	384.58	4	41535
32	tranza*	Definición	19	311.24	7	41394
33	cachiporra	Definición	25	271.62	6	40743
34	gafete	Definición	16	359.12	7	40221
35	zoquete	No encontrado	17	385.15	6	39285
36	fayuca	Definición	28	341.72	4	38273
37	aeromoza	No encontrado	15	278.49	9	37596
38	chutas	No encontrado	33	376.53	3	37277
39	chutar	No encontrado	23	395.75	4	36409
40	chota	Definición	40	298.00	3	35760

Por otra parte, además de las unidades de carácter coloquial y popular, se observan vocablos como: *monitorear* (rango 5), *accesar* (8), *cantinflear* (10), *internet* (11), *empatía* (15), *sustentable* (17), *agendar* (22) etcétera,

* Sorprende que la consulta *tranza* se haya llevado a cabo desde siete países. Tal vez los usuarios hayan querido buscar *transa* por lo que esta consulta podría haberse tomado como "Sí encontrada" (por similitud, lo que se detectaría mediante el algoritmo de Levenshtein). También, *tranza* puede verse como una forma flexionada del verbo *tranzar* (tampoco en el DEM), por lo que en esta tabla debería habérsele asignado la etiqueta "No encontrado". Sin embargo, aparece con la etiqueta "*Definición*" porque en el DEM ocurre en el interior de la palabra "Mastranza" en el ejemplo de la entrada *bravura* ("Se esfuerza por mandar a la **Maestranza** lo mejor de sus dehesas, en cuanto a *bravura* y presentación").

que no están todavía definidas en el diccionario, por no haber aparecido en la base estadística del CEMC, como *empatía*, o porque su uso ha empezado a generalizarse apenas recientemente, como *monitorear* o *internet*. El caso de *cantinflear* es interesante, ya que no es ni de uso popular ni coloquial, pero probablemente aparece aquí porque se percibe como tal.

Lo que hay que recalcar es que los otros vocablos que no son ni de uso popular ni coloquial se refieren a cuestiones con las que los hablantes estamos en contacto hoy en día, como aquellos relacionados con internet o los nuevos dispositivos electrónicos, y que por su frecuencia y dispersión de consulta merecen ser tomados en cuenta seriamente para su inclusión en las próximas versiones del diccionario.

Hacia una evaluación del diccionario

El ideal de un diccionario en línea es que a toda consulta corresponda una respuesta; esto es, que cada petición de búsqueda encuentre una coincidencia dentro del conjunto de entradas que conforman la nomenclatura o dentro del conjunto de locuciones o expresiones multipalabra que encabezan numerosas acepciones, dentro del conjunto de descripciones de pronunciación o formas de escritura alternativas o dentro del conjunto de palabras similares que se pueden determinar mediante distancias de Levenshtein.

Para evaluar esto, basta determinar el porcentaje de consultas acertadas en el sistema, respecto al total de consultas. Este porcentaje se conoce también como medida de precisión y es un valor de predicción positiva. Para el DEM en línea podemos calcularlo de varias maneras. Primero, se puede considerar el volumen de consultas acertadas respecto al total de consultas recibidas (122,061 de 180,164), lo cual resulta en un porcentaje de 67.75%. Segundo, se pueden considerar los tipos de consultas (16,505 tipos de consultas acertadas de un total de 24,944 tipos de consultas), lo que resulta en una precisión de 66.17%. Sin embargo, al examinar las consultas no acertadas, se observan palabras flexionadas (*haiga*, *bésame*, *fien*, *álbumes*, etcétera), nombres propios (*méxico*, *oaxaca*, *batman*, *chómpiras*, etcétera) y errores de los usuarios (*eitca*, *oe*, *hti*, etcétera) que no son fallas del diccionario, sino del usua-

rio. Por inspección se encontraron 696 tipos de consulta que deben considerarse errores y no fallas del diccionario. Así que una tercera manera de estimar la precisión sería mediante la proporción de tipos acertados respecto a 24,248 (24,944 – 696), lo que da un porcentaje de 68.07%, casi 7 de cada 10 consultas diferentes.

Observaciones finales

Como bien se puede apreciar en los datos presentados, en el *dem* en línea se registran consultas desde todo el mundo. La mayoría se hacen desde México, Estados Unidos y países de habla hispana. Es alentador que el 67.65% del volumen total de las consultas coincidan con la nomenclatura actual del diccionario. De hecho, esto puede verse como una medida de precisión para evaluar la base estadística que se obtuvo a partir del *cemc*. Además, al eliminar errores de los usuarios y consultas flexionadas, se observa una precisión de 68.06%.

Al considerar las consultas más frecuentes, hechas desde más países y más persistentes en el tiempo, se observan vocablos que ya son parte de la nomenclatura del *dem*. Estos vocablos son predominantemente de carácter popular y coloquial (*pinche*, *chido*, *chingar*, etcétera). Parece que los usuarios necesitan desafiar al diccionario en lo que perciben más mexicano y no lo perciben como un diccionario integral, sino de mexicanismos. De todas maneras, al considerar que de cada diez consultas tres no coinciden con la nomenclatura, este reto a la "mexicanidad" del diccionario logra enfrentarse dignamente.

Por otra parte, entre las consultas sin respuesta se observan vocablos como: *monitorear*, *accesar*, *internet*, *empatía*, etcétera, que se refieren a la realidad actual de los usuarios. Desde un ejercicio de adquisición léxica, estos vocablos, ordenados de mayor a menor índice de frecuencia y dispersión geográfica y temporal, merecen ser considerados a corto plazo para su inserción en la nomenclatura del diccionario.

Fuentes consultadas

cemc: *Corpus del español mexicano contemporáneo*, México: El Colegio de México (colmex). Disponible en línea en http://cemc.colmex.mx. Agosto de 2012-septiembre de 2013.

dem: Lara, L. F. (dir.) (2010*), Diccionario del español de México*, México: colmex. Disponible en línea en http://dem.colmex.mx. Agosto de 2012-septiembre de 2013.

Jurafsky, D., Martin, J. H. (2009), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2a ed., Englewood Cliffs, nj: Prentice Hall/Pearson.

Lara, L. F., Ham Chande, R., García Hidalgo, M. I. (1980), *Investigaciones lingüísticas en lexicografía*, México: El Colegio de México.

Manning, C. D., Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, ma: mit Press.

Fecha de recepción: 27 de marzo de 2014
Fecha de aprobación: 27 de agosto de 2014