3 Metodología

En esta sección se hablará del concepto de *fuentes* dentro del área de la lexicografía, y se definirán los tipos de fuentes que se pueden emplear en la labor lexicográfica. Al mismo tiempo se definirá en qué consiste el concepto de *fuentes primarias* y de *fuentes secundarias*, y se dirá cuáles son las fuentes primarias y secundarias particulares de este trabajo. Además se explicará cuáles han sido los procedimientos que se han seguido para obtener las fuentes secundarias y el tratamiento que se ha dado a las mismas.

3.01 Fuentes lexicográficas

Los diccionarios se basan en un sistema lingüístico que puede tener distintas procedencias, es decir, el material léxico que compone la fuente de información de un diccionario puede tener varios orígenes. En la práctica lexicográfica, la recolección de dicho material léxico puede darse por dos vías principales: a través de fuentes secundarias, recogiendo el conjunto de materiales útiles de otros diccionarios y estudios lexicográficos, o por medio de fuentes primarias, aprovechando un corpus de materiales originales como textos, grabaciones, encuestas, entre otros (Haensch, 1997, pp. 54-55; 1982a, p. 435). Sea cual sea el tipo de fuentes que se utilice o que se preferencie en la labor lexicográfica, la elaboración de un diccionario debe estar siempre precedida por la recopilación de dichas fuentes (Porto, 2002, p. 84).

3.02 Fuentes secundarias

De acuerdo con la teoría de la práctica lexicográfica, el primer paso a llevar a cabo en la elaboración de un diccionario es la revisión de las fuentes secundarias. En lexicografía

se entiende por fuentes secundarias a todos aquellos materiales que ofrecen vocablos potenciales a incluir en un nuevo diccionario y que contienen a la vez algún tipo de explicación metalingüística respecto de dichos vocablos (Haensch, 1982a, p. 436). Entre este tipo de materiales, resultan de gran utilidad los diccionarios existentes pues éstos cubren una gran cantidad de información (Sinclair, 1985, p. 82). Uno de los mayores méritos de las fuentes secundarias está en el hecho de que la información contenida en las mismas está ya organizada. Es así como una síntesis de buenos materiales, con la actualización y corrección necesarias, podría dar lugar a la elaboración de un diccionario confiable en un periodo de tiempo relativamente corto (Sinclair, 1985, p. 81). Claro está que la utilización de materiales lexicográficos previos debe tomar ciertas precauciones para no caer en el simple plagio y en el estatismo de la práctica lexicográfica.

Por ello, se deben tener en cuenta por lo menos tres desventajas inherentes al uso de las fuentes secundarias. Primero, es difícil conocer cuándo un término pierde vigencia, por lo cual es especialmente peligroso utilizar diccionarios ya existentes en la redacción de diccionarios de un estado de lengua contemporáneo. Segundo, los errores cometidos en la inclusión de vocablos de otros diccionarios son difíciles de detectar. Tercero, existen nuevos avances en la descripción lexicográfica actual, como la información pragmática, que no existían en la concepción teórica de trabajos anteriores (Sinclair, 1985, p. 81). Sinclair afirma que ningún material léxico debería incorporarse en la elaboración de un nuevo diccionario hasta que no se confirme su existencia de manera independiente a su registro en el diccionario fuente. Esta es la única manera, asegura Sinclair, en que la lexicografía se habrá de liberar gradualmente de dos tipos de elementos prescindibles: las voces caídas en desuso, definitivamente innecesarias en los diccionarios contemporáneos, y los materiales léxicos que son meros productos de la lexicografía y que no existen más allá

de ella, al menos en el sentido de que no se encuentra ninguna evidencia textual de los mismos (1985, p. 82). Con todo, los diccionarios existentes no dejan de ser una fuente importante en la elaboración de nuevos diccionarios y sus materiales deberían ser aprovechados. Sin embargo, el aprovechamiento de estos materiales deberá estar condicionado por la averiguación de: la vigencia de uso de la unidades léxicas en consideración, los posibles cambios de registro experimentado por éstas, la pertinencia de la inclusión de las mismas según la finalidad del diccionario en proyecto y la aparición de nuevas acepciones para cada una de ellas (Haensch, 1997, p. 55).

Siguiendo estas recomendaciones acerca de la utilización de las fuentes secundarias, el material lexicográfico existente que se tomó en cuenta en este estudio está conformado por los subconjuntos del léxico del subestándar de los diccionarios impresos en los últimos veinticinco años cuya extensión abarca de alguna manera el español en México. En cuanto a la forma en que estos diccionarios abarcan el español en México, aquí se ha partido únicamente de lo que los mismos diccionarios expresan en sus títulos, instrucciones de uso, lugares de publicación, entre otros tantos elementos que hacen patente la mencionada extensión de sus materiales. Tomando todo esto en consideración, los diccionarios de los que se han extraído los subconjuntos del léxico del subestándar son siete. Estos siete diccionarios, enlistados una vez más con su nombre completo, las siglas utilizadas en este trabajo para referirlos, el tipo de diccionario al que pertenecen, el nombre de su autor y el año de su publicació n, aparecen en la Tabla 3.1.

Tabla 3.1. Lista de diccionarios fuente del léxico subestándar del español en México

Nombre del diccionario	Siglas utilizadas	Tipo de diccionario	Autor	Año
(a) *Diccionario del español usual en México*	(*DEUM*)	general	Lara	(1996)
(b) *Diccionario de la lengua española*	(*DRAE*)	general	RAE	(2001)
(c) *Diccionario breve de mexicanismos*	(*DBM*)	regional	Goméz de Silva	(2001/ 2003)
(d) *Diccionario inicial del español en México*	(*DIME*)	escolar	Ávila y Aguilar	(2003)
(e) *Tumbaburro de la picardía mexicana: Diccionario de términos*	(*DTV*)	subestándar	Jiménez	(1999)
(f) *El chingolés: Primer diccionario del lenguaje popular mexicano*	(*PDLPM*)	subestándar	Usandizaga	(1994)
(g) *Así habla la delincuencia y otros más…*	(*AHDOM*)	subestándar	Colín Sánchez	(1987/ 2001)

Finalmente, falta hacer la aclaración de que en el presente estudio se decidió excluir en lo posible los elementos léxicos pertenecientes al ámbito jergal comúnmente incluido en el subestándar, en el que se encuentra la jerga de los estudiantes o la de los grupos humanos marginales. Esto se debe, principalmente a dos factores. En primer lugar, el corpus electrónico con que se va a trabajar, el *CREA*, es un corpus más bien de lengua general y estándar que, a pesar de tener como finalidad el llegar a ser "un corpus de referencia… lo suficientemente extenso para representar todas las variedades relevantes de la lengua en

cuestión [el español]" (RAE, 2004, noviembre), bien puede ofrecer una cierta representación del subestándar, parece poco probable que brinde resultados dignos de consideración respecto de la parte de la lengua que se denomina jergal. Además, tomando en cuenta que el *CREA* es el único corpus de dominio público en el español y que no existen todavía en esta lengua corpus electrónicos especializados, por lo menos no en el ámbito del subestándar, podríamos decir que las voces jergales pertenecientes a grupos marginales se encuentran lejos de la posibilidad de ser ejemplificadas con citas a través de este tipo de medio tecnológico. Esto parece ser cierto cuando menos en el caso de la lengua española. En segundo lugar, si se toma en cuenta que este proyecto tiene cortos alcances, ya que sus dimensiones están restringidas en tiempo y en recursos, resulta razonable restringir la variación diastrática o social a la propia del no estándar. Dicha restricción es comprensible si se tiene presente que la variación social del no estándar parece tener una mucho mayor posibilidad de representación en los textos que componen el *CREA* frente a la posibilidad de documentar la variación social jergal. Esta decisión de restringir la variación social se deriva de los criterios de construcción y composición de este corpus electrónico, que busca más bien recabar documentación de la lengua general y estándar (RAE, 2003).

3.03 Fuentes primarias

Ahora bien, si recordamos las precauciones que hacen Haensch (1997; 1982a) y Sinclair (1985) respecto de la utilización de las fuentes secundarias, podemos decir que éstas adquieren su validez en la labor lexicográfica sólo a través de las fuentes primarias. Las fuentes secundarias constituyen una reunión de material importante y útil en la elaboración de diccionarios; sin embargo, "los verdaderos progresos de la lexicografía se deben al aprovechamiento de fuentes primarias, es decir, de textos en sentido más amplio,

donde la unidad léxica que interesa aparece, por lo general, en un contexto" (Haensch, 1982a, p. 437).

Sin embargo, antes de hablar de este tipo de textos, se debe recordar que en las fuentes primarias se puede hacer una distinción de dos tipos: la introspección y la observación del lenguaje en uso. La introspección a su vez se subdivide en el examen de informantes y la introspección del lingüista (Sinclair, 1985, p. 81). El examen de informantes puede ser espontáneo, recogiendo unidades léxicas que aparecen en la lengua hablada, o sistemático, por medio de encuestas (Haensch, 1982a, p. 442). La recolección espontánea e informal continúa siendo controversial en cuanto a su validez, pues se trata de una forma de documentación con respaldos únicos y difíciles de corroborar en otras instancias igualmente espontáneas. Por su parte, la recolección sistemática o formal de materiales a partir de informantes, es decir la encuesta, presenta la limitante de que requiere una gran inversión de tiempo, preparación y recursos económicos. Por ello tiene pocas posibilidades de convertirse en una fuente importante de materiales primarios (Sinclair, 1985, p. 82). En cuanto a la introspección del lexicógrafo, Haensch hace notar su vigencia e importancia, pues aun con los grandes corpus electrónicos que han aparecido y siguen apareciendo en la actualidad, la carencia de representatividad de vocablos y la falta de interpretación que caracteriza a dichos corpus mantienen vivo el papel de este tipo de reflexión en la labor lexicográfica (1982a, p. 443). El redactor del diccionario, y por supuesto el lector del mismo, sigue siendo el único que puede interpretar una forma, un sentido o su caracterización. Con todo, en la elaboración de diccionarios ejemplificados, el mayor defecto de la introspección se deriva de que no proporciona evidencia del uso. Además, la introspección de un informante difícilmente proporcionará una distinción entre todos los posibles patrones lingüísticos de un elemento léxico. El uso de los vocablos no es

un elemento consciente en la competencia del hablante, de hecho es más bien un conocimiento implícito, y el tratar de definirlo partiendo de la introspección del lingüista podría dar lugar a que se registren ideas y no hechos acerca de la lengua. Sobre todo cuando se trata de hacer ejemplificación, el auténtico potencial de la introspección está en la evaluación de la evidencia del material léxico y no en su creación (Sinclair, 1985, p. 82). En este sentido, si bien el comentario de Sinclair está dirigido al campo de la lexicografía, vale la pena mencionar que la introspección del lingüista ha jugado un papel fundamental en otras áreas de la lingüística, como por ejemplo en el análisis sintáctico inscrito en la tradición chomskiana (Kochanski, 2003, Diciembre). Por ello, los comentarios de Sinclair sobre la introspección bien podrían aplicarse a otras áreas de la lingüística.

Por todo lo anterior, el tipo de fuente primaria que debería aparecer en primer lugar, según Sinclair, es la observación del lenguaje en uso, es decir, la utilización de un corpus (1985, p. 82). Respecto del corpus, este mismo autor distingue el uso de fichas de referencia, como el método tradicional, y lo separa del uso de concordancias, o listas de ejemplos de uso, proporcionadas por computadoras, o sea, de la utilización de corpus electrónicos. Además se declara a favor de este último método (Sinclair, 1985, p. 83).

El corpus electrónico utilizado en este trabajo es el *CREA*, que contaba con 140 millones de palabras en el momento en que se comenzaron a hacer las primeras pruebas con él, en octubre de 2003, y que alcanzó los 156 millones de palabras para cuando se terminó la documentación del muestreo aleatorio, en julio de 2004. Este corpus fue elegido porque, además de ser de uso público, se pensó que su composición geográfica podría resultar favorable para la exploración de la representatividad de los distintos subconjuntos del léxico subestándar en las fuentes secundarias trabajadas. Respecto de dicha composición, hay que hacer notar que el 50% del total del corpus está dedicado al español de América.

De esta porción el 40% pertenece al español de la “parte mexicana” del mundo hispano, que en el corpus está conformada por el español de México, del suroeste de los Estados Unidos, de Guatemala, de Honduras y del Salvador. Además, se debe mencionar que el periodo de tiempo a que pertenecen los materiales de este corpus va de la actualidad hasta veinticinco años atrás (RAE, 2003); periodo al que se ha restringido la composición de las fuentes secundarias de este estudio. Como una aclaración metodológica respecto del ámbito geográfico, quiero señalar que en este estudio se utilizaron únicamente aquellos documentos del corpus que pertenecen a México y no los que pertenecen a la “parte mexicana”.

3.04 Diseño

Para determinar la factibilidad de ejemplificación de los materiales léxicos recogidos, se exploró el nivel de representatividad que alcanzaban los subconjuntos del léxico subestándar que conforman las fuentes secundarias. Como el número de elementos léxicos que componen cada una de las fuentes secundarias sumaban un número demasiado alto para ser explorado en su totalidad en este trabajo, se hizo un muestreo aleatorio de cada uno de los diccionarios. Como el muestreo se llevó a cabo de forma separada para los distintos diccionarios que componen las fuentes secundarias, dicho muestreo resultó ser, además, un muestreo estratificado. Una vez hecha la exploración de la representatividad en el corpus, los resultados se han reportado en porcentajes. Esto se ha hecho separadamente para cada una de las muestras aleatorias del léxico subestándar de los distintos diccionarios con la finalidad de estimar el resultado que se obtendría si se explorara la representatividad del total de las fuentes. Así pues, el presente trabajo está basado en un diseño cuantitativo

que se apoya en un método estadístico simple que busca inferir el comportamiento de una población a partir del de una muestra aleatoria estratificada de la misma.

3.05 Expectativas y limitaciones

Este trabajo ha tenido como finalidad primordial especular acerca de las posibilidades de construcción de un diccionario ejemplificado del léxico subestándar del español en México por medio de un corpus electrónico de acceso público. Ante todo, se ha buscado predecir el resultado potencial de una empresa de este estilo. Desde un principio, se ha buscado poder mostrar la pertinencia de elaboración de un diccionario como el ya mencionado por medio de este tipo de recursos, y comparar los posibles resultados de su elaboración con los diccionarios existentes en otras variantes dialectales del español que cuentan con diccionarios similares, como es el caso del español peninsular, así como con diccionarios de otras lenguas como el inglés.

Por otro lado, también se ha pensado en que las posibles diferencias en el nivel de representatividad de los distintos diccionarios incluidos en la muestra podrían sugerir la exploración, en un futuro estudio, de algún tipo de influencia entre los procedimientos de presentación formal de los diccionarios y la representatividad del material léxico de los mismos en un corpus electrónico. Más allá, se podría intentar establecer algún tipo de correlación entre algún criterio individual de evaluación formal y de contenido de los diccionarios (Haensch, 1997, pp. 239-242) y el nivel de representatividad de sus materiales léxico en un corpus electrónico. Es decir, podríamos comenzar a especular acerca de qué tipo de características formales de un diccionario están en correlación con una mayor posibilidad de documentación de sus elementos léxicos por medios informáticos.

Con todo, hay que reconocer, como ya se hizo anteriormente, que el presente trabajo se enfrenta con fuertes limitaciones, encontrándose la principal de ellas en el hecho de que el subconjunto del léxico que se pretende explorar en el *CREA*, el léxico del subestándar, no es un elemento que defina a este corpus en su composición. En este sentido, sin embargo, habría que recordar que el *CREA* tiene como finalidad llegar a ser "un corpus de referencia ha de ser lo suficientemente extenso para representar todas las variedades relevantes de la lengua en cuestión [el español]" (RAE, 2004, noviembre). Así pues, hemos decidido utilizar al *CREA*, puesto que hasta el día de hoy continúa siendo el único corpus de dominio público que existe para el español (Porto, 2002, p. 129) y continúa siendo utilizado de manera primordial para el trabajo de actualización y realización de los diccionarios de la RAE (2004, noviembre). Además, nos parece pertinente seguir la ya referida recomendación de Sinclair acerca de la utilización de los medios disponibles en lexicografía, aun cuando éstos no aparezcan como ideales (1985, p. 86).

3.06 Procedimientos

Una vez seleccionados los siete diccionarios de los cuales se podrían extraer las posibles fuentes secundarias del léxico subestándar del español en México, se procedió a hacer dicha extracción. Ésta varió según el tipo de diccionario. Así en los diccionarios generales (el *DEUM* y el *DRAE*), en el escolar (el *DIME*) y en el de regionalismos (el *DBM*), se buscó únicamente el material perteneciente al subestándar, con excepción de los materiales léxicos propios de las jergas de grupos humanos marginales. Esto último, como ya se dijo, debido a que de acuerdo con la composición del *CREA* parecía ínfima la posibilidad de identificar léxico jergal en el mismo. De manera contraria, en los diccionarios del subestándar (el *DTV*, el *PDLPM* y el *AHDOM*) se intentó identificar todo

el léxico que no perteneciera al subestándar, además de que se buscó eliminar aquello que se considerara propio del léxico jergal.

Para identificar los lemas que respondiesen a los criterios de búsqueda del subestándar aquí planteados, se analizaron las informaciones metalingüísticas que acompañaban al lema, al encabezado de la definición o a la definición misma. Estas informaciones metalingüísticas pueden aparecer de diversas formas, ya sea como abreviaturas, comúnmente antes de la definición, o como datos desarrollados dentro de ella (Bajo, 2000, pp. 22-23). Este tipo de informaciones metalingüísticas se conocen comúnmente como *marcas*. En la definición de Porto, las marcas son "expresiones, utilizadas a lo largo de toda la obra, constituidas casi siempre por frases estereotipadas, abreviaturas, signos especiales o ciertos recursos gráficos (por ejemplo, un estilo o tamaño de letra especial), cuya misión es 'marcar' o destacar una palabra o acepción frente a otras que, por no presentar ninguna característica especial o, por el contrario, la que se considera normal o general, aparecen en el diccionario como elementos 'no marcados'" (2002, pp. 250-251). Así pues, para la identificación de las marcas que permitiesen extraer las unidades léxicas pertinentes a este estudio, o que permitieran la eliminación de las unidades léxicas no pertinentes, se tuvo que dar un tratamiento distinto a cada diccionario. De hecho, la identificación de marcas, propiamente hablando, se llevó a cabo únicamente en el *DEUM*, en el *DRAE*, en el *DBM*, en el *DTV* y en el *DIME*, pues sólo estos diccionarios contenían este tipo de información. Hay que hacer destacar que el trabajo de identificación y extracción de los materiales léxicos de las fuentes secundarias no lo realicé yo solo, sino que lo llevé a cabo, en dos etapas distintas, con la ayuda de un grupo de becarios[1].

[1] Estos becarios son estudiantes que, debido a que reciben descuentos en las colegiaturas de la escuela en que son alumnos, colaboran con los profesores de la institución en proyectos académicos.

En la primera etapa de dicho trabajo recibí apoyo de un grupo de ocho becarios, que trabajaron una hora diaria, cinco días a la semana, durante doce semanas. Entre estos primeros colaboradores había cuatro estudiantes de preparatoria (Anabel Barreda, Isabel Palafox, Paola Osorio e Iyari Zerón) y cuatro estudiantes de licenciatura (Edgar Esquivel, Daniel Márques, Claudia Montesinos y Gabriel de Santos). La primera etapa consistió en la identificación y creación de listas de papel y listas electrónicas de los seis primeros diccionarios que se describen en el siguiente subcapítulo (ver sección 3.07). En cuanto a los criterios de identificación, éstos fueron determinados por mí de acuerdo con la delimitación del subestándar antes presentada (ver sección 2.01.05). Debido a ello, el trabajo de los becarios consistió únicamente en buscar, subrayar y copiar los lemas con las marcas que les fueron indicadas; encomiendas que, si bien implicaban una gran cantidad de trabajo, no requirieron una extensa capacitación. Además, todas estas actividades fueron realizadas de manera múltiple e independiente, y sus resultados fueron cotejados para asegurar la validez del trabajo. Al mismo tiempo, este primer equipo de trabajo contribuyó a la inclusión del significado de los lemas y acepciones en las listas electrónicas finales del *DEUM*, del *DRAE* y del *DBM*. En cuanto a la inclusión del significado de estos diccionarios es sus respectivas listas, ésta fue la más fácil pues se usaron versiones electrónicas de dichos diccionarios y por tanto el proceso de inclusión consistió en un simple proceso de copiado y pegado de información. En la segunda etapa del trabajo de identificación y creación de las listas de las fuentes secundarias, tuve a mí cargo a un grupo de diez becarios, cuyos nombres son los siguientes: Luis Ángel González, María de Lourdes Cessa, Daniela Sánchez, Edgar Esquivel, Raúl Israel Fernández, Juan Carlos González, Claudia Montesinos, Gabriel de Santos, Ricardo Martínez y Yolanda Ruiz. De estos diez becarios, sólo el primero es estudiante de preparatoria y el resto son estudiantes de licenciatura. Este

segundo grupo de becarios trabajó de forma variada, pues algunos trabajaron tan sólo tres semanas en el proyecto y otros trabajaron hasta diez semanas, a razón de una hora diaria de lunes a viernes. Durante esta segunda etapa se identificaron los lemas pertinentes del *DIME* y se crearon las listas de dicho diccionario, que es el último de los presentados a continuación. Además, se capturaron manualmente los significados en las listas de todos los diccionarios de los que no se copiaron los significados durante la primera etapa.

3.07 Identificación del léxico del subestándar en los diccionarios fuente

En esta sección se presentarán los procedimientos utilizados para identificar la información a extraer de los diccionarios fuente. Cabe aclarar que, debido a las características particulares de estos diccionarios, los procedimientos de la mencionada identificación variaron ampliamente para cada uno de ellos. Así pues, los procedimientos de identificación se presentan en diferentes subsecciones tituladas con las siglas del diccionario al cual se refieren. En cada subsección, además, se menciona la creación de una serie de listas con la información identificada. Estas listas tuvieron como finalidad la construcción de una lista final, para cada diccionario, con la información pertinente al subestándar. Aquí las listas finales serán mencionadas sólo de paso, pues el tratamiento que se les dio a las mismas es el tema de la siguiente sección (ver sección 3.8).

3.07.01 DEUM. De este diccionario se localizaron tres marcas definitorias del tipo de léxico que se buscaba, es decir, del subestándar con exclusión de lo jergal: la marca *Popular*, la marca *Ofensivo* y la marca *Groser* (grosería). Para identificar las entradas del diccionario que incluyeran cualquiera de estas marcas, se comenzó trabajando en la versión impresa del diccionario. Se procedió a revisar uno por uno cada uno de los lemas y los números que indican las distintas acepciones de cada entrada del diccionario, los cuales

aparecen en negritas en el mismo. Durante este proceso, se resaltaron los lemas marcados y las marcas pertenecientes al subestándar. Fue de utilidad la coherencia de este diccionario en el hecho de que tanto el lema como el número que distingue a las entradas están siempre en negritas, además de que las marcas se encuentran siempre después de estos dos elementos. Resaltándolas con marcadores directamente sobre el texto del diccionario, se identificaron los materiales pertinentes para este estudio de entre las 14 mil entradas y 60 mil acepciones de este diccionario (Lara, 1996, p. 14). Después se hicieron independientemente dos listas a mano, que más tarde se cotejaron entre sí. Una vez hecho el cotejo de las dos listas, los datos resultantes del cotejo se capturaron en una hoja de cálculo de Excel. Al mismo tiempo, se utilizó la versión digital de este diccionario de la Biblioteca virtual Miguel de Cervantes, que se encuentra en Internet (Lara, 2000). No se utilizó la versión en Internet del Centro de Estudios Lingüísticos y Literarios del Colegio de México (Lara, 1996), pues en el momento en que se trabajó con la versión de la Biblioteca virtual Miguel de Cervantes, no conocíamos la existencia de esta versión con opciones de búsqueda avanzada. Así pues, utilizando la versión de la Biblioteca virtual Miguel de Cervantes, se revisó dos veces más el lemario completo del diccionario aplicando el buscador de Internet Explorer. Se crearon dos listas más, que se cotejaron entre sí y que después se compararon con la hoja de cálculo creada a partir de las listas del diccionario impreso. El resultado de la identificación de marcas en el *DEUM* dio un total de 633 marcas en 568 lemas. La Tabla 3.2 presenta los resultados, por letra, de la identificación de marcas en este diccionario. En total se encontraron 482 marcas *Popular*, 67 marcas *Ofensivo* y 84 marcas *Groser.*

Tabla 3.2. Conteo por letra de los lemas con marcas del subestándar en el *DEUM*

Letra	Popular	Ofensivo	Grosero	Subtotal
A	94	2	2	98
B	29	-	2	31
C	68	3	16	87
CH	32	13	15	60
D	17	1	-	18
E	16	1	-	17
F	27	9	8	44
G	13	6	-	19
H	13	3	1	17
I	1	4	-	5
J	29	12	10	51
L	13	-	2	15
LL	4	-	-	4
M	16	-	14	30
N	7	2	-	9
Ñ	1	-	-	1
O	7	-	-	7
P	24	3	11	38
Q	3	-	-	3
R	20	4	3	27
S	14	-	-	14
T	18	3	-	21
U	2	-	-	2
V	4	-	-	4
Y	8	-	-	8
Z	2	1	-	3
Subtotal	482	67	84	Total
				633

Marcas dobles	**33**
Marcas triples	**16**
Total de Lemas	**568**

3.07.02 DRAE. El procedimiento para este diccionario fue más simple, pues se utilizó desde un principio la versión en CD ROM (RAE, 2003). Así, recurriendo al Árbol de ámbito geográfico, en la subcarpeta de América, en la subcarpeta de Áreas geográficas, en la subcarpeta de México, se tuvieron que revisar de todo el diccionario tan sólo 2,905 marcas *Méx* en 2,444 entradas. De esta manera no se tuvo que hacer la búsqueda sobre las 88,431 entradas y 190,581 acepciones del diccionario impreso. Por ser éste un diccionario que incluye lemas de todo el mundo hispano, se buscó solamente la combinación de la marca geográfica del español en México, con alguna de las marcas que podrían reflejar unidades léxicas propias del subestándar, excluyendo las marcas propias de lo jergal. Así se decidió hacer la extracción de aquellos lemas que combinaran la marca *Méx* (México) con alguna de las siguientes: coloq. (coloquial), malson. (malsonante), vulg. (vulgar), despect. (despectivo), eufem. (eufemismo), e irón. (irónico). Aquí se debe recordar que, según se comentó en la subsección de La variación lingüística y el subestándar, de la sección Componentes teóricos del léxico del subestándar, de la Revisión bibliográfica, en este estudio se decidió incluir algunos elementos especiales de la variación social que depende de la valoración social del receptor hecha por el hablante (ver sección 2.01.05). Estos elementos, incluidos en este diccionario y en el siguiente que se presenta, no son comúnmente considerados parte del subestándar, pero se encuentran íntimamente relacionados con el léxico tabuizado, el cual sí es considerado por algunos teóricos como un elemento del subestándar. Las marcas que responden a esta inclusión en este diccionario son las marcas: despect. (despectivo), eufem. (eufemismo), e irón. (irónico). Para cuidar la calidad de la extracción de los lemas pertinentes, se llevaron a cabo de forma independiente un par de revisiones de la subcarpeta *México* del diccionario electrónico, y se generaron dos listas manuales que más tarde se cotejaron entre sí. La lista con los datos resultantes del

cotejo se comparó con la información contenida en el diccionario impreso y, finalmente, se capturó en una hoja de cálculo de Excel. Los resultados por letra de la identificación de marcas en el *DRAE* han sido presentados en la Tabla 3.3. Como se puede ver en esta tabla, para este diccionario se encontró un total de 215 combinaciones de marcas, en un total de 202 entradas. En dichas combinaciones, cuando menos una marca era una marca diatópica correspondiente al ámbito geográfico de México y otra marca era una marca diastrática-diafásica propia del subestándar. Así, se identificaron 155 combinaciones de la marca *Méx* con la marca *coloq*., veinticinco combinaciones con la marca *malson*., veinte combinaciones con la marca *vulg*., diez combinaciones con la marca *despect*., cuatro combinaciones con la marca *eufem*., y una combinación con la marca *irón.*

Tabla 3.3. Conteo por letra de los lemas con marcas del subestándar en el *DRAE*

Letra	coloquial	malsonante	vulgar	despectivo	eufemismo	irónico	Subtotal
A	6	2	-	-	-	-	8
B	13	-	-	-	-	-	13
C	27	2	2	1	-	-	32
CH	17	13	2	2	1	-	35
D	2	-	-	-	-	-	2
E	7	-	-	1	-	1	9
F	8	-	10	-	-	-	18
G	6	-	1	2	-	-	9
H	6	-	-	-	1	-	7
I	1	-	-	-	-	-	1
J	6	-	1	2	-	-	9
L	1	-	-	-	-	-	1
LL	-	-	-	-	-	-	-
M	14	7	1	-	-	-	22
N	-	-	1	-	-	-	1
Ñ	1	-	-	-	-	-	1
O	2	-	-	-	-	-	2
P	16	1	-	1	-	-	18
Q	-	-	-	-	-	-	-
R	5	-	-	1	-	-	6
S	5	-	-	-	-	-	5
T	6	-	2	-	1	-	9
U	-	-	-	-	1	-	1
V	5	-	-	-	-	-	5
Y	-	-	-	-	-	-	-
Z	1	-	-	-	-	-	1
Subtotal	155	25	20	10	4	1	Total
							215

Lemas con marcas dobles	**13**
Total de Lemas	**202**

3.07.03 DBM. Las marcas buscadas en este diccionario fueron muy similares a las buscadas en el *DRAE*. Esto resulta entendible si se toma en cuenta que este diccionario tiene como autor a Gómez de Silva (2001/2003), uno de los miembros de la Academia Mexicana, suscrita a la RAE. De este diccionario se buscaron las siguientes marcas: despect. (despectivo), irón. (irónico), malsonante, vulgar, eufemismo, e insulto. Al igual que se comentó respecto del *DRAE*, las marcas despect., irón., eufemismo e insulto no corresponden a la variación lingüística comúnmente considerada como parte del subestándar; sin embargo, su cercanía con el léxico tabuizado, así como sus escasas instancias en los diccionarios fuente, hizo que se incluyeran en este estudio. Cabe aclarar que algunas de estas marcas eran abreviaturas y otras eran datos desarrollados dentro de la definición. Además, ya que este diccionario registra exclusivamente voces del ámbito geográfico de México, no fue necesario buscar combinaciones de marcas como se hizo en el *DRAE.* Respecto del proceso de identificación de lemas marcados, desde un principio se comenzó a trabajar con la versión de Internet de este diccionario (Gómez de Silva, 2001/2003). Se utilizó el buscador de Explorer para identificar los lemas con marcas pertinentes y crear independientemente dos listas con los mismos. Las dos listas se cotejaron, se compararon con la información del diccionario impreso y se capturó la información revisada en una hoja de cálculo de Excel. Los totales por letra de esta búsqueda se presentan en la Tabla 3.4. Según se aprecia en la tabla, en el DBM se identificaron un total de 142 marcas en 141 entradas. Se encontraron quince marcas *despect*., dieciséis marcas *irón.*, 83 marcas *malsonante*, dos marcas *vulgar*, veintitrés marcas *eufemismo* y tres marcas *insulto.*

Tabla 3.4. Conteo por letra de los lemas con marcas del subestándar en el *DBM*

Letra	despectivo	irónico	malsonante	vulgar	eufemismo	insulto	Subtotal
A	1	1	-	-	3	1	6
B	-	1	-	-	-	-	1
C	1	2	9	1	3	-	16
CH	1	-	13	-	7	-	21
D	-	-	2	-	-	1	3
E	1	4	1	-	-	-	6
F	-	-	7	-	1	-	8
G	4	-	-	-	-	-	4
H	-	-	6	-	-	-	6
I	-	1	-	-	-	1	2
J	-	-	4	-	1	-	5
L	-	2	2	-	-	-	4
LL	-	-	-	-	-	-	-
M	2	-	24	1	2	-	29
N	-	-	-	-	-	-	-
Ñ	-	-	-	-	-	-	-
O	-	1	1	-	-	-	2
P	-	1	14	-	1	-	16
Q	-	-	-	-	-	-	-
R	2	1	-	-	1	-	4
S	1	1	-	-	-	-	2
T	1	-	-	-	3	-	4
U	-	-	-	-	1	-	1
V	-	1	-	-	-	-	1
Y	-	-	-	-	-	-	-
Z	1	-	-	-	-	-	1
Subtotal	15	16	83	2	23	3	Total
							142

Lemas con marcas dobles	**1**
Total de Lemas	**141**

3.07.04 DTV. Para este diccionario y los dos siguientes los procedimientos fueron inversos a los anteriores, pues más bien se buscaba eliminar las marcas jergales, las cuales se decidió excluir de la prueba de representatividad en el corpus por las razones ya antes comentadas (ver sección 3.02). Así, se comenzó contando las entradas. Después, se encontraron seis marcas a eliminar desarrolladas en las definiciones de los lemas de la siguiente manera: (1) jerga de la prostitución, (2) caló del hampa, (3) jerga carcelaria, (4) caló delincuente, (5) caló periodístico y (6) dominó a la mexicana. Las cuatro primeras, constituyen marcas jergales de grupos humanos marginales, mientras que la quinta es una marca jergal en un sentido distinto. Esta última marca puede considerarse una marca jergal laboral perteneciente a un grupo cuyos miembros comparten una profesión u oficio. La sexta marca resultó desconocida, y no se pudo encontrar su significado con ningún informante, por lo que se decidió también eliminarla. El total de entradas con marcas a eliminar fueron: una entrada con marca de *jerga de la prostitución*, 343 entradas con marca de *caló del hampa,* una entrada con marca de *jerga carcelaria*, una entrada con marca de *caló delincuente,* dos entradas con marca de *caló periodístico* y dos entradas con marca de *dominó a la mexicana.* La Tabla 3.5 muestra los totales por letra de entradas en el diccionario (2,307) y de entradas marcadas a eliminar (350), así como los totales de entradas a utilizar. Respecto de este último dato, para este estudio se emplearon 1,957 unidades léxicas o lemas de este diccionario.

Tabla 3.5. Conteo por letra de los lemas a utilizar del *DTV*

Letra	Entradas		
	totales	eliminadas	utilizadas
A	180	18	162
B	133	28	105
C	284	58	226
CH	136	20	116
D	121	18	103
E	146	13	133
F	64	10	54
G	82	10	72
H	63	9	54
I	25	1	24
J	50	4	46
K	4	1	3
L	71	8	63
LL	9	-	9
M	210	28	182
N	48	4	44
Ñ	7	1	6
O	31	4	27
P	247	38	209
Q	20	1	19
R	78	8	70
S	88	17	71
T	139	27	112
U	3	-	3
V	35	7	28
W	4	-	4
X	9	-	9
Y	6	1	5
Z	14	2	12
Total	**2307**	**336**	**1971**

3.07.05 PDLPM. Este diccionario presentó particularidades que ningún otro de los diccionarios fuente mostró. En la redacción de este diccionario no se lematizó, propiamente hablando, es decir, no se hizo la abstracción de la forma básica de la unidad léxica, que deja

de lado los cambios gramaticales o morfológicos de las misma (Biber, Conrad y Reppen, 1998, p. 29). Ciertamente en este diccionario aparecen lemas o encabezamientos para cada una de las entradas, pero estos son más bien frases en contexto. Así, para el que debería haber sido el lema *chingón, na*, aparecen un total de 286 frases en contexto, a las cuales se les dio una entrada, y cuyas definiciones son más bien enciclopédicas, es decir, definen la realidad no la palabra. Tomando en cuenta esta situación, de las 1,667 entradas que se contaron en este diccionario, se pudieron extraer tan sólo 34 formas lematizables, todas derivadas de la familia lexicogenésica del verbo *chingar*. Esto se debe a que además del caso de chingón, la forma *chingada*, incluida en dos locuciones adjetivas, *de la chingada* y *como la chingada*, aparece lematizada con distintos contextos 702 veces en todo el diccionario; la forma *chingadera* aparece lematizada en contexto 180 veces, el verbo *chingar* conjugado y en contexto aparece en 131 lemas o encabezados, el adjetivo *chingado, da* tiene 310 apariciones contextualizadas y lematizadas. La Tabla 3.6 muestra los resultados del conteo de entradas y el conteo de formas lematizables que se llevó a cabo en este diccionario. Como ya se dijo, sólo 34 formas lematizables se identificaron, con las cuales se creó una hoja de cálculo en Excel para proceder posteriormente a su exploración en el corpus. Hay que hacer notar que las formas lematizables son más bien reconstrucciones mías de los lemas potenciales que debería haber tenido este diccionario si se hubiesen seguido criterios lexicográficos en su elaboración.

Tabla 3.6. Conteo de entradas y formas lematizables en el *PDLPM*

Formas contadas	Totales
chingón (a)(es)(as)	286
de la chingada	402
como la chingada	300
chingaderas	180
chingar (conjugación)	131
chingo	14
chingado	138
chinga	10
chingonamente	8
chingada	172
hasta la chingada	26
chingadazo	12
historias y comentarios	24
otras formas	28
Total	1731
Formas múltiples que aparecen en una sola entrada	**64**
Total de entradas	**1667**
Formas lematizables de la familia lexicogenésica de chingar	**34**

3.07.06 AHDOM. Este diccionario realiza la lematización independiente de colocaciones, es decir de elementos léxicos que se combinan sin que exista entre ellos la exigencia de aparecer siempre juntos (Porto, 2002, p. 154). Por lo tanto, hay que considerar que muchas de las entradas que aparecen en su inventario léxico son colocaciones que deberían aparecer agrupadas en otras entradas del diccionario. Un ejemplo de lematización más convencional donde las colocaciones no son lematizadas de forma independiente lo

podemos encontrar en la primera definición de la entrada *allanar* de (Alvar, 1987). En esta entrada, después de la definición "Poner llana [una cosa]" es introducida con dos puntos y destacada en letra cursiva una serie de posibles colocaciones del verbo allanar, el cual es sustituido con una tilde nasal (~): "*~ unas cercas; ~ una piedra; ~ un monte; el terreno allana* o *se allana"* (p. 62). Por el contrario en el *AHDOM*, todas las posibles colocaciones de la entrada *aguantar* en su acepción de "ser paciente, tolerante" han sido lematizadas independientemente de ésta, de forma que de esta misma acepción nos encontramos con la serie de entradas siguiente: "**AGUANTAR**", "**AGUANTAR HORRORES**", "**AGUANTAR LA 'BARA'**", "**AGUANTAR LA RIATA**" y "**AGUANTAR LA VERGA**". Además de todas estas entradas para la acepción mencionada de *aguantar*, aparecen otras tantas entradas para otras posibles acepciones del mismo lema: "**AGUANTA**", "**AGUANTA UN PIANO**", **AGUANTADOR (SER)**", "**AGUÁNTAME TANTITO**", "**AGUANTE (QUÉ!)**" y "**AGUANTE (TENER MUCHO)**" (Colín Sánchez, 1987/2001, p. 24). Debido a este procedimiento de lematización excesivo, el conteo del inventario léxico del *AHDOM* aparenta ser muy prolijo. El conteo del total de entradas en este diccionario, que se tuvo que hacer manualmente, asciende a 9,613 entradas. El conteo por letra se presenta en la Tabla 3.7. No se encontraron marcas de ningún tipo que contribuyeran a la reducción de este inventario léxico, así que simplemente se creó una hoja de cálculo con el total de las entradas del diccionario que se incluyeron en la exploración del *CREA*.

Tabla 3.7. Conteo por letra del lemario del *AHDOM*

Letra	Entradas	**Letra**	Entradas
A	727	**N**	302
B	409	**Ñ**	17
C	1069	**O**	123
CH	454	**P**	1143
D	456	**Q**	106
E	546	**R**	356
F	233	**S**	507
G	321	**T**	647
H	252	**U**	63
I	96	**V**	209
J	156	**W**	4
K	7	**X**	12
L	335	**Y**	101
LL	42	**Z**	57
M	862		
Total	**9612**		

3.07.07 DIME. Aunque al listar los distintos diccionarios fuente se colocó a este diccionario después de los dos generales incluidos en este estudio, el *DEUM* y el *DRAE,* la extracción de los lemas del *DIME* se comenta hasta este momento debido a que éste fue el último diccionario hallado e integrado en este trabajo. En cuanto a la labor de extracción misma, en este diccionario, dedicado a las "niñas, niños y jóvenes mexicanos" (Ávila y Aguilar, 2003, p. v), se emplearon los lemas con las marcas siguientes: "uso coloquial", "uso popular", "uso ofensivo" y "uso grosero". Revisando el diccionario entrada por entrada, se identificaron dichas marcas y se creó al mismo tiempo una lista electrónica de los lemas marcados. Este procedimiento se llevó a cabo dos veces de manera

independiente. Las dos listas obtenidas se cotejaron entre sí y se verificaron las discrepancias contra el diccionario mismo. En cuanto a la identificación de marcas, al igual que sucedió con el DEUM y con el DBM, en este diccionario la marcación geográfica de "México" es más bien implícita. Esto encuentra su justificación en la acotación contenida en la introducción de este diccionario respecto de los destinatarios del mismo, que son las "niñas, niños y jóvenes mexicanos" (Ávila y Aguilar, 2003, p. v). Los resultados de la búsqueda de marcas se presentan en la Tabla 3.8 que aparece a continuación. En este diccionario se encontraron un total de 651 marcas "uso coloquial", 126 marcas "uso popular", 39 marcas "uso ofensivo" y 71 marcas "uso grosero". Con ello se identificaron un total de 887 marcas distribuidas en 821 acepciones. Aquí cabe señalar que en la composición de la lista de lemas a explorar en el corpus, en este diccionario se decidió hacer una lista más bien de acepciones que de lemas. Esta última decisión se tomó debido a los resultados de la exploración de otros diccionarios, como el *AHDOM* y el *PDLPM*, que muestran que la lematización de éstos es poco confiable. Por ello, se pensó que sería mejor trabajar con acepciones, sobretodo porque, en una posible elaboración posterior del diccionario cuya ejemplificación ha sido proyectada aquí, la mayoría de las entradas del *AHDOM*, que representan alrededor de tres cuartas partes del total de las fuentes secundarias, tendrían que ser lematizadas nuevamente.

Tabla 3.8. Conteo por letra de los lemas con marcas del subestándar en el *DIME*

Letra	Coloquial	Popular	Ofensivo	Grosero	Subtotal
A	45	4	-	-	49
B	26	1	-	-	27
C	85	34	5	17	141
CH	41	13	13	17	84
D	17	1	1	-	19
E	24	1	-	1	26
F	14	7	-	1	22
G	27	7	1	-	35
H	18	4	2	2	26
I	6	-	-	-	6
J	16	7	5	5	33
L	26	5	-	2	33
LL	2	-	-	-	2
M	68	10	2	12	92
N	2	1	-	-	3
Ñ	2	1	-	-	3
O	27	-	2	2	31
P	83	9	6	10	108
Q	9	-	-	-	9
R	24	5	-	1	30
S	28	6	-	-	34
T	33	9	1	-	43
U	4	-	-	-	4
V	18	1	1	1	21
Y	2	-	-	-	2
Z	4	-	-	-	4
Subtotal	651	126	39	71	Total
					887

Marcas dobles	**56**
Marcas triples	**5**
Total de Lemas	**821**

Una vez comentados los procedimientos para la extracción de las fuentes secundarias de este estudio, la Tabla 3.9 muestra finalmente el total de entradas o artículos de cada uno de los diccionarios, el total de materiales léxicos extraídos de los mismos (que constituyen las fuentes secundarias de este trabajo) y el tamaño requerido para obtener una muestra aleatoria representativa de cada uno de dichos subconjuntos del léxico del subestándar. Este último dato esta basado en lo que dicen Krejcie y Morgan respecto del número de elementos mínimos a incluir en una muestra para que ésta sea representativa de la población de que se le extrae (como se cita en Gay y Airasian, 2002, p. 113).

Tabla 3.9. Fuentes secundarias del léxico subestándar del español en México

Diccionario	**Total de**		**Entradas utilizadas**	**Tamaño requerido de la muestra**
	entradas	acepciones		
DEUM	14,000 (aprox)	60,000 (aprox)	568	226
DIME	13,000 (aprox)	22,000 (aprox)	821	163
DRAE	88,431	190,581	202	132
	Total de entradas			
DBM	6, 200 (aprox)		141	103
DTV	2,307		1,971	321
PDLPM	1,667		34	32
AHDOM	9,613		9,612	370
	Lista fusionada		**13349**	**1347**

En general para Krejcie y Morgan, las muestras para poblaciones con 100 elementos o menos se deberían acercar al total de la población, alrededor de los 500 elementos deberían contener aproximadamente el 50%, en torno a los 1,500 elementos se debería

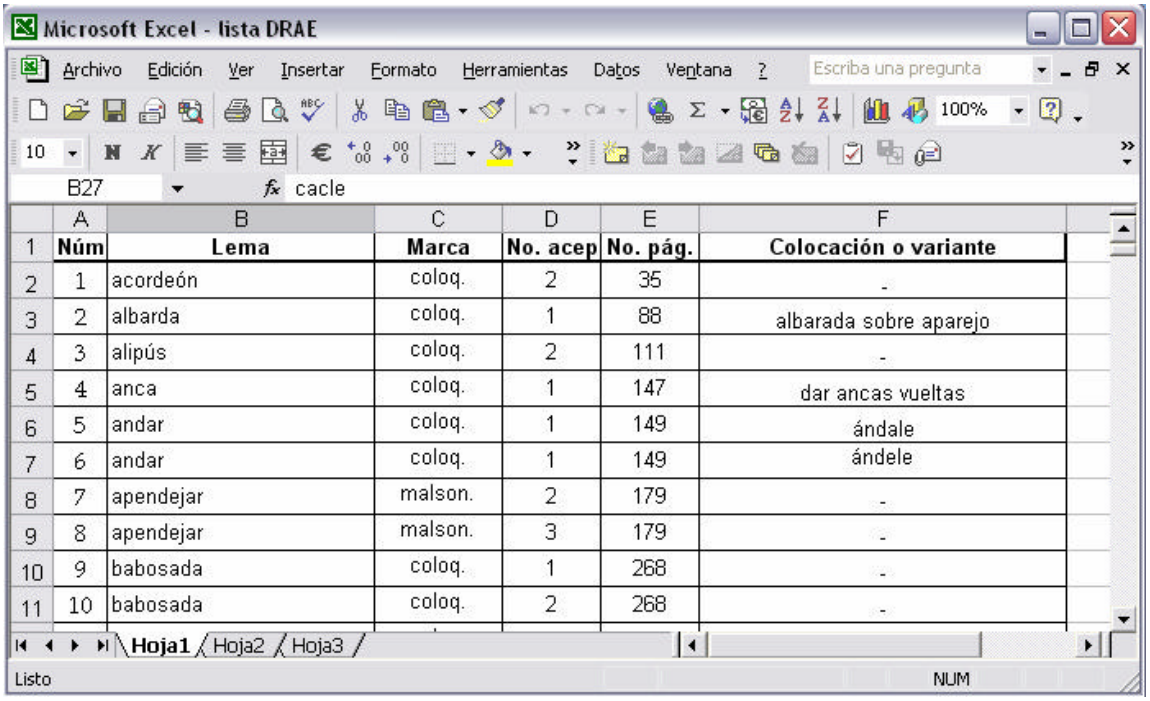
incluir más o menos el 20%, y por arriba de los 5,000 elementos en la población la muestra escasamente tendrá que rebasar los 400 elementos. Ahora bien, más allá de estas pautas generales para determinar la representatividad de un muestreo, el total de elementos que se incluyó en este trabajo respecto de la muestra de cada diccionario se determinó por medio de la utilización de la tabla, mucho más precisa, que proporcionan Krejcie y Morgan para poblaciones de diversos tamaños (como se cita en Gay y Airasian, 2002, p. 113). Así podemos ver que de todos los diccionarios fuente, de los cuales se extrajo un total de 13,349 entradas, se tuvo que hacer un muestreo total mínimo de 1,347 elementos para explorar la representación de las fuentes secundarias en el corpus. Para obtener el muestreo aleatorio representativo de los diccionarios, se ha aplicado un programa de cómputo de estadística en forma independiente al subconjunto del léxico subestándar de cada uno de los mismos. Debido a que la aplicación se ha hecho de manera separada para cada diccionario, se puede decir que el tipo de muestreo que se ha utilizado en este trabajo es un muestreo aleatorio estratificado. Este tipo de muestreo se ha utilizado con la finalidad de explorar la representatividad del *CREA* respecto de cada uno de los diccionarios de manera independiente. Con ello se espera poder hacer inferencias más certeras sobre el comportamiento potencial de toda la población que conforma las fuentes secundarias del léxico subestándar del español en México.

3.08 Manejo de las listas de las fuentes secundarias

La primera etapa en el manejo de la información extraída de las fuentes secundarias fue la extracción misma de los datos pertinentes para su documentación en el *CREA*. Para esta extracción, comentada de manera particular para cada diccionario en la sección anterior, se decidió crear listas en hojas de cálculo de Excel (ver sección 3.07). Estas

primeras listas que se crearon se nombraron con la palabra *lista* seguida de las siglas del diccionario de procedencia. Aunque los conceptos a manejarse en cada lista variaron de acuerdo al diccionario mismo del cual se extrajo la información, había algunos datos comunes a todas las listas. Éstos fueron: un número progresivo en la lista del lema o acepción que se extrajo del diccionario, el lema o encabezado de la entrada y el número de la página en la cual se encontraba el lema. La Figura 3.1 muestra un ejemplo de la *lista DRAE*, donde se puede ver que el número progresivo del lema extraído corresponde a la primera columna, o columna A, de la hoja de cálculo, el lema o encabezado se encuentra en la segunda columna, o columna B, y el número de página en el cual se encontró el lema dentro del diccionario aparece en la quinta columna, o columna E.

Figura 3.1. La *lista DRAE* en la primera etapa de trabajo con las fuentes secundarias



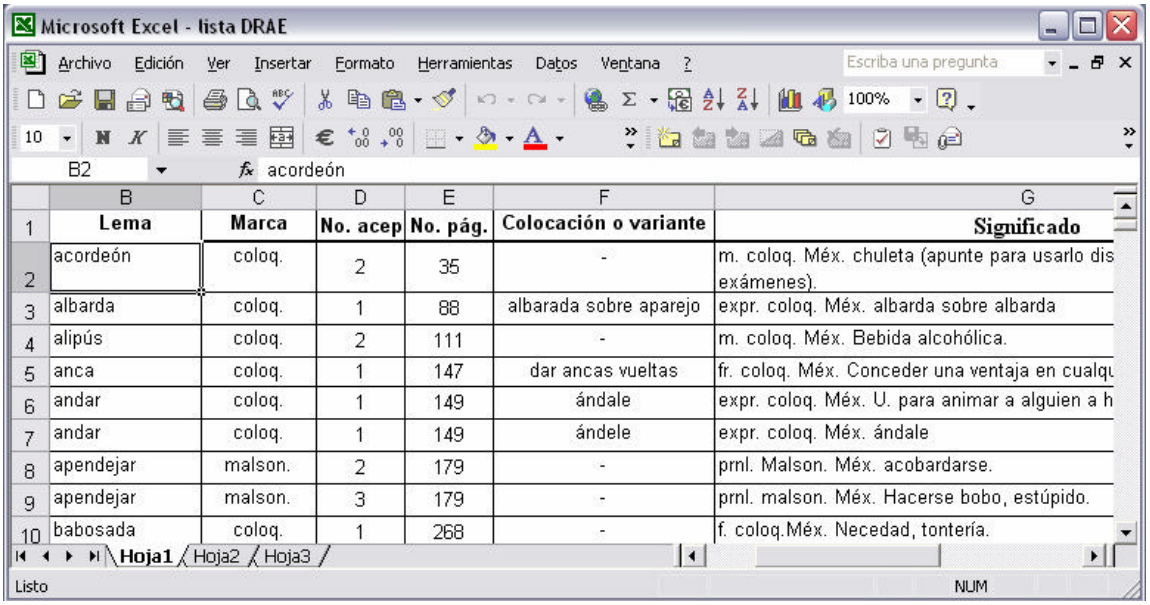
	A	B	C	D	E	F
1	**Núm**	**Lema**	**Marca**	**No. acep**	**No. pág.**	**Colocación o variante**
2	1	acordeón	coloq.	2	35	-
3	2	albarda	coloq.	1	88	albarada sobre aparejo
4	3	alipús	coloq.	2	111	-
5	4	anca	coloq.	1	147	dar ancas vueltas
6	5	andar	coloq.	1	149	ándale
7	6	andar	coloq.	1	149	ándele
8	7	apendejar	malson.	2	179	-
9	8	apendejar	malson.	3	179	-
10	9	babosada	coloq.	1	268	-
11	10	babosada	coloq.	2	268	-

Además, a las listas de la mayoría de los diccionarios, de hecho a todas ellas salvo a la del *DEUM* que fue el primer diccionario en trabajarse, se les añadió una columna llamada de "Colocación o Variante", donde se incluyeron varias de las palabras que acompañaban al lema. Esto se hizo sobre todo en el caso de las unidades pluriverbales más extensas. En la Figura 3.1 se puede ver que esta columna corresponde a la sexta de la *lista DRAE*. En esta columna se capturó también, en el caso de la *lista DBM*, algunas variantes no lematizadas que ofrecía el mismo diccionario inmediatamente después del encabezado de la entrada propiamente dicho. A la *lista DBM*, a la *lista DEUM*, a la *lista DRAE* y a la *lista DIME* se les añadió una columna con la marca que dichos diccionarios colocaban a sus lemas. En este respecto, las listas del *DEUM*, del *DRAE* y del *DIME* contienen también una columna con el número de acepción o acepciones afectadas por dichas marcas. En la Figura 3.1 se observa que la tercera columna de la *lista DRAE*, o columna C, contiene las marcas de los lemas que se encuentra en las segunda columna, o columna B. Por su parte, en la cuarta columna, o columna D, aparecen las acepciones afectadas por las marcas de la tercera columna. Así por ejemplo, si vemos el lema *alipús*, con el número progresivo tres en la *lista DRAE*, podemos saber que está considerado como una voz coloquial cuando se le utiliza con el segundo significado de los que lista el *DRAE*, que en este caso corresponde al de "bebida alcohólica (RAE, 2001, p.111). De manera distintiva, la *lista DTV* fue construida con una columna exclusiva para sus variantes, independiente a la columna de colocación. Esto último se debió sobre todo al gran número de variantes no lematizadas, mucho más numerosas que aquellas presentadas por el *DBM*, que el *DTV* ofrece respecto de algunas de sus entradas. Finalmente, la *lista PDLPM* tiene una columna con el número de apariciones de los lemas. Hay que recordar que este diccionario no ofrecía en muchos casos unidades léxicas lematizadas, propiamente hablando, sino una serie arbitraria de

concordancias o apariciones en varios contextos de un mismo lema. Así pues, esta columna muestra el número de veces que un lema, que más bien fue reconstruido por el autor de este trabajo para aparecer tal cual en la lista, se presentó en todo el diccionario. Esta columna buscaba también justificar la falta de información en la columna del número de página del lema, que en el caso de los lemas reconstruidos no podía ser llenada con ningún contenido. Todas estas listas fueron posteriormente ampliadas con una columna para el significado del lema o encabezado.

En una segunda etapa del manejo de la información extraída de las fuentes secundarias, se decidió añadirle a todas las listas el significado de las entradas que las conformaron. Las listas resultantes tenían el mismo nombre de las anteriores, pero fueron ubicadas en una nueva carpeta. La Figura 3.2 es un ejemplo de la *lista DRAE* con el significado incluido en la séptima columna, o columna G.

Figura 3.2. La *lista DRAE* en la segunda etapa de trabajo con las fuentes secundarias



	B	C	D	E	F	G
1	**Lema**	**Marca**	**No. acep**	**No. pág.**	**Colocación o variante**	**Significado**
2	acordeón	coloq.	2	35	-	m. coloq. Méx. chuleta (apunte para usarlo dis exámenes).
3	albarda	coloq.	1	88	albarada sobre aparejo	expr. coloq. Méx. albarda sobre albarda
4	alipús	coloq.	2	111	-	m. coloq. Méx. Bebida alcohólica.
5	anca	coloq.	1	147	dar ancas vueltas	fr. coloq. Méx. Conceder una ventaja en cualqu
6	andar	coloq.	1	149	ándale	expr. coloq. Méx. U. para animar a alguien a h
7	andar	coloq.	1	149	ándele	expr. coloq. Méx. ándale
8	apendejar	malson.	2	179	-	prnl. Malson. Méx. acobardarse.
9	apendejar	malson.	3	179	-	prnl. malson. Méx. Hacerse bobo, estúpido.
10	babosada	coloq.	1	268	-	f. coloq.Méx. Necedad, tontería.

La finalidad del trabajo de captura del significado fue ante todo el hacer de las listas materiales más independientes y manejables en el momento de la documentación. El tener prácticamente toda la información en la misma lista a documentar permitía trabajar más rápidamente con las hojas de cálculo e Internet, sin necesidad de recurrir a los diccionarios de papel de mucho más lento acceso. La única modificación mayor que implicó la captura del significado en las listas originales fue la partición de la *lista AHDOM* en dos listas, la *lista AHDOM 01* y la *lista AHDOM 02*. Esta división se llevó a cabo debido a que el tamaño de una sola lista con significados para las 9,612 entradas utilizadas de este diccionario volvía al archivo electrónico demasiado pesado y difícil de desplegar y manipular. Cabe aclarar que la ayuda proporcionada por los becarios en la construcción de las listas de las fuentes secundarias (ver sección 3.07) sólo tuvo lugar hasta aquí. El resto del trabajo relacionado con el manejo de las listas, y del que se hablará a continuación, corrió a cargo únicamente del autor de esta tesis.

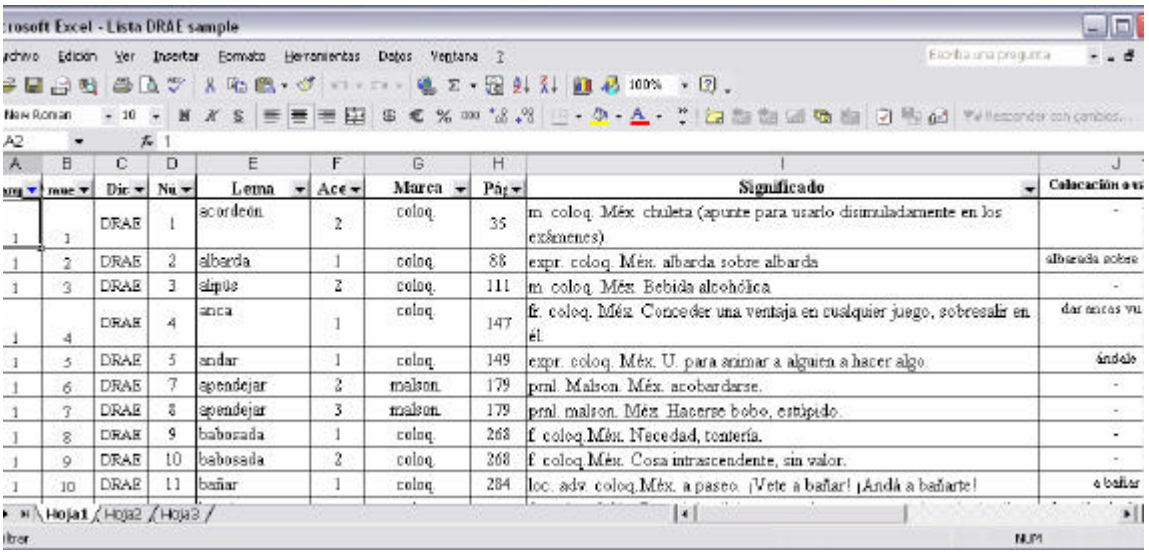
3.09 Organización del muestreo aleatorio estratificado

La tercera etapa en el manejo de las listas de las fuentes secundarias constituye el último paso de preparación de dichas listas antes de comenzar con el trabajo de documentación en el corpus electrónico. En esta etapa se llevó a cabo la creación de las listas del muestreo aleatorio. Al nombre de estas nuevas listas se le agregó únicamente la palabra *sample* para diferenciarlas de las anteriores. Los nombres de las nuevas listas quedaron, por tanto, de la siguiente manera: *lista DEUM sample*, *lista DRAE sample*, *lista DBM sample*, y así sucesivamente. Cabe aclarar de antemano que estas nuevas listas requirieron una vez más dividir la información de la *lista AHDOM*, ahora en tres listas: la *lista AHDOM 01 sample*, la *lista AHDOM 02 sample* y la *lista AHDOM 03 sample*. Todo

ello con la finalidad de manejar más fácilmente la información de este diccionario. También se debe mencionar que durante la confección de las listas sample, anterior al trabajo de documentación, se estandarizó el formato de las mismas. Así, se introdujeron en ellas columnas para todos los posibles datos que presentaban los distintos diccionarios, quedando las listas sample de la manera que se describe a continuación. Primero se colocó una columna titulada "sample", donde se indicó si la fila en cuestión había sido seleccionada o excluida por el programa de cómputo que llevó a cabo el muestreo aleatorio. Esta columna corresponde a la columna A de la *lista DRAE sample*, en la Figura 3.3. Esta primera columna estaba seguida de una columna donde se anotó un número progresivo para los lemas seleccionados durante el muestreo aleatorio, que en la Figura 3.3 se observa en la columna B. A estas dos columnas se le sumaron los siguientes datos en sendas columnas y en el orden aquí enunciado: las siglas del diccionario a que pertenecieran los lemas, el número progresivo del lema o encabezado en la lista general de lemas extraídos de dicho diccionario, el lema o encabezado en sí, el número de acepción del lema utilizado, la marca correspondiente, el número de página de procedencia del lema, el significado del lema o entrada y la colocación o variante. Todos estos datos aparecen, siguiendo el mismo orden, en las columnas C a la J de la Figura 3.3. Así por ejemplo, si tomamos el lema *bañar* de la *lista DRAE sample* en dicha figura, nos encontramos que en su columna A contiene un número uno, lo cual quiere decir que este lema fue elegido para ser parte de la muestra. Vemos también que su número progresivo en la muestra es el número nueve (columna B), que el diccionario al que pertenece es el *DRAE* (columna C), que su número progresivo en la extracción de lemas del diccionario fue el número diez (columna D), que su lema es *bañar* (columna E), que la acepción de este lema que es parte del subestándar es la primera del diccionario (columna F), que dicha acepción es parte del subestándar porque está

considerada como coloquial (columna G), que el lema se encuentra en la página 284 del *DRAE* (columna H), que de acuerdo con este diccionario significa lo mismo que “a paseo” (columna I), y que comúnmente aparece en el contexto “a bañar”, como cuando se dice “vete a bañar” (columna J).

Figura 3.3. La *lista DRAE sample* en la tercera etapa de trabajo con las fuentes secundarias



				Lema		Marca	Pág	Significado	
1	1	DRAE	1	acordeón	2	coloq.	35	m. coloq. Méx. chuleta (apunte para usarlo disimuladamente en los exámenes)	-
1	2	DRAE	2	albarda	1	coloq.	88	expr. coloq. Méx. albarda sobre albarda	albarda sobre
1	3	DRAE	3	alipús	2	coloq.	111	m. coloq. Méx. Bebida alcohólica	-
1	4	DRAE	4	anca	1	coloq.	147	fr. coloq. Méx. Conceder una ventaja en cualquier juego, sobresalir en él.	dar ancas en
1	5	DRAE	5	andar	1	coloq.	149	expr. coloq. Méx. U. para animar a alguien a hacer algo.	ándale
1	6	DRAE	7	apendejar	2	malson.	179	prnl. Malson. Méx. acobardarse.	-
1	7	DRAE	8	apendejar	3	malson.	179	prnl. malson. Méx. Hacerse bobo, estúpido.	-
1	8	DRAE	9	babosada	1	coloq.	268	f. coloq.Méx. Necedad, tontería.	-
1	9	DRAE	10	babosada	2	coloq.	268	f. coloq.Méx. Cosa intrascendente, sin valor.	-
1	10	DRAE	11	bañar	1	coloq.	284	loc. adv. coloq.Méx. a paseo. ¡Vete a bañar! ¡Andá a bañarte!	a bañar

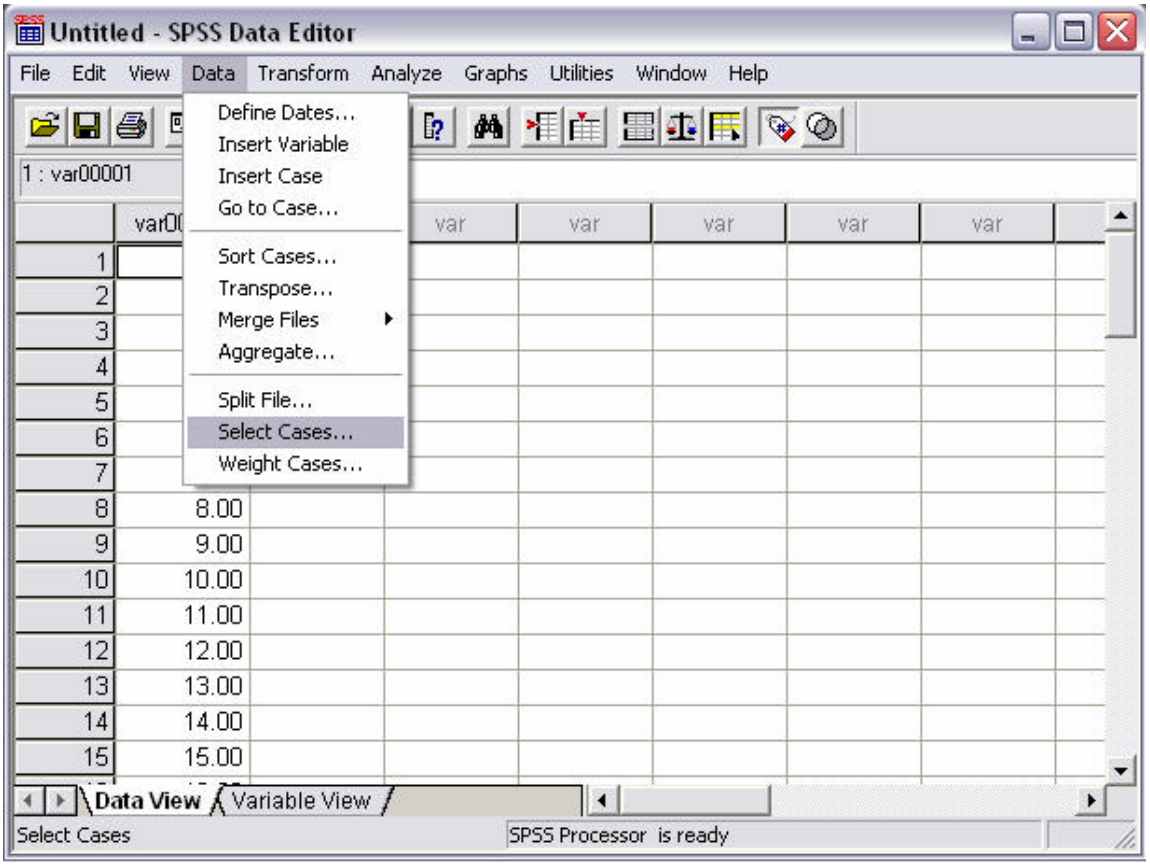
Además, se debe comentar que en las listas sample se añadieron cinco columnas, a llenar durante el proceso de documentación: una columna para anotar la formula de búsqueda en el *CREA*, una columna para anotar el resultado de dicha búsqueda, y tres columnas para indicar si, al analizar las concordancias arrojadas por el corpus, el lema quedaba sin ejemplo, con dos ejemplos o con un sólo ejemplo. Estas columnas aparecen en las listas sample de la columna K a la columna O; sin embargo, en la Figura 3.3 no se aprecian pues no tendría sentido mostrarlas aquí ya que durante esta etapa estas columnas

no fueron llenadas con ninguna información. Finalmente, hay que decir que las distintas columnas fueron igualadas en su ancho y en el tipo y tamaño de letra de su contenido en las listas de todos los diccionarios.

Ahora bien, con la finalidad de llevar a cabo un muestreo aleatorio estratificado del lemario tentativo obtenido de todas las fuentes existentes del léxico subestándar del español en México, se tuvo que hacer una selección aleatoria de los elementos mínimos para la obtención de muestras representativas de los distintos diccionarios. El tamaño de las muestras representativas se determinó, según se indicó arriba, siguiendo las recomendaciones de Krejcie y Morgan, como se citan en Gay y Airasian (2002). La selección aleatoria en sí se hizo con la ayuda del programa de cómputo para estadística SPSS para Windows, en su versión 10.0.1 Estándar (1999).

Para hacer la selección aleatoria se hizo lo siguiente con todas las listas. Se comenzó corriendo el programa SPSS para Windows. En la ventana de bienvenida se inició seleccionando la opción “Type in data” ante la pregunta “What would you like to do?”. Esta primera instrucción abre una hoja de datos, como la que se muestra en la Figura 3.4. Para comenzar a introducir datos en la hoja nueva de SPSS para Windows, se hizo un copiado de la columna en la hoja de Excel que contenía el número progresivo de los lemas extraídos en la lista general del diccionario en cuestión. Esta columna se pegó en la primera columna de variantes “var” de la hoja de datos de SPSSS para Windows, la cual después del copiado recibe automáticamente el nombre de “var00001”. A continuación se seleccionó la opción “Select cases” mostrada al hacer clic sobre la opción “Data” de la barra del menú principal. Las operaciones hasta aquí descritas se pueden ver también en la misma Figura 3.4.

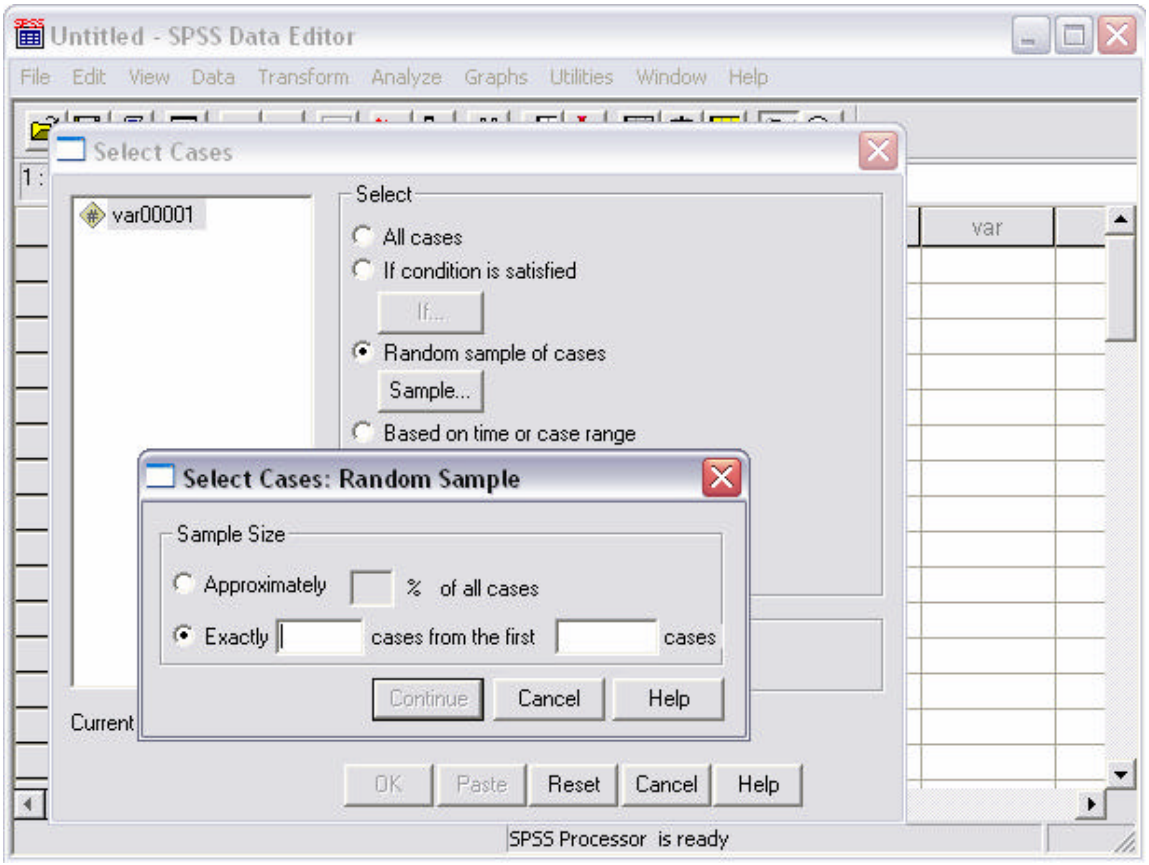
Figura 3.4. Hoja de datos del programa SPSS para Windows



Después de seleccionar la opción “Select Cases” en el menú “Data” de la barra de menú principal, el programa arroja una ventana llamada igualmente “Select cases”, la cual mostraba ya seleccionada la columna “var00001”. Esta nueva ventana aparece en la Figura 3.5. En esta ventana, se seleccionó la opción “Random sample of cases” para obtener un muestreo aleatorio de dicha columna, exportada desde la lista de los diccionarios. Antes de llevar a cabo el muestreo aleatorio en sí, se hizo clic en el botón “Sample” habilitado por la selección de la opción “Random sample of cases”, para desplegar la ventana “Select cases: Random Sample”. Esta ventana ofrece la posibilidad de obtener un muestreo con un

porcentaje preciso del total de la población, en su opción “Aproximately”, o un muestreo con un número específico de casos de entre un número determinado de casos iniciales de la columna seleccionada, en su opción “Exactly…”. Para el muestreo requerido en este trabajo, se optó por la segunda opción. Todas estas posibles opciones de SPSS para Windows en la obtención de una muestra aleatoria se pueden visualizar también en la Figura 3.5. Respecto del tamaño de la muestra solicitado en la opción “Exactly…”, se indicó primero el número de casos necesarios para obtener una muestra representativa de la lista del diccionario en cuestión, siguiendo las recomendaciones de Krejcie y Morgan (como se citan en Gay y Airasian, 2002), y se pidió su extracción del total de casos de la lista del diccionario. Una vez introducidos estos dos datos se hizo clic en el botón “Continue” que cierra la última ventana “Select cases: Random Sample”, mostrada igualmente en la Figura 3.5. Después se hizo clic en el botón “OK” de la ventana “Select cases”. Como resultado de esta operación el programa SPSS para Windows generaba una columna después de la columna “var00001”, llamada “filter_$”, donde asignaba un número uno a los casos seleccionados y un cero a los casos eliminados del muestreo. La columna “filter_$” del programa SPSS para Windows se copió a la columna titulada “sample” de las listas del muestreo aleatorio en Excel.

Figura 3.5. Opciones para la obtención de un muestreo aleatorio en SPSS para Windows



Finalmente, en las listas sample que contenían la columna con la selección aleatoria del programa SPSS para Windows, se aplicó la función “Autofiltro” de Excel. Esta función se activó eligiendo la opción “Filtro” mostrada en el menú “Datos” de la barra de menú principal. Con esta opción se pudo desplegar una lista que contuviera únicamente aquellas filas que en su columna “sample” contuvieran un número uno, es decir, aquellas que hubiesen sido seleccionadas por el programa de estadística. El resultado de esta operación

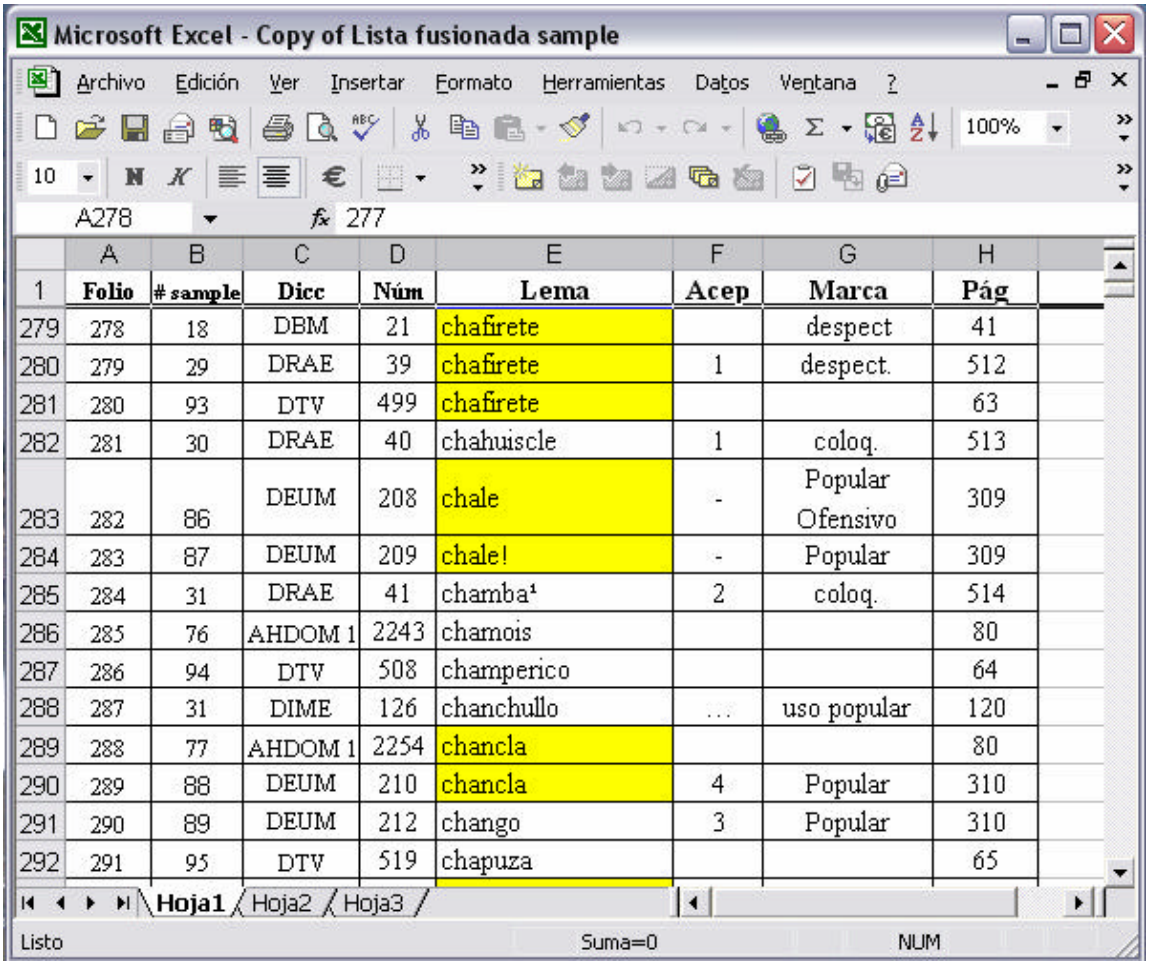
se puede ya apreciar desde la Figura 3.3, donde los números progresivos de la muestra (columna B) no coinciden con los números progresivos de la extracción de lemas (columna D), debido a que de los lemas extraídos sólo están desplegados aquellos seleccionados por el programa SPSS para Windows. En este respecto, cabe hacer notar que en la columna B, donde se anotó el número del lema seleccionado para la muestra, éste se introdujo manualmente, puesto así lo requería la lista una vez que se le aplicó la función "Autofiltro". Esta última numeración se llevó a cabo para corroborar que el tamaño de la muestra representativa arrojada en la columna "filter_$" de la hoja de datos de SPSS Windows coincidiese efectivamente con el número de filas mostradas por el "Autofiltro" de Excel. En este sentido, no hubo ninguna discrepancia. Todo este procedimiento, como se dijo anteriormente, se realizó de manera individual para las distintas listas de los diccionarios fuente. Con esto se terminó la preparación de las listas de las fuentes secundarias.

3.10 Documentación de las listas del muestreo

Una vez preparadas las listas definitivas para llevar a cabo la documentación, me preocupó la posibilidad de tener casos en que algunos lemas con las mismas acepciones se presentaran en diferentes diccionarios, y que por ello, la documentación de dichos lemas se llevara a cabo de forma múltiple y repetitiva. Para evitar esto, se construyó una lista con los lemas seleccionados durante el muestreo aleatorio procedentes de todos los diccionarios. Esta lista fue sorteada por orden alfabético para poder detectar las posibles coincidencias de entradas entre las distintas muestras. En esta lista, llamada *lista fusionada sample* y presentada en la Figura 3.6, se incluyeron sólo los datos necesarios para la identificación de los lemas repetidos en la muestra. Así las columnas de esta lista contenían los siguientes datos: una numeración progresiva general de los lemas que compusieron la lista fusionada

misma (columna A), el número del lema seleccionado durante el muestreo aleatorio de acuerdo a cada diccionario (columna B), el nombre del diccionario del que procediese el lema en cuestión (columna C), el número del lema o encabezado en la lista general de lemas extraídos de dicho diccionario (columna D), el lema (columna E), el número de acepción del lema (columna F), la marca del mismo (columna G), y el número de página en que se halló el lema dentro del diccionario (columna H). Cuando se conformó la *lista fusionada sample* con estos datos, se sombrearon las celdas en la columna de los lemas donde había coincidencias entre los mismos. Así por ejemplo, en la parte de la *lista fusionada simple* que se muestra en la Figura 3.6, se puede ver que el lema *chafirete*, que se encuentra sombreado, está presente en varios diccionarios, el *DBM*, el *DRAE* y el *DTV*. De la *lista fusionada simple* se imprimió una copia física para consultar durante la documentación individual de las listas sample.

Figura 3.6. La *lista fusionada sample* utilizada durante la documentación



Microsoft Excel - Copy of Lista fusionada sample

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

A278 277

	A	B	C	D	E	F	G	H
1	**Folio**	**# sample**	**Dicc**	**Núm**	**Lema**	**Acep**	**Marca**	**Pág**
279	278	18	DBM	21	chafirete		despect	41
280	279	29	DRAE	39	chafirete	1	despect.	512
281	280	93	DTV	499	chafirete			63
282	281	30	DRAE	40	chahuiscle	1	coloq.	513
283	282	86	DEUM	208	chale	-	Popular Ofensivo	309
284	283	87	DEUM	209	chale!	-	Popular	309
285	284	31	DRAE	41	chamba[1]	2	coloq.	514
286	285	76	AHDOM 1	2243	chamois			80
287	286	94	DTV	508	champerico			64
288	287	31	DIME	126	chanchullo	...	uso popular	120
289	288	77	AHDOM 1	2254	chancla			80
290	289	88	DEUM	210	chancla	4	Popular	310
291	290	89	DEUM	212	chango	3	Popular	310
292	291	95	DTV	519	chapuza			65

Hoja1 Hoja2 Hoja3

Listo Suma=0 NUM

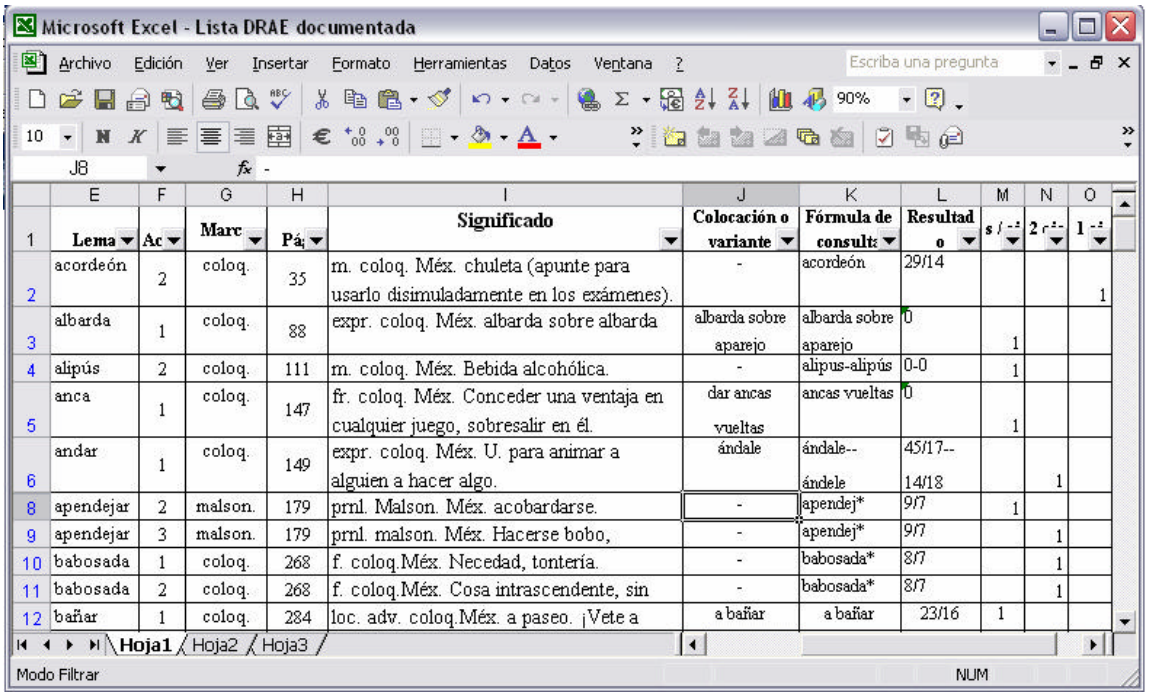
Para llevar a cabo la documentación en sí, se siguieron los pasos enunciados a continuación. Primero se consultó la *lista fusionada sample* y se identificaron las coincidencias de lemas entre distintas listas. Esto se llevó a cabo con la finalidad de hacer una sola búsqueda en el corpus electrónico para los lemas comunes y generar posibles fichas de ejemplo compartidas. Así, si se encontraban coincidencias entre lemas, se procedía a verificar si la coincidencia se hallaba también a nivel del significado. Si la coincidencia se encontraba sólo a nivel del lema y no del significado se podía generar por lo menos una fórmula de búsqueda común y revisar las concordancias del corpus teniendo

en cuenta los varios significados presentes en las distintas listas. Si la coincidencia se daba también a nivel del significado se podía incluso generar fichas de ejemplos comunes para varios casos en distintas listas. La coincidencia de significado se corroboró con las listas *sample* completas o directamente con los ejemplares de los diccionarios. En los casos en que se detectaron coincidencias de lemas y significados, se procedió a escribir una definición simplificada al margen de la *lista fusionada sample*, para tener en cuenta dicha definición durante el análisis de las concordancias. Una vez que se llevó a cabo esta revisión de las coincidencias entre distintas listas sample se podía llevar a cabo una primera búsqueda de los lemas similares, con fórmulas y significados comunes.

Tanto para los casos similares como para los que no lo eran, lo primero que se hizo, en el proceso mismo de documentación, fue desplegar la lista sample en la que se estuviera trabajando y salvarla con un nuevo nombre para guardar la información reunida durante este proceso. De esta manera, las listas pasaron a ser renombradas con la palabra *lista*, más el nombre del diccionario al que pertenecieran, y la palabra *documentada*. Con la nueva nominación, las listas adquirieron nombres tales como *lista DEUM documentada*, *lista DRAE documentada*, *lista DBM documentada*, y así sucesivamente. En estas listas se identificaba en primer lugar el lema del que se buscarían ejemplos de uso en el corpus. Después, se anotaba la fórmula o las fórmulas de búsqueda que se consideraba que pudiesen desplegar más eficientemente las concordancias pertinentes para las distintas variaciones morfológicas del lema buscado. Aquí hay que recordar que las concordancias son las listas de ejemplos de uso que proporcionan las computadoras que manejan las grandes bases de datos llamadas corpus electrónicos (Sinclair, 1985, p. 83). Para algunos lemas se utilizó más de una fórmula de búsqueda, y algunos de ellos alcanzaron a tener más de cinco fórmulas distintas. La Figura 3.7 muestra un ejemplo de las últimas columnas de la

lista DRAE documentada. En esta figura se puede ver cómo del lema *andar* y del lema *alipús* se utilizaron dos fórmulas de búsqueda distintas, mientras que del resto de los lemas que aparecen en esta figura se utilizó solamente una fórmula.

Figura 3.7. Llenado de la *lista DRAE documentada*



	E	F	G	H	I	J	K	L	M	N	O
1	Lema	Ac	Marc	Pá	Significado	Colocación o variante	Fórmula de consult	Resultado	s /	2	1
2	acordeón	2	coloq.	35	m. coloq. Méx. chuleta (apunte para usarlo disimuladamente en los exámenes).	-	acordeón	29/14			1
3	albarda	1	coloq.	88	expr. coloq. Méx. albarda sobre albarda	albarda sobre aparejo	albarda sobre aparejo	0	1		
4	alipús	2	coloq.	111	m. coloq. Méx. Bebida alcohólica.	-	alipus-alipús	0-0	1		
5	anca	1	coloq.	147	fr. coloq. Méx. Conceder una ventaja en cualquier juego, sobresalir en él.	dar ancas vueltas	ancas vueltas	0	1		
6	andar	1	coloq.	149	expr. coloq. Méx. U. para animar a alguien a hacer algo.	ándale	ándale--ándele	45/17--14/18		1	
8	apendejar	2	malson.	179	prnl. Malson. Méx. acobardarse.	-	apendej*	9/7	1		
9	apendejar	3	malson.	179	prnl. malson. Méx. Hacerse bobo,	-	apendej*	9/7		1	
10	babosada	1	coloq.	268	f. coloq.Méx. Necedad, tontería.	-	babosada*	8/7		1	
11	babosada	2	coloq.	268	f. coloq.Méx. Cosa intrascendente, sin	-	babosada*	8/7		1	
12	bañar	1	coloq.	284	loc. adv. coloq.Méx. a paseo. ¡Vete a	a bañar	a bañar	23/16	1		

Después, de generar un fórmula de búsqueda para el lema en cuestión, ésta se anotaba en la lista documentada y se introducía en el buscador de la pantalla principal del *CREA* (esta pantalla aparece más adelante en el segundo elemento de la Figura 3.8). Cabe aclarar que al corpus electrónico aquí utilizado, el *CREA*, se puede acceder a través de Internet en la página principal de la RAE (disponible en http://www.rae.es). Para acceder al *CREA* se debe hacer clic en el hipervínculo del lado izquierdo de la página llamado "Consulta banco de datos", que despliega en su parte inferior dos opciones: "Corpus actual"

y “Corpus histórico”. La primera de estas opciones es la que da acceso directo a la pantalla de consulta del *CREA*. La página principal de la RAE, con los hipervínculos mencionados, se muestra en el primer elemento de la Figura 3.8. Como se mencionó recién, en un segundo elemento de esta misma figura se ha incluido la pantalla del buscador del *CREA*.

Figura 3.8. Página principal de la RAE y del *CREA*

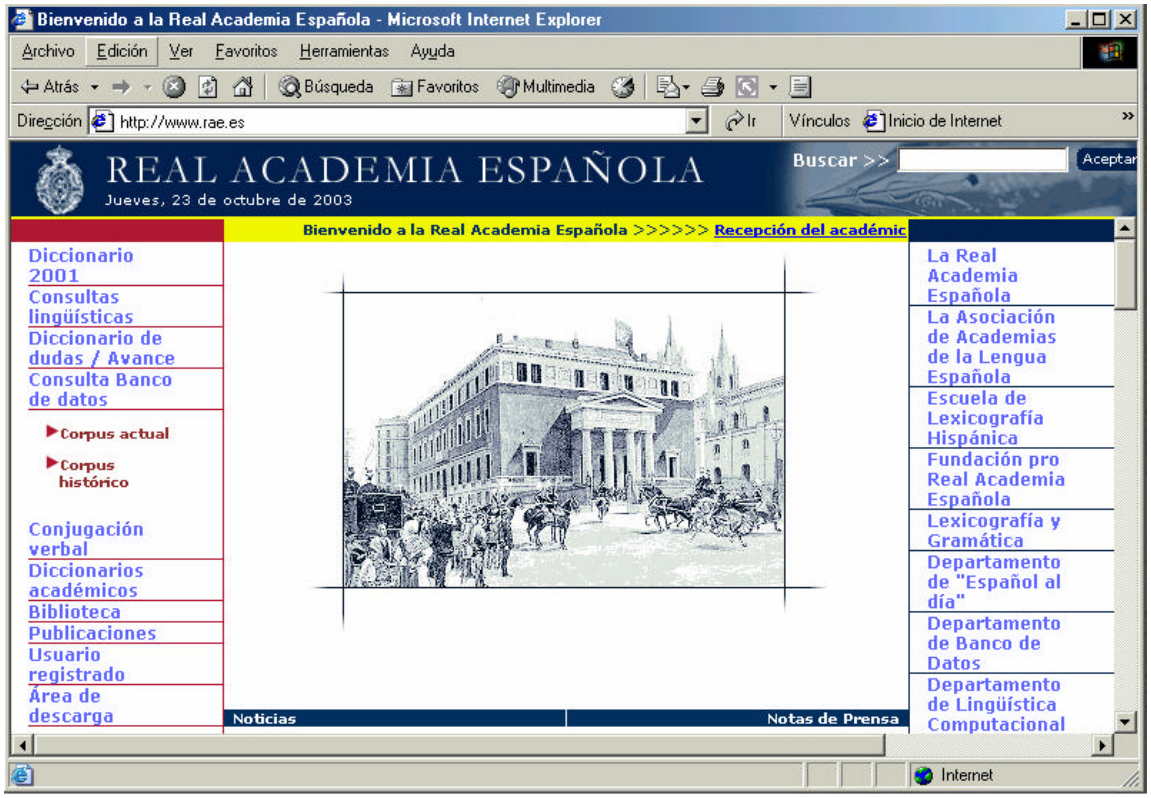
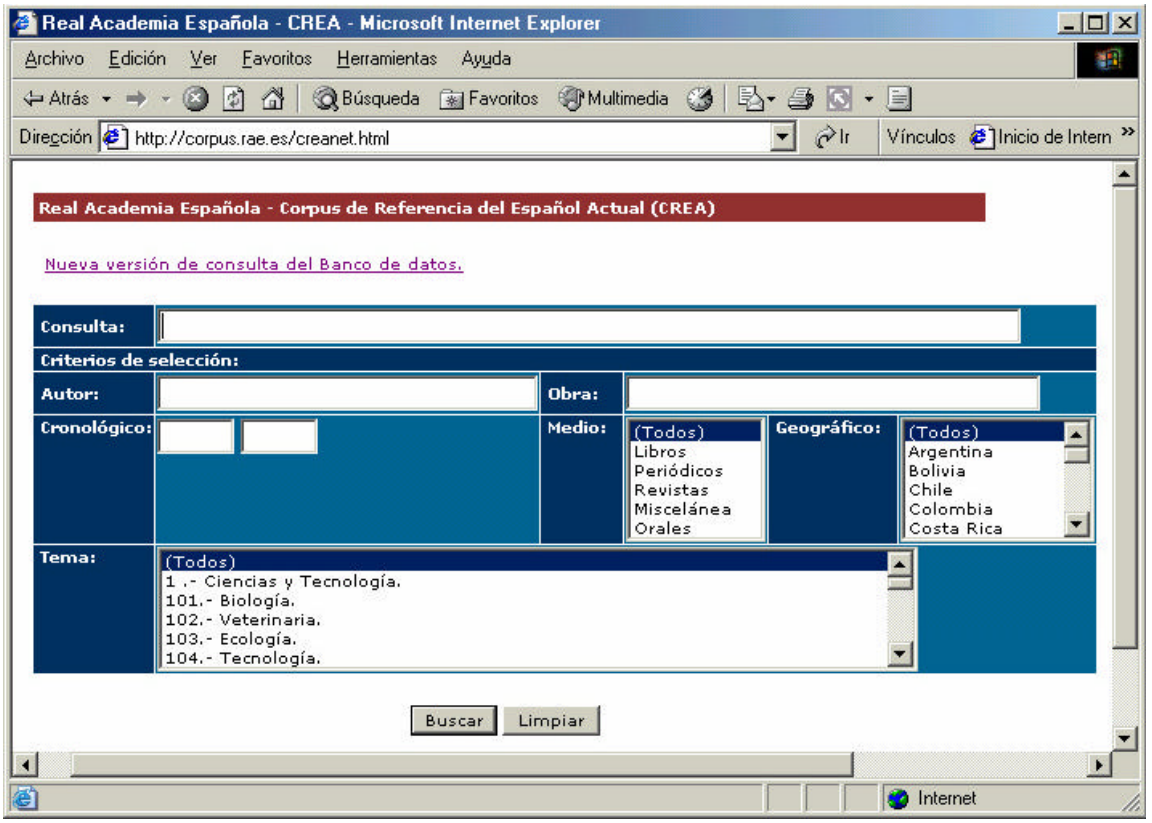


Figura 3.8. (*continuación*)



Para definir las fórmulas de consulta más eficientes respecto a la recuperación de concordancias, se tomaron en cuenta las anotaciones sobre la "Sintaxis del lenguaje de consulta" que ofrece el *Manual de consulta* del *CREA* (RAE, 2003, octubre). La búsqueda, además, se llevó a cabo estableciendo un criterio restrictivo de selección del ámbito geográfico que recuperase tan sólo aquellas concordancias que perteneciesen a documentos de origen mexicano. Este criterio de selección se puede indicar seleccionando la opción "México" con la barra de desplazamiento del criterio "Geográfico", mostrado en la parte central derecha del segundo elemento de la Figura 3.8. Una vez solicitada una búsqueda, el corpus muestra el total de casos o concordancias, en la base de datos del *CREA*, que coinciden con la fórmula. Además, indica el total de documentos en los cuales se

encuentran estos casos. Todos estos datos son desplegados en una ventana como la que se muestra en el primer elemento de la Figura 3.9. Las concordancias en sí se pueden desplegar haciendo clic sobre el botón “Recuperar” de la sección “OBTENCIÓN DE EJEMPLOS”, en la parte inferior de la mencionada ventana. En el ejemplo de búsqueda cuyos resultados se encuentran en el primer elemento de la Figura 3.9, vemos que la fórmula de búsqueda “ching*” (que permitiría recuperar, entre otras cosas, todas las formas conjugadas del lema *chingar*) arroja un total de 462 concordancias, en un total de 72 documentos distintos. El despliegue de estas concordancias se puede ver en el segundo elemento de la misma Figura 3.9. A su vez, si desea ver un contexto más amplio de una concordancia en concreto, se puede hacer clic sobre la fórmula resaltada de la concordancia en cuestión, la cual despliega una pantalla como la mostrada en el tercer elemento de la Figura 3.9. En el caso del ejemplo incluido en este último elemento de la figura, se puede ver que en éste se ha desplegado un contexto más extenso respecto de la concordancia número dos de las que aparecen listadas en el segundo elemento de la misma figura.

Figura 3.9. Resultados de consulta en el *CREA*

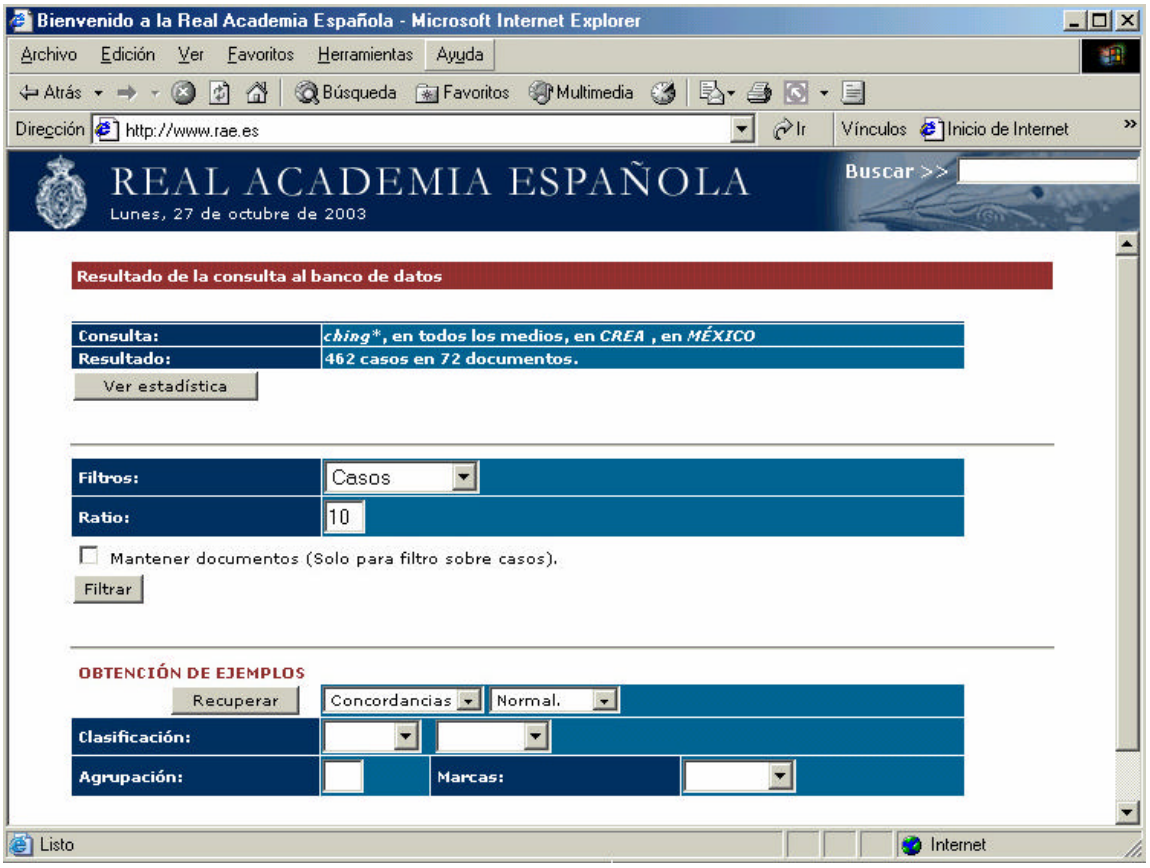


Figura 3.9. (*continuación*)

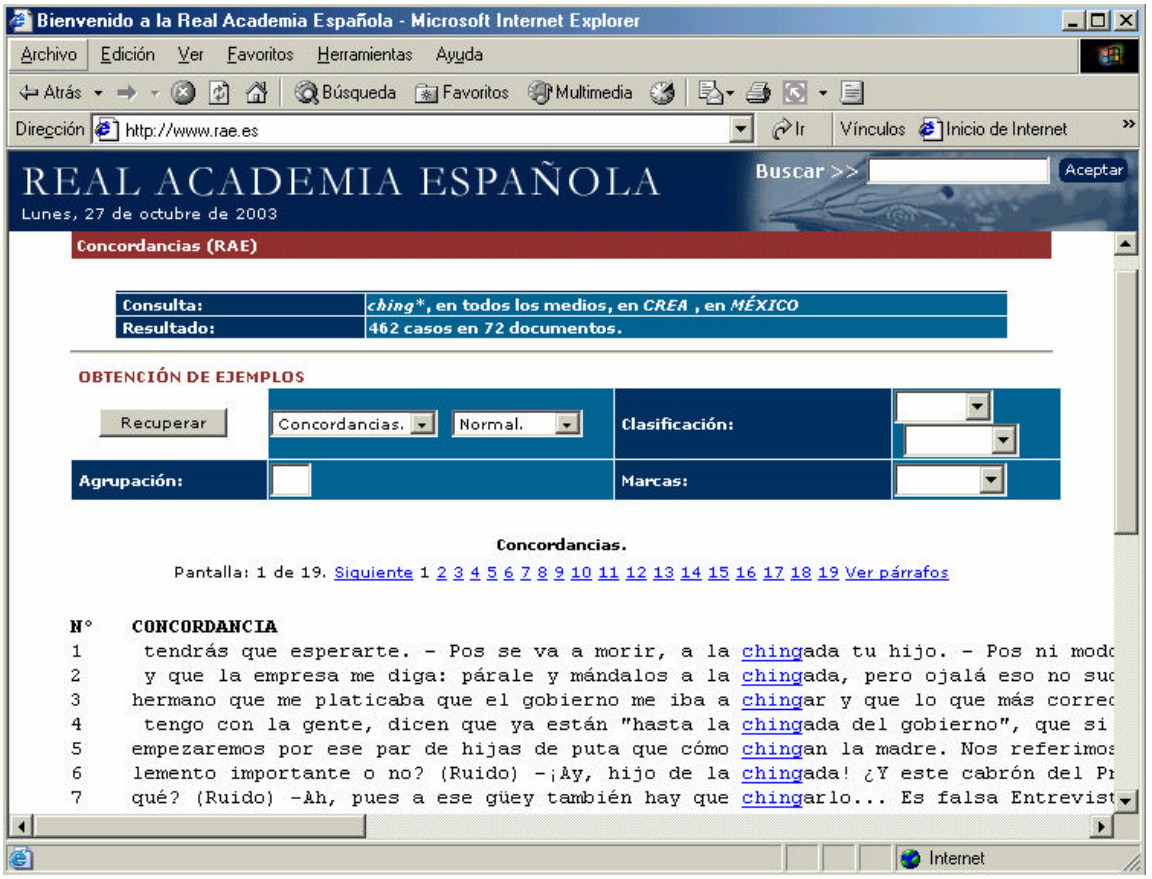
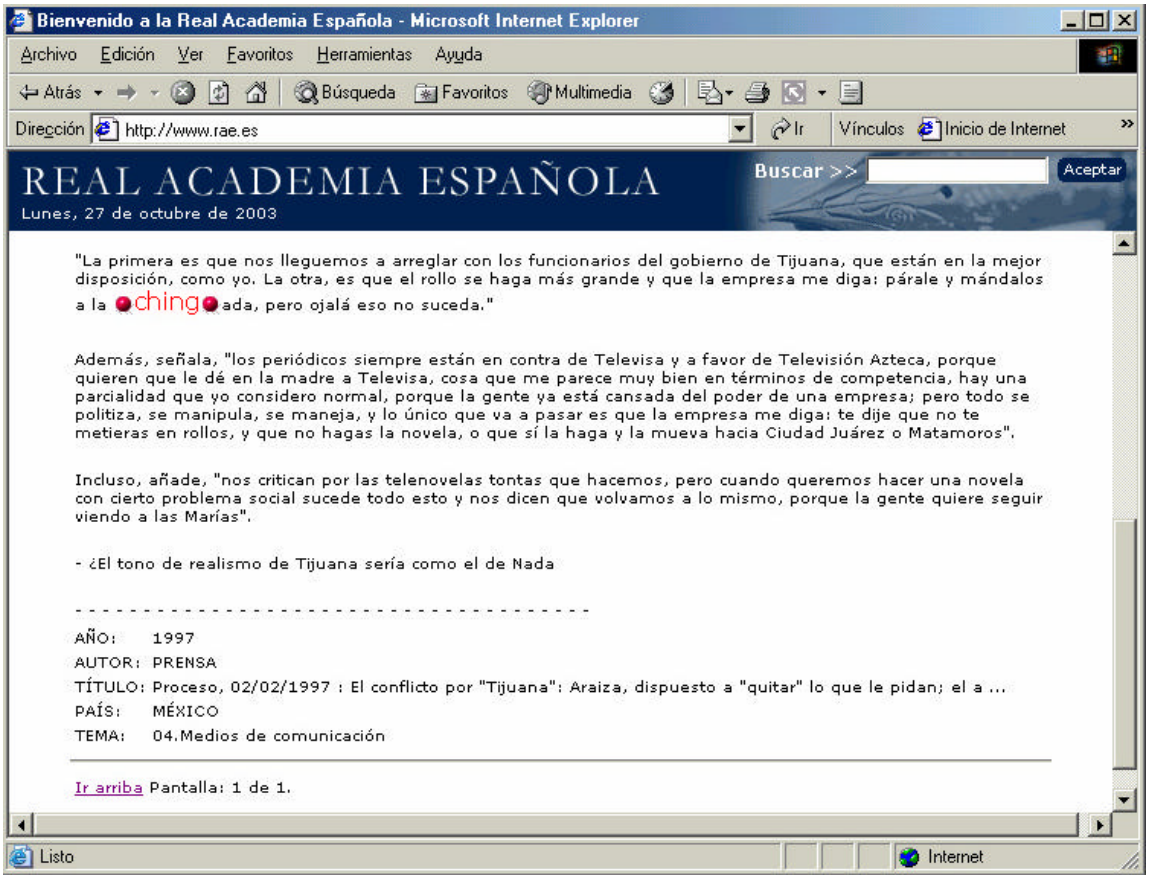


Figura 3.9. (*continuación*)



Después de obtener las concordancias de una fórmula en particular, se procedió a hacer el análisis dichas concordancias una por una hasta encontrar un máximo de dos ejemplos, por lema y por significado o acepción. Terminada la revisión de todas las concordancias, se llenaron los datos del lema trabajado, anotando cuántas concordancias o ejemplos contenía el corpus, en cuántos documentos distintos se hallaban dichas concordancias y cuántos de dichos ejemplos habían correspondido al lema y acepción buscada. En este último dato se reportó únicamente si se habían encontrado dos ejemplos, un ejemplo o ninguno. Si regresamos a ver la *lista DRAE documentada* en la Figura 3.8, podemos ver cómo en la columna M, la columna N y la columna O, se reporta que de los

lemas *andar*, *apendejar* (en su tercera acepción) y *babosada* (en su primera y segunda acepción) se encontraron dos ejemplos en el corpus; del lema *acordeón* se encontró un solo ejemplo; y de los lemas *albarda*, *alipús*, *anca*, *apendejar* (en su segunda acepción) y *bañar* no se encontró ningún ejemplo.

En cuanto al hallazgo de ejemplos pertinentes al lema y la acepción buscados, decidí que sería mejor crear fichas para una posible utilización posterior en caso de que se llevase a cabo la elaboración del diccionario. Estas fichas fueron creadas con base en los siguientes criterios. Se elaboraron fichas sólo para aquellos lemas de los que se hubiesen encontrado ejemplos de uso entre las concordancias del corpus. Una vez que se detectaba algún ejemplo de uso que pareciera reflejar el mismo significado o alguno relacionado con el presentado en el diccionario fuente, se procedía a desplegar todo su contexto en el corpus (ver el tercer elemento de la Figura 3.9). Esto se llevaba a cabo haciendo clic sobre la fórmula de búsqueda resaltada con un hipervínculo en la concordancia (ver el segundo elemento de la Figura 3.9). Si el contexto completo corroboraba la coincidencia entre el uso del lema y el significado plasmado por el diccionario fuente, se hacía un copiado de todo el contexto desplegado por el corpus, junto con los datos del documento de procedencia, a un documento nuevo de Word. Esto se intentaba llevar a cabo en un máximo de dos ocasiones por lema o acepción. Sin embargo, en algunos casos en que las concordancias pertinentes al lema trabajado resultaban abundantes, llegué a copiar más de dos ejemplos, sobre todo cuando me parecía encontrar ejemplos especialmente interesantes por los distintos matices de significación que reflejaban. Con todo, el máximo de ejemplos reportados por lema en el conteo final de la ejemplificación nunca fue más de dos. Hay que anotar también que, aun en los casos en que se halló un solo ejemplo de uso, también se generó una ficha electrónica. Ahora bien, una vez copiados los ejemplos recuperados por el corpus en el

documento de Word, se le ponía un encabezado a este último con los datos que permitieran identificar a qué lema le correspondían dichos ejemplos. Para llevar a cabo esto, se llenaban, en la lista documentada, las últimas cuatro columnas de la fila del lema trabajado. En la primera columna se anotaban dos datos, cuántas concordancias o ejemplos contenía el corpus respecto de la fórmula de búsqueda y en cuántos documentos distintos se hallaban dichas concordancias. En las tres columnas restantes se anotaban cuántos ejemplos, entre dichas concordancias, se habían encontrado para el lema en cuestión. Como se dijo anteriormente, respecto de este último dato sólo se indicó si se habían obtenido uno o dos ejemplos para aquellos lemas de los que se elaboró una ficha, y cero ejemplos en los casos en que no se elaboró ninguna.

Una vez llenada la fila correspondiente a un lema, se intentó incluir todos sus datos en el encabezado de los ejemplos contenidos en la ficha. Para hacer esto de manera ágil decidí crear un documento de Excel, llamado *Transposición* donde en una primera columna se colocase el nombre de los contenidos de cada una de las columnas de la lista documentada, pero en forma vertical en vez de horizontal. Hecho esto, lo único que se tenía que hacer era un copiado de la fila del lema trabajado y "trasponer" sus elementos de una fila horizontal a una columna vertical, en este caso en la columna subsiguiente a la primera del documento llamado *Transposición*. Esta operación se realizó utilizando la opción "Trasponer" de la ventana desplegada por la opción de "Pegado especial" del menú "Edición", en la barra del menú principal de Excel. Después de crear este documento con todos los datos del lema al que le correspondían los ejemplos de una ficha, se copiaba como encabezado de la misma. Los cambios hechos sobre el documento *Transposición* no se guardaban para poder seguir utilizándolo como una plantilla para nuevas fichas. Las fichas se iban salvando con un nombre conformado por las siglas del nombre del diccionario al

que pertenecieran, un guión bajo y el número progresivo, a cuatro cifras, del lema correspondiente en la lista general del diccionario. La nominación resultante dio ejemplos como los de las fichas *DEUM_0004*, *DRAE_0001*, *DBM_0003*, *AHDOM_0113*, entre otras tantas. Finalmente, resulta pertinente comentar un par de casos especiales que se presentaron durante la creación de las fichas y que requirieron nombrarlas de forma distintiva.

La primera situación se mencionó con anterioridad y tiene que ver con el hecho de que, como el muestreo se llevó a cabo de manera estratificada, era común que un lema se presentara en varios diccionarios con el mismo significado. El problema más notorio derivado de esta situación era una posible documentación múltiple que inflara los resultados de la representatividad del corpus. Además, este trabajo repetitivo implicaba una pérdida de tiempo y esfuerzo. Para evitar esta situación, como ya comentamos, se crearon fichas compartidas entre dos o más diccionarios. Las fichas resultantes fueron nombradas igual que las convencionales, pero se les añadió un segundo guión bajo seguido de la partícula "comp", tal como se puede apreciar en las fichas *DEUM_0117_comp*, *DRAE_0008_comp*, *DBM_0021_comp*, entre otras. Así pues, si tomamos como ejemplo a las fichas *DRAE_0039_comp*, *DBM_0021_comp*, y *DTV_0499_comp*, comentadas al presentar la *lista fusionada sample* en la Figura 3.6, podemos ver que estas tres fichas comparten los mismos ejemplos. Esto se debe a que estas fichas se refieren al lema común *chafirete*, el cual tiene un significado aproximado en las tres fichas, "mal chofer, mal conductor de vehículo automóvil" en el *DBM*, "chófer", en el *DRAE*, y "conductor de taxi, camión carguero o autobús" en el *DTV*. Los ejemplos de uso hallados en el *CREA* fueron un total de tres, de los cuales se copiaron todos, si bien tan sólo se reportaron dos. Estos ejemplos, en una versión condensada son:

(a) COZUMEL, Quintana Roo, 31 de agosto.- El veterano Efraín Payán disparó tres incogibles, incluso un cuadrangular, para encabezar el ataque con que los Rockies vencieron 4-3 a los Taxistas, para tomar ventaja en la serie final de la Liga Municipal Asterio Tejero.
Jonrón de Payán, ex jugador de los Leones de Yucatán en la Liga Mexicana, puso en ventaja a los Rockies en la primera entrada, pero jit de Alfonso Martín empató la pizarra.
El propio Martín, con sencillo en la sexta entrada, puso el pizarrón 3-1 a favor de los chafiretes, pero en la novena los Rockies atacaron con tres para llevarse el valioso triunfo.
AÑO: 1996
AUTOR: PRENSA
TÍTULO: Diario de Yucatán, 01/09/1996 : Truena el fusil de Payán en Cozumel
PAÍS: MÉXICO
TEMA: 05.Deportes

(b) Llegamos a La Perla Negra, y apenas entramos mi papá me murmuró al oído: "está igualito a como lo dejé". En la pista había un conjunto de música tropical que se parecía al Combo de Lobo y Melón; frente a él, dos rumberas que bien hubieran podido ser las Dolly Sisters, se tropezaban en lo redondo de una pequeña pista; un mulato de cabello afro cantaba acompañado del resto del combo: "yo soy el ruletero (que sí, señor, el ruletero). Yo soy el chafirete (que sí, señor, el chafirete). Yo soy el macalacachimba (que sí, señor, el macalacachimba). Yo soy el icuiricui (que sí, señor, el icuiricui)".
AÑO: 1985
AUTOR: Alatriste, Sealtiel
TÍTULO: Por vivir en quinto patio
PAÍS: MÉXICO
TEMA: 07.Novela

(c) Cosa rara: Cuca, la cuñada del Cachorro, no le respondió el saludo ni se detuvo, pasó de largo metiéndose a su casa, el nueve. Pero más tardó en entrar, que en salir disparada por un cachetadón del cuñadito. Maximina intervino en su defensa, para impedir que su marido le rompiera la crisma.
- Oye, si porque nos mantienes crees tener derecho de tranquear a mi hermana, te equivocas.
- Ningún mantienes, bien que me saca raja, ¿no me pone a que les ayude en todo?
Él, dirigiéndose a su esposa a modo de justificación:
- No vino a dormir, es una puta.
- Eso quisieras, güey. Me pegas porque soy mujer. Ganas me dan de traerte a mi novio para que te rompa...
- Anda, tráemelo. Me gustaría medirme con ese chafirete de la mudanza foránea, te apuesto a que se la pasó con él. Claro, con ese camión cerrado cualquiera no. Si trae su hotel sobre ruedas.
AÑO: 1993
AUTOR: Hayen, Jenny E.
TÍTULO: Por la calle de los anhelos
PAÍS: MÉXICO
TEMA: 07.Novela

Ahora bien, el conteo de estas fichas con ejemplificación compartida se tuvo que tomar en cuenta para hacer las inferencias correspondientes acerca de los posibles resultados finales del comportamiento de la población. Es decir, por un lado, si se generaron un cierto número de fichas compartidas, había que considerar que el número de lemas ejemplificados será inferior al total del número de fichas obtenidas. Esto se tomó en cuenta en el momento que se hizo el análisis de los datos. Por otro lado, las fichas comunes se tuvieron que incluir de alguna manera en el conteo final de la representatividad del corpus para no minar los resultados en este respecto.

El segundo problema que se presentó al crear las fichas tiene que ver con la diferenciación entre lemas y acepciones en los distintos diccionarios fuente. Anteriormente se llamó la atención sobre el hecho de que del *DIME*, que fue el último diccionario del que se extrajeron lemas, se creó una lista más bien de acepciones y no de lemas. Todo esto, como se explicó, fue hecho al momento de percatarse de las deficiencias en la lematización de varios diccionarios. Sin embargo, esta creación de listas de acepciones y no de lemas no se hizo desde un principio. Por ello, a la hora de construir las fichas resultó necesario dar cuenta del hecho de que algunas de ellas, que eran el resultado de trabajar un solo lema, ejemplificaban más de una acepción. En consideración a esto se decidió afectar también el nombre de las fichas que presentaran esta ejemplificación múltiple para su posterior contabilización, pues ignorar estas fichas implicaría subestimar la representatividad del corpus. La nominación resultante dio ejemplos como el de la ficha *DEUM_0510_triple*, el de la ficha *DBM_0018_doble*, entre otros. Un ejemplo de una ficha con más de una acepción, y con varios ejemplos para dichas acepciones, es precisamente el de la ficha *DBM_0018_doble*, donde el lema *carajo* presenta dos significados distintos (“interj. [interjección] que denota gran enfado o disgusto. | **del carajo**. 1. loc. adj. [locución

adjetiva] Malo, difícil, complicado) ejemplificados de manera múltiple. En el caso del primer significado se encontraron varios ejemplos, dos de los cuales son:

(a) Dice vehemente que "si en 1988 nos hubiéramos decid ido por una alianza estratégica, por una defensa del voto mancomunada; si no hubiéramos lanzado aquel llamado a la legitimidad, me pregunto yo si a partir de ahí no hubiera cambiado el destino del país. Nos hubiéramos evitado el paso de seis años por la historia de México de (Carlos) Salinas. ¡No hubiera estado en nuestra historia Salinas, carajo! Y ese solo hecho hubiera cambiado muchas de las cosas que hoy son tan dolorosas en este país".
Luego de recordar que a lo largo de la historia del PAN ha habido corrientes a favor de la participación en alianzas electorales con otras fuerzas, Fox dice no entender por qué Carlos Castillo Peraza "cambió repentinamente" su posición al respecto y hoy se opone a las alianzas.
AÑO: 1997
AUTOR: PRENSA
TÍTULO: Proceso, 19/01/1997: Fox insiste en la alianza por la Cámara "para echar a esos barbajanes de Los P...
PAÍS: MÉXICO
TEMA: 03.Política

(b) En lo que se refiere a su pretensión de que nunca soborna a los periodistas porque no asiste a banquetes y francachelas con ellos, parece un poco contradictorio cuando párrafos abajo leemos: "Está bien ir a comer, emborracharte con ellos en una feria, prestarles un carro para que se vayan si están muy borrachos y conseguirles boletos para el palenque; eso no es deshonestidad, no estoy comprando a nadie con eso. No es cierto que haya corrupción, carajo".
Agregaré a lo anterior que el mismo Herrerías me confesó en una ocasión, cuando decía que éramos amigos (?), que a los periodistas de un diario fotograbado les seguiría dando su sueldo usual.
AÑO: 1996
AUTOR: PRENSA
TÍTULO: Proceso, 15/12/1996 : De Enrique Guarner
PAÍS: MÉXICO
TEMA: 05.Tauromaquia

Dos ejemplos del segundo significado de este mismo lema son:

(a) CARGADOR 2 Pero que ya sigue mejor, me dijo el Guacho.
Página 63
CARGADOR-JEFE Quién sabe. Lo que sí es que yo no sé de qué la va a hacer el pobre cuate.
CARGADOR 2 De albañil, ésa era su chamba.
CARGADOR-JEFE Ya ni para eso. (Transición.) Bueno, pues ese mismo día, todavía luego del accidente, se armó un escándalo con el pinche dueño de la casa.

Empistolado el cabrón... A plomazos nos quería enfriar a todos el hijo de su rechingada. A ver nomás.
CARGADOR Bueno, ¿y por qué?
CARGADOR-JEFE Por sus pinches güevos... No, si te digo que fue un sábado del carajo.
Cargador-jefe se interrumpe. Jorge y Sara están entrando. Jorge va adelante, llevando libros que coloca junto a los que trajo anteriormente. Detrás, Sara llevando la base de una lámpara y algunos otros objetos.
AÑO: 1979
AUTOR: Leñero, Vicente
TÍTULO: La mudanza
PAÍS: MÉXICO
TEMA: 07.Teatro

(b) No Sergio, no es por ahí, el Mauro. Si quieren conseguir la base, eso cuesta, y no dinero, ciertamente, sino esfuerzo, sacrificio. Te mandé llamar porque, aunque no lo creas, me preocupa su situación. Podemos llegar a un arreglo, si quieres, tú y yo. Te conseguimos una buena liquidación y te vas por donde viniste sin problema alguno. Claro, yo entiendo que la situación está del carajo y que cada quien quisiera tener seguridad para su familia. Uno lucha por lo mejor para ella. El asunto es que no siempre es posible lograr, de momento, lo que uno pretende. Pongo por ejemplo mi caso. Ve esa fotografía. Quizá no te diga nada, pero para mí es conmovedor, francamente conmovedor ver desde dónde empezamos.
AÑO: 1990
AUTOR: Montaño Hurtado, Alfredo
TÍTULO: Las cenizas de los sueños
PAÍS: MÉXICO
TEMA: 07.Relatos

Así pues, debido a la ejemplificación múltiple que se presentó en algunos lemas, el resultado de la exploración del corpus, que se reportará en la siguiente sección de este trabajo (ver sección 4.01), se hará en acepciones y no en lemas, ya que para estimar el número total de estos últimos se necesitaría hacer una relematización de los materiales pertenecientes a las fuentes secundarias. Sin embargo, esta relematización, que sería parte necesaria del trabajo de elaboración del diccionario, rebasa los alcances del problema de investigación que se ha presentado aquí.