1 Introducción

1.01 Proyecto de investigación

En este trabajo se pretende explorar hasta qué punto resulta factible llevar a cabo la ejemplificación de un diccionario de léxico subestándar por medio de un corpus electrónico. Por ello, comenzaremos revisando cómo se definen dentro de la lexicografía los conceptos de *subestándar*, de *corpus electrónico* y de *diccionario ejemplificado* o *con citas.*

Entre los distintos tipos de obras lexicográficas se encuentran los llamados diccionarios de la lengua hablada o diccionarios del subestándar. El subestándar en lexicografía es entendido como un ámbito de la lengua en donde se mezclan una serie de niveles, tales como el coloquial, el familiar, el popular o el vulgar. Además, en las obras lexicográficas hispanas es común encontrar bajo el mismo rubro un número de vocablos de origen jergal, es decir sociolectal. Estos vocablos proceden de una variedad de sociolectos, como el de la delincuencia, el de los jóvenes estudiantes o el de grupos humanos marginales (Haensch, 1997, p. 94). En este sentido, hay que aclarar que un sociolecto es el conjunto de elementos que identifican la forma de hablar de un grupo de hablantes que comparten ciertas características socioculturales (Briz, 1996, pp. 15). De esta manera, en la mayoría de los trabajos lexicográficos del subestándar, es común que coexista léxico perteneciente a diferentes esferas. Por un lado, nos encontramos con un léxico que tiene una marcación diastrática, es decir, que caracteriza la manera de hablar de ciertos grupos sociales. Por otro lado, aparece un tipo de léxico que tiene una marcación diafásica, es decir, un léxico cuyo uso depende de la situación en que es utilizado. En este mismo

sentido, también es común encontrar incluido en los diccionarios del subestándar el léxico que tiene una marcación propia del tabú lingüístico, es decir, que forma parte de las groserías. Respecto a este último tipo de léxico, es pertinente mencionar que éste no corresponde a un nivel de lengua o diastrato, sino que representa una connotación especial en el uso social; es más bien una variación diafásica o pragmática, y por tanto, su utilización depende de las características del contexto en que se usa y de la adecuación entre el significado de las voces tabuizadas y la intención del hablante. De cualquier manera, siendo parte de la variación diafásica, es muy frecuente que se incluya el léxico tabuizado en el léxico del subestándar (Haensch, 1982b, p. 144). Así pues, la consideración de estos dos tipos de marcas, las diastráticas y las diafásicas, como un solo conglomerado, al menos dentro del campo de la lexicografía, parece normal e incluso justificada (Haensch, 1997, p. 94). Ahora bien, como hace notar Haensch (1982a), el problema que enfrenta la lexicografía en el tratamiento tanto del léxico subestándar como del léxico tabuizado está en el hecho de que muchos de los vocablos que conforman estos dos tipos de léxico, incluyendo aquellos vocablos que tienen una frecuencia de aparición muy alta, se hallan representados de manera sumamente irregular en los distintos textos escritos. En consecuencia, documentarlos de forma escrita es una tarea que en muchos casos ha dependido más bien de la casualidad (Haensch, 1982a, p. 442). Cabe recordar que la *causalidad*, según la Real Academia Española (RAE), es la "combinación de circunstancias que no se pueden prever ni evitar" (2001, p. 477). Es decir, lo que Haensch critica como "el gran problema de muchos diccionarios generales descriptivos" es la falta de regularidad que se ha tenido para la "documentación escrita" de "las unidades léxicas propias del lenguaje subestándar (familiar, popular, vulgar, tabuizado o jergal)", que por tanto ha resultado "en gran parte fruto del azar" (1982a, p. 442).

En cuanto a la documentación de las voces del diccionario, una de las herramientas tecnológicas utilizadas para la búsqueda y el hallazgo de posibles instancias de una palabra es el corpus electrónico. Para entender qué es un corpus electrónico, hay que comenzar definiendo el concepto de *corpus* en lexicografía. El corpus es el conjunto de textos, tanto orales como escritos, en que aparecen las palabras que se tomarán en cuenta para su inclusión en el diccionario. Estos textos conforman los materiales lexicográficos en que se basará la elaboración de la obra (Porto, 2002, p. 84). Para el almacenamiento de estos materiales, en la lexicografía tradicional es común el empleo de fichas textuales (Sinclair, 1985, p. 83), las cuales conforman, en conjunto, el llamado fichero de referencia. Ahora bien, en informática un corpus es simplemente una cadena de textos electrónicos enlazados de forma secuencial y etiquetados de manera que puedan ser identificados según una serie de datos. Los datos que se incluyen en el etiquetado, como el autor, el título de la obra, entre otros, son aquellos que se consideran necesarios para la búsqueda e identificación posteriores de los materiales que conforman el corpus (Porto, 2002, p. 131). Así, a través de recursos informáticos, un corpus electrónico tiene la capacidad de localizar y desplegar todos los contextos en que una palabra ha sido usada en una gran base de datos (Biber, Conrad y Reppen, 1998, p. 22). Esto hace que el corpus electrónico supere cuantitativamente al uso de fichas textuales. En este sentido, hay que tener presente que las fichas textuales representan tan sólo los contextos que un lector humano, o un grupo de éstos, puede registrar sobre una palabra dada. Tales contextos, además, corresponden en muchos casos a los usos menos comunes de un vocablo, y eliden los usos estadísticamente más significativos (Biber, Conrad y Reppen, 1998, p. 26).

Por todo lo anterior, el corpus electrónico constituye una herramienta única para la lexicografía, proveyéndole con un material no sólo cuantioso sino además vigente, pues la

mayoría de los corpus que se crean en la actualidad están en constante actualización. Esto último hace que el corpus pueda ser representativo de un estado de lengua determinado (Lara, 2002, p. 5). Una ventaja más del corpus electrónico es que, debido a que es de cuño reciente y se apoya en el uso de las computadoras, puede recabar y almacenar una gran cantidad de datos de las más variadas fuentes, incluyendo materiales orales (Biber, Conrad y Reppen, 1998, p. 22). Por todo esto, los elementos provenientes de los corpus representan usos lingüísticos naturales y espontáneos que pueden contrastarse con las intuiciones del autor, como hablante nativo, tanto a nivel de uso del término como del significado o significados del mismo (Biber, Conrad y Reppen, 1998, pp. 24-25). Con todo, no hay que olvidar que los corpus electrónicos también presentan algunas limitaciones desde el punto de vista de la labor lexicográfica.

Una de las principales carencias de los corpus electrónicos la constituye la posible falta de representación de vocablos. Esta falta de representación puede tener lugar aun cuando dichos vocablos y sus acepciones sean de los más comunes. Así pues, muchas veces los diccionarios que han sido elaborados dando una marcada preponderancia a la intuición del autor registran vocablos y acepciones que no se encuentran en los corpus electrónicos. Esto se da de tal forma que puede incluso haber casos en que una acepción en concreto sea de las primeras en aparecer en la mente de los hablantes, y a pesar de ello no sea frecuente en el uso diario (Biber, Conrad y Reppen, 1998, p. 41). Si bien es posible que estos casos resulten aislados, su existencia debe ser tomada en cuenta. Debido a lo anterior, es necesario reconocer que las intuiciones del lexicógrafo no deberían ser descartadas en su totalidad, sino que más bien deberían considerarse como un complemento a los resultados obtenidos por medio de los corpus electrónicos (Biber, Conrad y Reppen, 1998, p. 41). Siguiendo este razonamiento, Biber, Conrad y Reppen nos dicen que en el caso concreto de

los diccionarios diafásicos y, por extensión, de los diccionarios del léxico subestándar, la utilización de fichas textuales podría continuar siendo recomendable (1998, p. 26). Ahora bien, sea cual sea el medio con que se documenten los vocablos del diccionario, la búsqueda de la documentación es condición necesaria para llevar a cabo la elaboración de un diccionario ejemplificado. Para entender mejor en qué consiste este tipo de diccionario, habría que comenzar por explicar primero en qué consisten los llamados *diccionarios con citas.*

Un diccionario con citas es aquel en que los textos en que ocurren las palabras que forman parte del diccionario aparecen en el mismo como ejemplos (Porto, 2002, p. 84). Las citas son referencias textuales utilizadas para justificar, por un lado, la inclusión de un vocablo en un diccionario y, por otro, la definición dada de dicho vocablo (Bajo, 2000, p. 48). El término de *diccionario ejemplificado,* por su parte, es un concepto más general que engloba tanto al diccionario en que los ejemplos han sido creados por quien lo redacta, como a aquel en que los ejemplos han sido extraídos de algún corpus (Bajo, 2000, p. 49). En este trabajo en particular, cuando se habla de la elaboración de un diccionario ejemplificado, se está haciendo referencia a esta última posibilidad. Es decir, por diccionario ejemplificado aquí se alude a un diccionario con citas procedentes de un corpus, específicamente de un corpus electrónico.

En cuanto al mérito de los diccionarios ejemplificados, una de sus principales ventajas está en el hecho de que a partir de un ejemplo concreto del discurso de un individuo es posible hacer toda una serie de conclusiones sobre el modo de empleo de los elementos utilizados en esa realización particular del discurso (Werner, 1982, p.75). En otras palabras, el usuario del diccionario puede inferir una gama de informaciones sobre el uso de un vocablo a partir de un ejemplo concreto de dicho uso. Biber, Conrad y Reppen

afirman que el aprendiz de una lengua necesita algo más que una definición del significado y una lista de sinónimos más o menos aproximados; quien busca el significado de una palabra en un diccionario necesita saber además cómo se usa de forma real el vocablo en cuestión (1998, p. 53), y una forma de conseguir que lo descubra es proporcionándole un ejemplo. Otros posibles beneficios de la ejemplificación, mencionados por Bajo (2000), son el recordatorio de la norma lingüística, de algunas formas difíciles y del funcionamiento sintáctico; la indicación de combinaciones habituales; y la inclusión de información enciclopédica (p. 50). Sin embargo, esta autora menciona también la existencia de varios riesgos implícitos en la ejemplificación, tales como el cuidado en la selección de ejemplos para voces con marcas de restricciones de uso, la manipulación de definiciones *ad hoc* y la proyección de ideología en la selección de los ejemplos (Bajo, 2000, pp. 50-51).

Para este estudio, cuyo contexto es el subestándar, el principal inconveniente es, como se ya se mencionó, la dificultad para obtener los ejemplos a través de un corpus electrónico (Biber, Conrad y Reppen, 1998, p. 41). Esta situación se deriva del hecho de que la lingüística moderna quizá todavía no ha desarrollado una forma suficientemente sistemática, exhaustiva y verificable para el reconocimiento de este tipo de léxico (Lara, 1997, p. 249). Sin embargo, como Sinclair hace notar, la utilización de recursos electrónicos como los corpus no debería posponerse en la labor lexicográfica. Por el contrario, se deben aprovechar aquellos recursos electrónicos que se tengan disponibles e intentar obtener el mayor rendimiento posible de ellos (1985, p. 86).

1.02 Relevancia del proyecto

Así pues, a pesar de las limitantes, la realización de obras lexicográficas que registran un léxico subestándar o tabuizado, es decir, popular o grosero, es completamente

factible. Haensch hace énfasis en este hecho, y comenta la existencia de este tipo de obras en otras lenguas como el francés (1982b, p. 150). A pesar de ello, este mismo autor llama la atención acerca de los raquíticos avances de la lexicografía hispana respecto de este tipo de diccionarios diastrático-diafásicos (1982b, p. 150; 1997, p. 95). En cuanto al caso concreto de este tipo diccionarios en México, hay que señalar que es posible encontrarse con un par de trabajos lexicográficos que resultan al menos ricos en materiales, el de Colín (1987/2001) y el de Jiménez (1999). Sin embargo, si a la observación de Haensch acerca de la pobreza en los avances de la lexicografía hispana del subestándar le añadimos la posibilidad de contar con un diccionario ejemplificado de este tipo de léxico para el español en México, el progreso en ese sentido resulta simplemente nulo. Por otro lado, la relevancia de la elaboración de un diccionario ejemplificado del subestándar no se desprende únicamente de la carencia de un trabajo similar previo, sino de la importancia, recién mencionada, que tiene la ejemplificación en sí para cualquier diccionario (Bajo, 2000, pp. 50-51; Biber, Conrad y Reppen, 1998, p. 53; Werner, 1982, p. 75).

1.03 Planteamiento del problema

En la busca de subsanar la carencia de un diccionario ejemplificado del español en México, este trabajo pretende utilizar uno de los pocos corpus electrónicos que existen para el español actual, de hecho, el único disponible al público en Internet, el *Corpus de referencia del español actual*, *CREA* (Porto, 2002, p. 129). Por medio de la utilización de este corpus, se intenta explorar hasta qué punto resulta viable llevar a cabo la ejemplificación de una lista exhaustiva de palabras o lemas pertenecientes al léxico subestándar del español en México. Esta lista de lemas se ha obtenido a partir de todas las fuentes secundarias, u obras lexicográficas predecesoras, que han surgido en los últimos

veinticinco años de labor lexicográfica respecto del español en México. La pregunta acerca de la posibilidad de llevar a cabo la ejemplificación de dichas fuentes se intenta contestar por medio de la documentación de un muestreo aleatorio de las mismas. Con la documentación del muestreo se pretende estimar el volumen potencial de la ejemplificación total del conjunto de lemas del subestándar provenientes de las obras lexicográficas predecesoras de este trabajo (ver sección 2.07).

Aquí cabe aclarar que por lema se debe entender la reducción de "todas las formas de una palabra a la forma paradigmática considerada fundamental (el infinitivo para todas la formas verbales del mismo paradigma, el singular para los sustantivos…)" (Bajo, 2000, p. 16). El lema es por tanto el encabezamiento de la *entrada* o unidad autónoma mínima del diccionario en la cual aparece cada una de las unidades léxicas incluidas en el mismo. Para mostrar mejor este concepto he decidido poner a continuación un ejemplo de entrada tomado del diccionario de Moliner (1998, p. 154):

> **amachinarse** (de «a-[2]» y «machín», nombre aplicado al dios cupido) **1** (Am. C., Col., Méj.) prnl. recípr. **Amancebarse.* **2** (Guat., Pan.) prnl. **Abatirse, acobardarse.*

En esta entrada, nos encontramos con que el lema corresponde a la palabra *amachinarse* que aparece en letras redondas, minúsculas y resaltadas en negritas. La entrada está constituida por el párrafo, separado del resto del texto, que conforma el ejemplo en su totalidad. Si bien esta entrada, con tres líneas en su fuente original, se antoja corta, algunas otras entradas pueden ser mucho más largas. En este mismo diccionario, por ejemplo, nos encontramos la entrada *alzar*, aparecida poco antes de la entrada *amachinarse*, que alcanza casi las 60 líneas (Moliner, 1998, pp. 153-154). Igualmente, nos podemos

encontrar con la entrada *Am* con apenas cinco palabras en total. Ahora bien, aludiendo al concepto de entrada, nos encontramos con que en ésta, además del lema, aparece una definición del mismo, así como cualquier otra información relativa a la unidad léxica (Haensch, 1997, pp. 39-40). También habría que apuntar que en la definición del lema puede aparecer uno o más significados o acepciones. En el ejemplo de Moliner (1998) aquí presentado, constituido por la entrada de *amachinarse*, podemos ver que la definición contiene dos acepciones que están marcadas por los números uno y dos en negritas. Las acepciones están destacadas en este diccionario con letras cursivas, y el comienzo de cada una de ellas está indicado por un asterisco. En este mismo ejemplo nos encontramos con algunas de las informaciones relativas a la unidad léxica que Haensch (1997) menciona. Después del lema *amachinarse* aparece un paréntesis con información etimológica, relativa a los orígenes de la palabra[1]. De manera similar, entre el número que indica el conteo progresivo de cada una de las acepciones del lema y la definición de cada acepción en sí, aparece otro paréntesis, de ámbito geográfico, que indica los lugares en donde tiene efecto la acepción subsiguiente[2]. Además, las informaciones que se encuentran después de estos paréntesis de ámbito geográfico y antes de la definición misma son informaciones de carácter gramatical[3]. Así pues, una vez explicados los conceptos de lema y entrada, se puede entender más fácilmente el concepto de *lemario*, el cual es el conjunto de todos los lemas de un diccionario, y cuya contabilización equivale al total de entradas incluidas en el

[1] Respecto de este paréntesis, en él aparecen dos envíos o remisiones a otras entradas del diccionario. En el caso del primer envío, además, se incluye un superíndice, el número dos. Este superíndice indica qué número de acepción de la entrada a la que se remite es el que corresponde al significado con que se utiliza el término enviado en el paréntesis.

[2] En la primera acepción aparece el grupo de abreviaturas Am. C., que significa América Central, y las abreviaturas simples, Col. (Colombia) y Méj. (México). En la segunda acepción aparecen las abreviaturas Guat. (Guatemala) y Pan. (Panamá).

[3] La abreviación prnl. que aparece en las dos acepciones presentadas en el ejemplo corresponde a la indicación *pronominal* que se puede hacer respecto de un verbo. La indicación recípr., aparecida únicamente en la primera acepción, significa *recíproco*.

diccionario. El lemario, junto con sus entradas correspond ientes, constituye por tanto la parte más substanciosa del cuerpo del diccionario.

La lista utilizada en este trabajo constituye tan sólo un lemario tentativo para un diccionario del español subestándar en México. Este lemario provisorio ha sido obtenido haciendo una revisión de todos los diccionarios (fuentes secundarias) aparecidos en los últimos veinticincos años, que por cuya construcción contuvieran lemas pertenecientes al léxico del subestándar del español en México. A este respecto debemos recordar que el subestándar aquí debe ser entendido como la combinación de la variación diastrática y la variación diafásica que se mencionó con anterioridad.