ESTADÍSTICA Y LINGÜÍSTICA DE CORPUS: IMPLICACIONES PEDAGÓGICAS EN LA ENSEÑANZA Y EL APRENDIZAJE DEL LÉXICO

Mª del Carmen Ávila Martín
Universidad de Granada (España)

Resumen: Los estudios estadísticos del léxico se remontan al siglo pasado pero, en el momento actual, el desarrollo de la lingüística de corpus nos proporciona nuevas herramientas para su utilización en los métodos de aprendizaje de la lengua. En la situación actual, el desarrollo de métodos informáticos está permitiendo que los estudios cuantitativos del léxico renueven su importancia para la didáctica de la lengua tanto materna como extranjera. Sin embargo, el estudio cuantitativo del léxico debe ser puesto en relación con las necesidades de aprendizaje para conseguir un mejor rendimiento cualitativo.
Palabras clave: Lexicología, lingüística de corpus, estadística léxica, léxico.

Abstract: Statistical studies of lexicography date back to the last century, but today the development of corpus linguistics provides us with new tools to use in language learning methods. In its current situation, the development of computer technology methods allows quantitative lexical studies to regain its importance for language teaching whether it is a native or foreign language being taught. Nevertheless, a quantitative lexical study needs to be assessed in relation to learning needs in order to achieve a greater qualitative result.
Key words: Lexicology, Corpus Linguistics, Lexical Statistics, Lexicon.

Résumé: Les études statistiques du lexique ont commence le passé siècle mais, en ce moment, le développement de la linguistique de corpus nous fournit de nouveaux outils pour l'utilisation dans les méthodes d'apprentissage de la langue. Dans l'actuelle situation, les méthodes informatiques perment que les études quantitatives du lexique renouvèlent son importance pour la didactique de la langue tant maternelle comme étrangère. Cependant, l'étude quantitative du lexique doit être mise en relation avec les nécessités d'apprentissage pour obtenir des meilleurs resultats.
Mots-clés: Lexicologie, Linguistique de corpus, Statistique lexicale, lexique.

1. INTRODUCCIÓN

Los estudios estadísticos del léxico se han aplicado a la enseñanza y aprendizaje de las lenguas desde comienzos del siglo XX. En el momento actual, la aparición de nuevos métodos de recuento de las unidades léxicas ofrece una mayor precisión en la recopilación de palabras frecuentes. La importancia de la frecuencia en la selección del léxico en los métodos de enseñanza y aprendizaje de las lenguas se ha puesto de manifiesto en muy diversas ocasiones, si bien no se contaba con herramientas tan precisas en el recuento de unidades como las que tenemos actualmente a disposición de cualquier profesional.

La aplicación de estos métodos estadísticos al aprendizaje de lenguas ha sido

completada en la actualidad con métodos que incorporan el contexto y la situación comunicativa o aspectos cognitivos como factores determinantes en el aprendizaje. Aunque la importancia de la frecuencia ha sido cuestionada y no ha sido tenida en cuenta en muchos métodos de aprendizaje, debemos seguir preguntándonos por su validez. La frecuencia de uso de una unidad léxica no es el único dato relevante y se debe poner en relación con otros tales como el de *disponibilidad* o con conceptos tales como el de *riqueza léxica*. Nuestro objetivo en este trabajo es buscar una respuesta a la pregunta de cuánto vocabulario hay que aprender.

2. LOS ESTUDIOS ESTADÍSTICOS DEL LÉXICO

Los recuentos léxicos y los vocabularios estadísticos han tenido varias aplicaciones disciplinares y diversos objetivos científicos en las sucesivas etapas históricas en que se han realizado.

En la década de los años veinte se iniciaron estas recopilaciones para el español destinadas especialmente a la enseñanza del español a extranjeros. El primero que se conoce es el de H. Keniston, «Common words in Spanish» (1920)[1]. Su intención era establecer un inventario de lengua cotidiana e incluyó obras de teatro, periódicos, revistas así como cuentos y novelas. La lista incluye 1322 palabras divididas en ocho secciones, ordenadas por frecuencias. Según señala Keniston, las 185 palabras que aparecen en la primera lista se encuentran al menos en el 80% de los textos estudiados, mientras que las 221 de la lista octava, aparecen en el 33% de esos textos.

Con finalidad pedagógica se crearon también para el inglés listas de frecuencias para uso de los enseñantes (Alvar 2005: 20). E. L. Thorndike (1921) creó una lista de diez mil palabras:

> The Teacher's Word Book is an alphabetical list of the 10,000 words which are found to occur most widely in a count of about 625,000 words from literatura for children; about 3,000,000words from the Bible and English classics; about 300,000 words from elementary-scooll text books; about 50,000 words from books about cooking, sewing, farming, the trades, and the like; about 90,000words from de daily newspapers; and about 500,000 words from correspondence. Forty-one different sources were used (1921: 7).

En esa misma década se publica también *Graded Spanish Word Book*, de Milton A. Buchanan (1927), basado en un recuento de lenguaje escrito, que gozó de cierto prestigio en los años siguientes[2].

---

[1] Esta obra tuvo una segunda versión en Keniston, H. (1929) *Spanish Idiom List*. New Cork, Macmillan.

[2] También se publicó en esta época un listado de carácter plurilingüe: EATON, H. (1940) *An English-French-German-Spanish Word Frequency Dictionary*. New Cork: Dover Publications.

En la década de los cincuenta se publican varios trabajos, tanto en el ámbito del español como en el ámbito del francés, siempre con un interés pedagógico. En 1952 se publicó el *Recuento de vocabulario español* de Rodríguez Bou. La lista realizada incluía palabras del lenguaje oral y escrito de niños y adultos –composiciones y conversaciones transcritas de niños, periódicos, revistas, programas de radio, literatura religiosa y textos escolares–.

A mediados de los años cincuenta el criterio de frecuencia se utilizó en Francia para la creación de manuales de francés elemental cuya finalidad era facilitar el aprendizaje de esta lengua a los inmigrantes franceses y a los habitantes de las excolonias francesas. El equipo estaba dirigido por Georges Gougenheim, y de él formaban parte René Michéa, Paul Rivenc y Aurélien Sauvageot, y publicó en 1954 la obra *Français Fondamental* «estudio estadístico sobre la frecuencia de las palabras concebido como instrumento eficaz para la difusión amplia y rápida de la lengua francesa», que además se basaba en el lenguaje oral y que se considera pionera de este tipo de estudios. Para la selección del léxico se basaron únicamente en el criterio de la frecuencia y cuando analizaron las listas se dieron cuenta de que en ellas faltaban palabras muy normales como *timbre, lettre* y otras como *auto* o *métro* tenían frecuencias muy bajas; por lo que, siguiendo ese criterio no entrarían a formar parte del francés elemental. Quedaba así patente que la frecuencia sola no podía ser el único criterio para la selección del léxico. Los investigadores franceses se dieron cuenta de que había dos tipos de palabras: unas que son frecuentes independientemente del tema –Michéa las llamó *atemáticas*– y otras cuyo uso y, por lo tanto, su frecuencia, depende del tema de que se trate –*temáticas*, según Michéa–. Estas últimas son el objeto de estudio de la disponibilidad léxica[3].

La recopilación de listas de palabras atendiendo a la frecuencia y destinadas al ámbito pedagógico se inició también en España en los años 50. Uno de los primeros trabajos fue el *Vocabulario usual, común y fundamental* de Víctor García Hoz, publicado en 1953. Esta obra recoge un vocabulario aproximado de 15 000 palabras que era la cantidad que se suponía debía dominar un adulto. Se selecciona un corpus de unas 400 000 unidades de cuatro aspectos: a) la vida familiar –cartas privadas–, b) la vida social indiferenciada –periódicos–, c) la vida social regulada –documentos oficiales del estado y eclesiásticos, así como folletos de organizaciones sindicales–, y d) la vida cultural –textos de los libros más vendidos–.

A partir de dicho corpus, García Hoz estableció tres estratos o niveles léxicos: vocabulario usual –12 428 voces–, común –1971 voces, presentes en los cuatro bloques seleccionados– y fundamental –208 voces–. En cuanto a la aplicación de este vocabulario común en la lectura ya señalaba este autor en los años cincuenta:

[3] Sobre los numerosos trabajos de disponibilidad léxica que se están realizando para el español puede consultarse la página http://www.dispolex.com/.

> Está en manos de los autores de libros escolares, quienes, en vez de usar palabras cultas, que en muchas ocasiones reflejan una retórica rebuscada y empalagosa, debieran emplear con preferencia las palabras de uso corriente.

El estudio estadístico del léxico no solo se ha realizado con fines didácticos. Recordemos por ejemplo los análisis estadísticos de Charles Müller (1964) que tenían como finalidad el estudio estilístico de autores literarios y que se ha empleado con profusión en el recuento de unidades léxicas empleadas por un autor o en una obra en concreto.

El diccionario de frecuencias más citado de nuestro ámbito es el *Frequency Dictionnary of Spanish Word* de Juilland y Chang-Rodríguez, que se publicó en 1964. Utilizó textos pertenecientes al drama, narrativa, ensayo, literatura técnica y científica y periodismo de los años comprendidos entre 1920 y 1940. Está basado en un corpus de 500 000 unidades y obtiene un vocabulario básico de 5 000 palabras. Estos autores utilizaron 100 000 ocurrencias en cada uno de los géneros citados, por lo que además de indicar la frecuencia absoluta de una palabra, se daban indicaciones de la frecuencia relativa en cada género, y se podían establecer comparaciones de género a género de la importancia de una palabra (Lara 2006:69).

Esta obra influyó en la realización de otras en el ámbito hispánico, así El *Léxico básico del español de Puerto Rico*, realizado por Amparo Morales en 1986, con un interés didáctico. Utilizó los mismos criterios que la obra anterior, pero circunscrito al español de Puerto Rico y a los años 1948 -1970.

Algunos trabajos sobre vocabularios de frecuencias son estudios basados en el lenguaje producido por los adolescentes en su lengua materna. Así, la obra de Fernando Justicia, *El desarrollo del vocabulario. Diccionario de frecuencias* (1995) se basa en un corpus recogido entre 1985 y 1988 basado en producciones escritas de 3 402 niños –de 6 a 13 años– de las provincias de Almería, Granada, Jaén y Málaga. El universo léxico total se compone de más de medio millón de ocurrencias –528 544– para obtener un vocabulario básico de 8 937 palabras –vocabulario común a los tres niveles más los vocabularios diferenciales a cada nivel–. Según diversos autores, la amplitud media del vocabulario del niño en esas edades se sitúa en torno a las 8 500 palabras, cifra tomada como referencia de la cantidad que se debía estimar. Siguiendo las pautas establecidas por Fernando Justicia, Jorge Molero Huertas publicó *El vocabulario usual de nuestros alumnos* (1999). La recopilación se basó en 412 433 palabras para obtener 9 887 vocablos procedentes de 1864 composiciones escritas de 932 alumnos de 14 y 15 años de Granada.

Los últimos diccionarios de frecuencias del siglo XX se publican en la universidad de Oviedo (Alameda *et al.* 1995) y en la Universidad de Barcelona (Sebastián *et al.* 2000).

Los repertorios léxicos que se realizaban a principio de siglo eran recuentos manuales y en la primera época del manejo de los ordenadores el número de textos que se

introducían era escaso porque había que introducir los textos manualmente. A partir de los años noventa, el desarrollo de la informática empezó a posibilitar el escaneo automático de textos que permite que el uso de material sea mucho más extenso. Los recuentos anteriores se han basado mayoritariamente en fuentes escritas, fundamentalmente debido a la preponderancia de lo escrito sobre lo oral en la tradición estadística. Sin embargo, los programas de reconocimiento de voz están posibilitando la introducción de textos hablados y su incorporación a los corpus, así como la misma realización de corpus orales.

3. LA LINGÜÍSTICA DE CORPUS

El desarrollo de los métodos informáticos ha propiciado en nuestros días la aparición de una nueva disciplina lingüística denominada lingüística de corpus[4] (Manuel Ávila, 1999: p. 12 y ss). Esta disciplina se basa en una técnica de trabajo que propicia el acopio masivo de datos, por medio de herramientas informáticas. Esto permite la realización de investigaciones empíricas del lenguaje más que el estudio introspectivo del lenguaje. Es el mismo principio que guiaba los estudios estadísticos del lenguaje, pero las aplicaciones de la lingüística del corpus son mucho más ambiciosas debido a los progresos que la informática ha experimentado desde los años 60 hasta nuestros días. Así los campos de trabajo de la lingüística de corpus tienen que ver con la traducción automática, la comprensión del lenguaje humano por parte de un ordenador, incluyendo generación y síntesis de habla; la investigación del lenguaje mismo y la comprobación y mejora de las descripciones lingüísticas actuales mediante la aplicación a los corpus.

Se han elaborado y se están elaborando un buen número de corpus del español en la actualidad con diferentes finalidades descriptivas. El manejo de gran cantidad de datos léxicos tiene como consecuencia la elaboración de listados de frecuencias. No podemos enumerar la gran cantidad de trabajos que se están realizando en esta área, pero nos detendremos en las consecuencias que se extraen de algunos de estos estudios y que revelan datos de interés pedagógico. Algunos de ellos destacan por tratarse de descripciones de lengua oral, mientras que otros se utilizan en la elaboración de obras lexicográficas.

Entre los trabajos que describen el lenguaje oral en los recuentos léxicos del español podemos citar la obra de Antonio Manuel Ávila Muñoz (1999) titulada *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Esta obra está basada en textos hablados –54 horas y media de grabación– correspondientes a cinco tipos discursivos. Participan 291 sujetos de la ciudad de Málaga y se tienen en cuenta las variantes sociológicas de sexo,

[4] El Primer Congreso Internacional de Lingüística de Corpus se celebró en Murcia en mayo de 2009.

edad y educación. El corpus analizado consta también de 500 000 palabras para llegar a un vocabulario básico de 5 228 lemas.

También tiene en cuenta el lenguaje oral el recuento realizado por Marcial Terrádez Gurrea titulado *Frecuencias léxicas del español coloquial* (2001). Está basado en un total de 100 000 palabras procedentes de 25 conversaciones coloquiales grabadas en Valencia y su área metropolitana desde 1991 a 1999 por el grupo Val.Es.Co. Las conversaciones tratan de temas variados –familia, trabajo, tabaco, deporte, amigos, salidas, estudios, política, relaciones amorosas, enfermedad, comida, carné de conducir...–. Se tuvieron en cuenta las variables de sexo, edad y nivel sociocultural.

Entre los corpus que se han realizado para la elaboración de obras lexicográficas, encontramos el *Corpus Cumbre* (2001) cuyo director fue Aquilino Sánchez Pérez. Este corpus está basado en veinte millones de palabras del español escrito y oral de España e Hispanoamérica. La distribución de textos en el corpus es la siguiente: España 65% e Hispanoamérica 35%. En la parte de España: 70% escrito y 30% oral –la mitad pertenece a muestras grabadas de la radio y televisión y la otra mitad a grabaciones de la vida diaria–. En la parte de Hispanoamérica: 60% escrito y 40% oral. Las muestras orales se recogieron entre 1993 y 1994 y las escritas datan de la segunda mitad del siglo XX.

El corpus Cumbre se ha utilizado en la elaboración del *Gran diccionario de uso del español* (2001) en el que se indica por primera vez en un diccionario del español la frecuencia de las palabras mediante la marcación numérica de la frecuencia –del 1 al 5–.En muchas ocasiones los corpus no son accesibles al gran público, por pertenecer a editoriales u otros organismos, o están publicados en papel. Sin embargo, en los últimos años se encuentran al alcance del público varios corpus. En primer lugar, señalaremos el corpus disponible en la red de Mark Davis, (http://www.corpusdelespanol.org), realizado en 2002, con una nueva versión en 2007.

Este corpus se basa en unos 100 millones de palabras procedentes de más de 20000 textos del español, desde 1200 a 1900. Para el siglo XX ha realizado cuatro divisiones de igual tamaño –5 millones de palabras– para los distintos registros: oral, ficción, periodismo y texto académico.

Basándose en este corpus, Davis ha publicado *A frecuency dictionary of Spanish* (2006) que recoge las 5000 palabras más frecuentes del español. Se basa en el español del siglo XX –del que recoge 20 millones de palabras en su corpus–. El índice principal del diccionario recoge los lemas, ordenados por frecuencia, con la indicación de la categoría gramatical, la traducción en inglés, un ejemplo extraído del corpus, el rango total de la frecuencia –*range count*–, es decir su presencia en los distintos bloques de registros del corpus y el número total de veces que aparece –*raw count*–. Finalmente, se incorpora una indicación sobre en qué tipo de registro aparece con más frecuencia. (*cf.* http://davies-linguistics.byu.edu/personal/spanlex.asp).

El más extenso de los corpus del español hasta el momento está también disponible en la red. El CREA –Corpus de Referencia del Español Actual– de la Real Academia de la Lengua Española. Este corpus está accesible en la página de la RAE (http://www.rae.es) y en su última versión cuenta con 154.279.050 formas, pertenecientes a textos procedentes de todos los países hispánicos y producidos entre 1975 y 2004. En esta versión de 2008, se incorporaron también las listas de las 1.000, 5.000 y 10.000 formas más frecuentes del español, así como la lista de todas las formas presentes en el CREA, con indicación de su frecuencia absoluta y normalizada. El hecho de que esté disponible para todo el público que desee consultarlo, el estar realizado por una institución de tanto prestigio, y la gran extensión de este corpus, lo convierte en el corpus de referencia del español actual.

4. IMPLICACIONES PEDAGÓGICAS

La frecuencia de aparición de una palabra en un corpus tiene un interés pedagógico que parece innegable. Si una palabra se utiliza con frecuencia tendrá mayor utilidad comunicativa que otra que se utilice menos. Pero los datos que reflejan los corpus, aunque no son coincidentes en gran medida, sí arrojan datos sorprendentes que nos interesa señalar por sus implicaciones pedagógicas.

La importancia comunicativa de una palabra no se rige solo por el dato de frecuencia, sino que hay que tener también en cuenta la dispersión y el uso. Dos unidades pueden tener la misma frecuencia pero tener una distribución diferente en los distintos tipos discursivos. Es decir, una palabra que aparezca muchas veces en un mismo tipo de texto es menos importante que una palabra que tenga la misma frecuencia pero que aparezca constantemente en diversos tipos de texto. Así se mide, además de la frecuencia, el rango, es decir, se mide la dispersión de un término en periodos temporales diversos y en la serie de fuentes, mundos o tipos de texto que constituyen el corpus. La dispersión nos señala el valor general de un término al determinar su presencia en conjuntos o tipos de textos de la muestra total. La dispersión mide, por consiguiente, la estabilidad de la frecuencia, aportando así información de interés sobre la utilidad de una unidad léxica en la lengua (L. F. Lara 2006:170).

Sabemos, por los datos ofrecidos en los corpus, que el aumento de palabras en un corpus no implica necesariamente un aumento del número de unidades léxicas diferentes. En cualquier corpus de textos, más o menos la mitad de las palabras sólo aparece una vez, mientras que la mayor parte de las restantes aparece menos de diez veces, según señala Manuel Ávila (1999: 14). Por su parte, Luis F. Lara explica que en el análisis de un corpus:

> el número de *ocurrencias* –cada aparición de una palabra– va creciendo una a una hasta terminar su recuento; en cambio el número de *tipos* –cada palabra eliminando las repeticiones– va creciendo menos, porque muchos de ellos se repiten, y el número de *vocablos* –representante de un paradigma– resultante todavía menos. Después de analizar por ejemplo 100 000 *ocurrencias* de palabras, la cantidad de vocablos que hayamos reconocido será mucho menor.

El corpus se vuelve suficiente a partir del momento en que siguen creciendo las ocurrencias y es cada vez más difícil encontrar vocablos nuevos (Luis F. Lara 2006:156).

Para este autor la cantidad de ocurrencias puede no ser significativa, pero sí lo será la riqueza y variedad de textos que se incluya en el corpus.

La cantidad de palabras que conoce un alumno tiene implicaciones en su competencia comunicativa y en su capacidad de comprensión de los textos. Todos tenemos la experiencia de textos que no son comprendidos por los alumnos incluso en niveles avanzados de la enseñanza materna. Según algunos autores hay una influencia evidente de la competencia léxica o conocimiento del léxico en la comprensión escrita. Se ha establecido la existencia de un umbral o conocimiento mínimo de vocabulario que posibilita la comprensión de un texto auténtico. Dicho umbral se establece en las 3.000 familias de palabras más frecuentes frente a las 2.000 según otros autores a partir de los cuales el alumno podrá aplicar las estrategias de inferencia (Izquierdo 2004:64).

Estas apreciaciones parecen corroborarse por estudios estadísticos más recientes. En las listas de frecuencia elaboradas por Antonio Ávila las primeras 5.000 palabras del corpus cubren el 99,7% de un texto dado. Estos datos están aplicados al léxico de frecuencia del español hablado en la ciudad de Málaga. Son datos que corroboran que 5.000 palabras son suficientes para las necesidades básicas de un hablante y nos pueden ayudar a entender cualquier conversación –pues es un corpus de habla oral–.

La tabla que refleja esta cobertura de un texto por parte de las unidades más básicas es la siguiente (Ávila, 99:92):

Tabla: Cobertura total del corpus	
50 primeras palabras	63%
100 primeras palabras	73 %
500 primeras palabras	89%
1.000 primeras palabras	93%
1.500 primeras palabras	95 %
2.000 primeras palabras	96,6%
2.500 primeras palabras	97 %
.....	
5.000 primeras palabras	99,7%

Según este autor,

> Aunque el vocabulario es algo que vamos adquiriendo con el paso del tiempo, parece que las palabras más usadas que sirven de base en la comunicación están suficientemente adquiridas ya en la preadolescencia.» «De todas maneras, una afirmación como esta debe asentarse sobre principios más sólidos y bases más estables que la simple comparación entre dos listados elaborados con distinto criterio. Sin embargo, el dato nos parece interesante por la posibilidad de que la tendencia observada pueda corroborarse en estudios futuros[5].

En el corpus elaborado por Mark Davis se establece la siguiente tabla de porcentajes de aparición de las primeras 1000, 2000 y 3000 palabras recogidas en su corpus en tres registros diferentes (Davis 2005:109):

Table 3. Percent coverage of tokens by groups of types/lemma

	Non-fiction	Fiction	Oral
1st thousand	76.0	79.6	87.8
2nd thousand	8.0	6.5	4.9
3rd thousand	4.2	3.5	2.3
FIRST 3000	88.2	89.6	94.0

Como se puede ver en la tabla, las primeras 3000 palabras de los índices de frecuencias cubren el 88.2% de los textos de no ficción, el 89.6% de los textos de ficción y el 94% de los textos orales. Según Davis (2005) el 90% de un texto está constituido por aproximadamente 2.600 nombres, 230 verbos, 980 adjetivos y 50 adverbios. Es decir, un total de 3.800 formas. Si se añaden las palabras que indican función –determinantes, preposiciones, conjunciones, etc.–, se obtiene el dato de que con 4.000 palabras se cubre el 90% de una conversación típica[6].

Los datos no son totalmente coincidentes, pero sí son significativos en cuanto al número de unidades que aparecen en los textos y que constituyen por tanto las unidades empleadas en la mayoría de las comunicaciones. Los diferentes resultados tienen que ver con la composición del corpus y el objetivo del mismo.

La primera reflexión que se extrae de los estudios estadísticos del lenguaje es que los hablantes elaboramos gran parte de nuestros mensajes con un número muy limitado de vocablos.

[5] En nota, en Antonio Ávila (1999)

[6] Davis (205: 109) señala también las 2000 primeras palabras cubre entre el 80% y el 90% de un texto e en español, inglés y alemán. Los diferentes porcentajes se deben a diferencias en la recogida o representación de los materiales fundamentalmente.

Podríamos estar ante una demostración más de la tendencia a una cierta economía de esfuerzo lingüístico que parece afectar de manera evidente al nivel léxico. El hecho de que los hablantes utilicen un número relativamente bajo de palabras diversas podría indicar una cierta pobreza léxica en general. De todas maneras, esta última afirmación también requiere algunas matizaciones y, ante todo, un estudio detenido del léxico en función de algunas variables sociológicas (Ávila 99: 93).

La enseñanza de lenguas extranjeras motivó muchos trabajos sobre la frecuencia léxica, por una razón práctica. Aunque ese criterio de frecuencia ha sido matizado con el de la disponibilidad. El léxico disponible es el conjunto de palabras que los hablantes tienen en el lexicón mental y cuyo uso está condicionado por el tema concreto de la comunicación. Lo que se pretende es descubrir qué palabras sería capaz de usar un hablante en determinados temas de comunicación. Se diferencia del léxico básico en que éste lo componen las palabras más frecuentes de una lengua con independencia del tema tratado.

Mientras que el léxico básico está formado en su mayoría por palabras gramaticales, las que no lo son, pertenecen, en orden decreciente de frecuencia, a verbos, adjetivos y sustantivos de significado general. En el léxico disponible abundan los sustantivos que aluden a realidades concretas. Como ha señalado Humberto López Morales, el léxico disponible es un léxico potencial, no actualizado; mientras que la frecuencia sólo trabaja con léxico actualizado. La suma de ambos tipos de léxico constituye el léxico fundamental de una lengua.

La llegada de los métodos comunicativos en los años 80 hizo que se abandonara el criterio de frecuencia y se diera más importancia a la intencionalidad comunicativa. El aprendizaje del léxico se circunscribe en estos métodos a los textos reales y a situaciones comunicativas que condicionan el aprendizaje. Desde el punto de vista de la frecuencia léxica, este método no interfiere en el conocimiento del léxico más frecuente de una lengua, pues la exposición a textos reales –*realia*– garantiza la repetición de ese léxico básico. Si se dominan las 5.000 palabras más frecuentes, se conoce el 90% de un texto y el resto, el léxico incidental, puede ser trabajado específicamente o con la consulta del diccionario.

En la enseñanza de lenguas extranjeras se tiene en cuenta el criterio de frecuencia en el método de lecturas graduadas y textos simplificados. La lectura facilita el método de aprendizaje. Estas lecturas graduadas recogen entre 300 y 2.500 palabras según el método de la editorial que las publica.

En la enseñanza de la lengua materna ha primado la enseñanza de lo escrito y de otros aspectos como el histórico literario, y no se ha tenido en cuenta el criterio de frecuencia. Todo ello se relaciona con el dominio léxico y el desarrollo psicosocial de nuestros alumnos. Parece lógico pensar que la escuela debe garantizar el uso y manejo de un número de unidades léxicas aceptable para todos los alumnos con el que se pueda

mantener una comunicación fluida. Ese número se establece en torno a las 5.000 unidades. La estadística así lo establece para los hablantes de una lengua –el número coincide tanto en los listados sobre lengua oral como en los listados de lengua escrita– . Esa cantidad parece pequeña para un hablante nativo, porque a ella habrá que añadir el léxico disponible y los léxicos de especialidad, que no son léxicos frecuentes.

Estas consideraciones nos deben llevar a replantearnos la cantidad de vocabulario que deben aprender los alumnos. No se trata de grandes listas de palabras y, además, las que establecen el entramado comunicativo ya están presentes en su lexicón mental. Además una palabra nos se aprende de forma aislada sino en su contexto y en la estructura lingüística de la que forma parte. De ahí que la reflexión gramatical sobre la estructura del lenguaje parece imprescindible. Pero ciñéndonos al aspecto léxico, habrá que enseñar también cómo adquirir el léxico más incidental, de ahí que la enseñanza del manejo del diccionario se convierte en un aspecto fundamental.

Con todo, el elemento cuantitativo es imprescindible e incluso hay investigaciones que corroboran que la selección de unidades léxicas debe ser adecuada. Según señala Izquierdo (2004:162), en Francia se han hecho estudios que demuestran la existencia de una sobrecarga léxica en el marco escolar francés. Por ejemplo los recuentos de Lyeury y su equipo señalaban la cantidad excesiva de unidades léxicas en los manuales escolares –17.000 unidades léxicas en 4ème y 25.000 unidades en 3ème–. Este autor consideraba que, para facilitar el aprendizaje, se debería, en primer lugar, desinflar el programa léxico de todas las asignaturas, y en segundo lugar, multiplicar las fuentes de información sobre un mismo tema con el fin de fijarlo en la memoria.

CONCLUSIONES

Los estudios sobre la frecuencia léxica encuentran en la actualidad una potente herramienta en la lingüística de corpus. El aspecto cuantitativo del estudio léxico es fundamental en sus aplicaciones pedagógicas, si bien debe ser matizado por otros conceptos tales como el de disponibilidad o el de riqueza léxica. Los resultados de los más recientes corpus del español arrojan datos cuantitativos similares en cuanto al número de unidades que manejan los hablantes en la comunicación cotidiana. Esos datos, aunque puedan sorprender, pueden indicar la existencia una sobrecarga léxica en la enseñanza y aprendizaje de la lengua. La respuesta a la pregunta de cuánto vocabulario hay que conocer para aprender una lengua no da como resultados, basándonos en estos datos, una cifra demasiado elevada. Es una cantidad que se puede graduar en su adquisición, por etapas educativas, y que debe garantizar el conocimiento de un determinado léxico por parte de los alumnos. El léxico incidental deberá aprenderse con la consulta de obras lexicográficas.

No existe una planificación en la enseñanza y aprendizaje del léxico porque la utilización de textos reales garantiza el conocimiento del léxico más frecuente. A mayor número de textos trabajados, más garantía de que el léxico frecuente de la lengua se conozca mejor, en mayor diversidad de situaciones. Sin embargo, los datos cuantitativos nos ayudan también a entender mejor la naturaleza de los textos que manejamos y la naturaleza del lenguaje, que no funciona, tampoco en el nivel léxico, con un número ilimitado de unidades.

REFERENCIAS BIBLIOGRÁFICAS

ALVAR EZQUERRA, Manuel (2005) «La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera», en *Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: deseo y realidad*, *Actas del XV Congreso Internacional de ASELE*. Sevilla, Universidad de Sevilla, pp. 19-33.

ALMELA, R., CANTOS, P., SÁNCHEZ, A., SARMIENTO, R. ALMELA, M. (2005): *Frecuencias del español contemporáneo. Fundamentos, metodología y análisis*. Madrid, SGEL.

ÁVILA MUÑOZ, Antonio Manuel (1999) *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga, Servicio de Publicaciones de la Universidad de Málaga.

DAVIS, Mark (2005): «Vocabulary Range and Text Coverage: Insights from the Forthcoming *Routledge Frequency Dictionary of Spanish»,* en *Selected Proceedings of the 7th Hispanic Linguistics Symposium*. Ed.David Eddington. Somerville, MA, Cascadilla Proceedings Project. , pp. 106-115.

— (2006) *A frecuency dictionary of Spanish*. New York, Routledge.

GARCÍA HOZ,Víctor (1953) *Vocabulario usual, común y fundamental.* Madrid, CSIC.

IZQUIERDO GIL, Mª. del Carmen (2004) *La selección del léxico en la enseñanza del español como lengua extranjera. Su aplicación en el nivel elemental de estudiantes francófonos*. Universidad de Valencia, Servicio de Publicaciones.

JUILLAND, A. y CHANG-RODRÍGUEZ (1964) *Frecuency Dictionary of Spanish Words*. The Hague, Mouton.

JUSTICIA, Fernando (1995) *El desarrollo del vocabulario. Diccionario de frecuencias*. Granada, Universidad de Granada.

KENISTON, H. (1920) «Common words in Spanish», *Hispania*. 3, pp. 85-108.

LIEURY, Alain (1992) *Des méthodes pour la mémoire*. Paris, Dunod.

LÓPEZ MORALES, Humberto (1999) *Léxico disponible de Puerto Rico*. Madrid, Arco Libros.

— (2000) «La vitalidad del léxico» en Manuel Alvar*, Introducción a la lingüística español.* Barcelona, Ariel, pp. 523-545.

MOLERO HUERTAS, Jorge (1999) *El vocabulario usual de nuestros alumnos*. Granada, Consejería de Educación y Ciencia de la Junta de Andalucía.

MORALES, Amparo (1986) *Léxico básico del español de Puerto Rico*. San Juan de Puerto Rico, Academia Puertorriqueña de la Lengua Española.

MÜLLER, Charles (1964) *Esai de statistique lexicale*. Paris, Bibliotèque française et romaine.

NATION, I.S. P. (1990) *Teaching and Learning Vocabulary*. Boston, Heinle Publisher.

SÁNCHEZ, Aquilino (dir.) (2001) *Gran diccionario de uso del español actual*. Madrid: SGEL.

TERRÁEZ GURREA, Marcial (2001) *Frecuencias léxicas del español coloquial: análisis cuantitativo y cualitativo*. Valencia, Universidad de Valencia.

TORNDIKE, Edgard L. (1921) *The Teacher Word Book*. New York Teachers College, Columbia University.