Los recuentos léxicos con indicación de la frecuencia en español

M. del Carmen Ávila Martín
Universidad de Granada

1. INTRODUCCIÓN

Las listas de palabras basadas en la frecuencia léxica se han utilizado en español desde hace más de un siglo. Estos materiales se han considerado fundamentales para la enseñanza de lenguas, y, también, en la descripción lingüística. Más recientemente, la creación de grandes corpus del español proporciona listados de frecuencias que se elaboran con métodos informáticos.

La presentación de estos listados en forma de diccionarios de frecuencias se ha realizado en español con métodos manuales desde principios del siglo XX. La irrupción de la informática y la creación de corpus para el estudio del lenguaje han dado lugar a una conjunción entre diccionarios, corpus y listas de frecuencias, que, además, se debe poner en relación en el ámbito del estudio de los discursos, en los que el género textual y la temática de los mismos también están implicados. El desarrollo de la Lingüística de Corpus está propiciando la aparición de este tipo de obras con mayor rapidez.

El objetivo que se propone es, en primer lugar, dar cuenta de las recopilaciones léxicas con indicación de la frecuencia que se han publicado en español, desde una perspectiva historiográfica; en segundo lugar, se realiza un análisis de esos recuentos léxicos y las características que presentan. Las distintas recopilaciones proceden de corpus elaborados con diversos materiales, lo que marca diferencias entre unas listas y otras. La representatividad de los corpus es una de las cuestiones que se plantean, así como la posibilidad de creación de corpus específicos según las utilidades que se quiera dar a las listas obtenidas.

Por otro lado, la importancia de la frecuencia léxica se debe poner en relación con otros aspectos del funcionamiento léxico. Así, junto con la frecuencia, hay que considerar la disponibilidad, la morfología o sus combinaciones. En cierta manera, los listados de frecuencias ponen de manifiesto que el número de unidades léxicas no es inabarcable, sino que la comunicación se produce con un número limitado de unidades. Además, el léxico que aparece en ellas está también relacionado con el tipo de corpus del que se extrae.

2. LOS RECUENTOS DE PALABRAS CON INDICACIÓN DE LA FRECUENCIA EN ESPAÑOL

Las listas de palabras con indicación de la frecuencia en español han estado relacionadas fundamentalmente con la enseñanza de las lenguas. Parece lógico pensar que las palabras más frecuentes se deben incluir antes en el aprendizaje que las menos frecuentes. Por otro lado, la frecuencia se utiliza también como un elemento descriptor de la lengua y, más recientemente, se ha puesto en relación con determinados aspectos

del funcionamiento del lenguaje.

En este sentido, Almeida (2013) tiene en cuenta la frecuencia de las palabras para el análisis los procesos de variación y cambio, en concreto en las realizaciones fonéticas de la /d/ intervocálica y la /ʧ/ en el español hablado de Santa Cruz de Tenerife (Canarias). Este autor mantiene la tesis de que el fonema /d/ se elide más en palabras frecuentes que en palabras no frecuentes. Parece confirmarse lo que este autor denomina «la hipótesis de la frecuencia» en español, y también en otras lenguas; aunque como la diferencia no es muy grande, se plantea que posiblemente existan otros factores lingüísticos, sociales y actitudinales en esos procesos. En el caso del fonema africado sordo /ʧ/ también señala una diferencia significativa en las realizaciones tensas de esta consonante en el grupo de palabras de alta frecuencia (Almeida 2013).

En la enseñanza de lenguas, la frecuencia léxica ha sido objeto de interés desde muy temprano. En el siglo XIX Eduardo Benot en su gramática, *La arquitectura de las lenguas* (1887), señalaba que no hablamos con un número muy grande de unidades léxicas. A pesar de que los diccionarios son obras que intentan recoger el mayor número de elementos léxicos como sea posible, hablamos con un número inferior de unidades, lo que hace suponer que en su época ya se habían realizado estos recuentos. Este autor presenta, en época tan temprana, referencias a resultados en los recuentos léxicos que se habían realizado ya en ese momento.

> El hombre, con sus limitadas facultades, no podría hablar si para cada objeto y para CADA UNA de sus mudanzas hubiese querido tener una palabra especial.
>
> Y todavía el portento es más sorprendente y se hace más digno de explicación, atendiendo a que las 40.000 palabras de una lengua son propiedad del Diccionario solamente, pero no de ningún poeta ni del orador más abundante.
>
> El niño habla con muy pocos vocablos: su vocabulario oscila entre 300 y 400 términos muy usuales. Lenguas hay en que no existen tantas raíces. El libreto de de una ópera italiana no pasa regularmente de 650 voces. Del gran poeta Racine se ha dicho que le bastaron 1.200 vocablos para escribir todas sus tragedias (lo que parece cuestionable). Contados con celo religioso los vocablos de la Biblia, correspondientes al Antiguo Testamento se ha visto que son 5.642. Un periodista elegante apenas hace uso de más, y un hombre de buena sociedad no emplea nunca tantas, ni con mucho, en su conversación. El orador más copioso suele no llegar a 7.000; y, por exceder este número en algunos millares, se citan como portentos de facundia y de riqueza, entre todos los escritores, a CERVANTES, a LUTERO y a Shakespeare, especialmente a este último, cuyo vocabulario se acerca a 15.000.- ¿Cómo, pues, hablamos? (Benot, 1887: 31-32)

Las primeras referencias sobre un recuento léxico que permitiera realizar estas apreciaciones se remonta a 1898 en el que F. W. Kaeding, publicó un listado de frecuencias del léxico alemán basado en un corpus de once millones de palabras, fruto de la recopilación y análisis manuales (Atkins y Zampolli 1994:21, en Almela *et alii* 2005:12). Esta obra, según Rojo (2008), se realizó con la ayuda de miles de colaboradores y estaba destinada a mejorar la taquigrafía y su enseñanza. Por ese motivo, recogía también las combinaciones de letras y las sílabas más frecuentes, datos que siguen interesando en muchos recuentos de frecuencia en la actualidad por su interés en Lingüística Clínica.

En el siglo XX se realizaron varios recuentos léxicos con indicación de la frecuencia que se sucedieron con cierta lentitud puesto que se realizaban con métodos manuales de recopilación. Según Almela *et alii* (2005), "es conocida la moda o la pasión por los listados de frecuencias léxicas en la década de los años 30 y posteriormente en la década de los años sesenta (propiciada por la metodología audio-oral en la enseñanza de lenguas extranjeras)" (Almela *et alii* 2005: 12). Estos autores se remiten a las listas de palabras creadas por Thordike y Lorge en 1944 (*The Teacher's wordbook of 30.000 words*) y la *General Service List of English Words* de Michael

West (1953).

Por su parte, los estudios de frecuencias en Francia dieron también como resultado la publicación de listas de vocabulario para la enseñanza de lenguas. La distinción entre *léxico frecuente* frente a *léxico disponible* surgió de los estudios realizados en el país galo a mediados de los años cincuenta. El desarrollo de los estudios de frecuencia léxica para ser aplicado a la enseñanza de lenguas puso de manifiesto que determinadas palabras, que no eran frecuentes, formaban parte del léxico disponible de los hablantes según la temática de la que se tratara. Estas apreciaciones han tenido amplia repercusión entre los investigadores y se han desarrollado numerosos trabajos sobre el léxico disponible en el proyecto panhispánico dirigido por Humberto López Morales (*cf*. http://www.dispolex.com).

En el ámbito de la enseñanza del español, contamos con algunos estudios de conjunto sobre las obras disponibles en el siglo XX, recuento que se realizó en los *Boletines de la Asociación Europea de Profesores de Español como Lengua Extranjera*, antecesora de ASELE. En esta publicación Ezquerra (1974) recoge hasta finales de los años setenta un total de nueve trabajos ordenados cronológicamente: Buchanan (1927), Keniston (1929), Rodríguez Bou (1952), García Hoz (1953), Juilliand y Chang Rodríguez (1964) y Rojo Sastre (no indica el año). Y en un trabajo posterior (Ezquerra 1977) recoge tres obras más: Márquez Villegas (1975), García Hoz (1976) y Díaz Castañón (1977). En estos recuentos este autor analiza los siguientes parámetros: la extensión del corpus, fuentes literarias, fuentes no literarias, objetivos más bien pedagógicos, noción de disponibilidad, corpus de frecuencia, corpus disponible, homogeneidad geográfica, homogeneidad temporal, norma claramente definida, recuento separado de flexiones, recuento de todos los vocablos e índice de clasificación.

De entre todas las obras que cita, deja fuera de sus comentarios el trabajo de Keniston, *Spanish Idiom List*, (1929) por tratarse "de una lista de frases e idiomatismos" (1977:52). Sin embargo, Keniston es, precisamente, uno de los primeros autores en recoger listas de palabras usando criterios estadísticos y que publicó en *Hispania* (revista de la *American Association of Teachers of Spanish and Portuguese*) a comienzos de la década de 1920 (Ávila Martín 2010). Keniston (1920) preparó un inventario de lengua cotidiana a partir de un corpus en el que se incluyen obras de teatro, periódicos, revistas, cuentos y novelas. En la lista aparecen 1322 palabras ordenadas en ocho grupos por el índice de frecuencia y con indicación de la categoría gramatical. Este trabajo tiene interés, no solo por ser temprano, sino porque en él se establece también el dato de que, según 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. Es decir, los datos de Keniston corroboran las afirmaciones realizadas por Benot unas décadas antes en las que indicaba que no hablamos con un número excesivo de palabras.

El desarrollo de la informática a partir de los años 80 y 90 supone la elaboración de una gran cantidad de corpus con diferentes finalidades. Como ha señalado G. Rojo (2008), la historia de la Lingüística de Corpus se inicia cuando se construye el primer corpus construido para una computadora, y que este autor sitúa en 1964. Esta disciplina proporciona métodos rápidos y eficaces para la obtención de listas de frecuencias. La evolución tecnológica, así como la finalidad de los corpus elaborados ha dado lugar a que aparezcan a partir de esta época listados de frecuencias con diferentes objetivos.

Las listas para la enseñanza de la lengua (Morales 1986) dan paso también a diversos corpus destinados al estudio del vocabulario de los jóvenes (MEC 1989, Justicia 1995), así como a las investigaciones en el ámbito de la adquisición del lenguaje. El recuento realizado por Alameda y Cuetos (1995) se construyó con dos

millones de palabras de diferentes tipos de textos escritos entre 1978 y y1993. Una segunda lista se publicó en la universidad de Barcelona, LEXESP (Sebastián, Martí, Carreiras y Cuetos 2000) a partir de un corpus de cinco millones de palabras.

En el ámbito de la Psicolingüística, la frecuencia se considera un aspecto importante en el reconocimiento experimental de palabras, base de numerosas investigaciones. El tamaño de los corpus es un elemento esencial, pero el registro en el que se basa el corpus se considera más importante que el tamaño. Así en un trabajo posterior de Cuetos (2011) se propone que es mejor utilizar listados de frecuencias procedentes de los subtítulos de películas de cine o televisión en el reconocimiento de experimental de palabras que el uso de los listados de frecuencias basados en libros y prensa.

Efectivamente, no recogen el mismo léxico los listados de frecuencias de un corpus oral, que un corpus del habla infantil. Tampoco recogen el mismo tipo de unidades los listados de frecuencias realizados a partir de los textos escritos por los niños, como en el recuento del MEC (1989), o Justicia (1995), como los realizados a partir de los textos que lee un niño, como el caso de Martínez y García (2004). En el recuento realizado por estos autores se señala que las palabras más frecuentes son similares en los recuentos de lengua escrita, aunque puede haber diferencias en las frecuencias relativas. Este dato depende mucho de cómo se ha hecho la selección de los recuentos. Y las diferencias con el diccionario de Alameda y Cuetos (1995), realizado a partir de muestras de lenguaje dirigido a adultos, están en los términos que denominan de clase abierta. En relación al diccionario de Justicia (1995) señalan estos autores que entradas como *fray*, *claramente*, *decimal*, *fortuna* o *frasco* aparecen una sola vez y se sitúan entre el 10% de palabras menos frecuentes. Y señalan que "todas ellas, sin embargo, aparecen entre las primeras 3.500 palabras más frecuentes en nuestro diccionario basado en lo que los niños leen" (Martínez y García 2014:39).

A finales del siglo XX se publican también léxicos de frecuencias de la lengua hablada, como el de Ávila Muñoz (1999) sobre el léxico de frecuencias de la ciudad de Málaga, o el de Terráez (2001) sobre la lengua coloquial.

En la actualidad, el número de corpus ha aumentado considerablemente. Se pueden encontrar diversas recopilaciones de corpus del español que nos muestran el interés por estas recopilaciones consideradas como el punto de partida de cualquier investigación sobre el lenguaje (Briz y Alberda 2009, Rojo 2016 y en la página de Llisterri de la Universidad de Barcelona). Los corpus que se han realizado y que se están realizando presentan características muy diversas y se clasifican en función de la extensión, el tipo de textos tratados, el tipo de transcripción o la finalidad de su elaboración. Muchos se han finalizado y se pueden consultar en la red, y otros forman parte de proyectos que no se han completado todavía. No todos presentan las listas de frecuencias léxicas, aunque algunos las publican en listas complementarias y son pocos todavía los que analizan los datos que proporcionan. En otros casos, las listas de palabras no se han publicado o no se pueden consultar por diversos motivos, sea por la propia presentación del corpus, o porque, en algunos casos, tienen un valor comercial, o no forman parte del objetivo del corpus.

Por ejemplo, se dispone ya de un número superior al medio centenar de corpus orales del español o que utilizan muestras orales (según Briz y Alberda 2009), aunque no todos están concebidos del mismo modo, ni presentan las mismas características. Como señalan estos autores, la utilidad de los corpus está relacionada con la intencionalidad para la que han sido concebidos.

Las listas de frecuencias que presentan los corpus están muy relacionadas con la finalidad y la selección textual que se ha realizado para la elaboración del corpus. De

ahí que las listas presenten diferencias. La comparación entre estas listas de frecuencias nos proporciona ejemplos de esa disparidad, aunque, por otro lado, muestra ciertas coincidencias en el funcionamiento del lenguaje.

Los corpus elaborados con finalidades lexicográficas aportan también muchos elementos de reflexión sobre el funcionamiento cuantitativo del léxico. Hablamos con un número reducido de unidades. Y tampoco hace falta un corpus muy extenso para obtener resultados representativos en cuanto al número de unidades. Este rasgo ya era señalado por Alvar Ezquerra (2005) a propósito del diccionario de Juilland y Chang-Rodríguez (1964) y ha sido señalado para otros corpus, por ejemplo, el Corpus del español de México de L. F. Lara, o el corpus Cumbre.

Otro aspecto señalado es que la frecuencia no es el único aspecto que se debe tener en cuenta para valorar la importancia comunicativa de una unidad léxica. Como indica Lara (2006:165) sobre la composición del Corpus del Español Mexicano Contemporáneo (CEMC), hay que tener en cuenta también la dispersión y el uso. Este autor utilizó estos cálculos estadísticos para la elaboración del *Diccionario del español de México* (DEM) y señalaba que un corpus se vuelve suficiente a medida que siguen creciendo las ocurrencias y es cada vez más difícil encontrar vocablos nuevos.

Por otro lado, Alvar Ezquerra (2005) en sus recuentos del corpus Vox Biblograf (de 10 352 337 millones de palabras) aporta también datos que corroboran que con un número muy reducido de unidades se alcanza casi el 90% de un texto. Este autor plantea las diferencias que puede haber entre unos corpus y otros debido a que están realizados de forma muy dispar. También señala que las palabras más frecuentes coinciden, pero no aparecen en el mismo orden en todos los listados.

Alvar Ezquerra señala diferencias entre unos recuentos y otros, en algunos casos, por razones históricas, puesto que el léxico es cambiante. Entre otros factores distorsionantes señala también la representatividad de la muestra, y la dispersión en los campos de interés. Ese dato, que no suele estar disponible en los corpus, no permite adscribir una unidad léxica a un ámbito concreto que es un aspecto que interesa a los lexicógrafos para poder marcar el uso de las unidades léxicas en ámbitos determinados.

Efectivamente la comparación entre las listas de frecuencias de distintos corpus muestra que no solo las palabras frecuentes son pocas, si bien es verdad que son las más polisémicas, sino que un número elevado de palabras aparece solamente una vez en el conjunto de un corpus.

Es lo que se constata en los análisis que presentan Almela *et alii* (2005) del corpus Cumbre que consta de 20 millones de palabras y cuyo objetivo es también de carácter lexicográfico. Los autores de la obra han publicado análisis de los materiales que vienen a coincidir con las apreciaciones realizadas por otros autores en cuanto al funcionamiento cuantitativo del léxico. Por ejemplo, el hecho de que todas las lenguas naturales utilizan un número reducido de unidades en la comunicación. O que un número elevado de palabras aparece solamente una vez en el conjunto de un corpus.

> Analizando la distribución de las formas en el Corpus *CUMBRE* se observa que casi el 42% de todas las palabras diferentes aparecen solamente una vez en el conjunto. Este modelo de distribución léxica que se da en todas las lenguas naturales- muestra que el léxico del que nos valemos tiende a concentrar su peso comunicativo en un reducido número de palabras. Y así lo confirma el hecho de que las formas que más aparecen en los textos, en el tramo de frecuencias de más de mil ocurrencias, son solamente 1.851 (de las cuales, la forma *de* por sí sola acapara más de un millón) (Almela *et alii*. 2005:2).

También se había señalado por parte de los autores de otros corpus, como Lara (2006), que a medida que avanza el número de palabras de un corpus no aparecen palabras diferentes.

Según se ha demostrado ya (Sánchez y Cantos 1997), el aumento de nuevas voces y de las frecuencias de las voces no es lineal, sino que, más bien, sigue la curva de una hipérbola (sic). Eso significa que cuanto más aumenta el número de palabras de un corpus, más disminuye el número de palabras diferentes (Almela et alii 2005:3).

Para los autores del *Cumbre*, la frecuencia léxica es un dato importante porque demuestra que a pesar de que los datos sobre las unidades léxicas recopilan para el español un número muy elevado de unidades, "el hecho es que la persona adulta y con buen nivel de educación no suele valerse de más de 20.000 palabras en sus contextos de comunicación. Y desde luego el lenguaje oral y habitual dista mucho de esa cifra" (Almela *et alii.* 2005: 4).

Las frecuencias también interesan a los lexicógrafos para realizar una organización de las acepciones que esté más acorde con el uso. Los autores del corpus Cumbre señalan que es muy difícil que se puedan recoger todas las muestras de uso de una lengua en un momento histórico concreto, pero es importante que el corpus sea representativo.

La selección del material que se realice para construir el corpus siempre determina sus características. Estos autores comparan su propia muestra con la recopilación del léxico de García Hoz (1953). Las diferencias que encuentran en el ámbito de la letra A (que equivale según estos autores a un 10% del total de corpus) alcanzan una representatividad del 58%. Es decir, hay un 42% del léxico que no aparece en el corpus Cumbre y sí en el de García Hoz. Las diferencias se justifican por el momento histórico en el que se hizo el corpus y los documentos incluidos, lo que da una lista de palabras que incluye los "términos *adorar*, *adornar*, *agrupación*, *alabar*, *altar*, *amado*, *apóstol*, *arca*, *artillería*, *ascender*,... mientras que no figuran voces como *aborto*, *accidente*, *acceso*, *actor*, *adiós*, *aeropuerto*, *agrícola*, *alarma*, *amenaza*, *antena*, *asamblea*, *asesinato*, *asesino*, *atentado*, *autonomía*, *avión*..." (Almela *et alii* 2005: 7).

La composición del corpus debe ser representativa, pero hay otros muchos factores que influyen en la aparición de unas unidades u otras. En el ejemplo de estos mismos autores, las diferencias cronológicas entre un corpus y otro y los cambios culturales y sociales marcan diferencias importantes en las unidades que aparecen.

Por otro lado, los análisis que realizan de su propio corpus les permiten, por ejemplo, demostrar que existe una relación "inversamente proporcional entre la frecuencia y el número de palabras con esa misma frecuencia". Asimismo, como se muestra en un gráfico, "el número de palabras poco frecuentes (235.680 con frecuencia baja) es muy elevado, mientras que el de palabras muy frecuentes es realmente bajo (1.196 con frecuencia muy alta)" (Almela *et alii.* 2005:20). Estas apreciaciones coinciden de nuevo con los análisis que se han realizado sobre otros corpus en sus aspectos cuantitativos. También se corrobora otra afirmación sobre la correlación entre lemas y número de significados puesto que en la "correlación entre frecuencia general y polisemia: las palabras tienden a presentar un abanico de sentidos tanto más variado cuanto mayor sea su frecuencia de uso" (Almela *et alii.* 2005:27).

Las diferencias entre unos corpus y otros no están solo en la composición y en la selección de materiales. También en las medidas de frecuencia que se señalan para las diferentes unidades. Por ejemplo, en el corpus Cumbre, a diferencia de otros corpus, el rango no se considera una medida totalmente fiable, y se prefiere utilizar la desviación estándar. Según sus autores:

"Este índice es mucho más preciso que el rango ya que considera no solamente los datos extremos, sino todos los valores observados con respecto a la media de los mismos. Para su cálculo es

preciso obtener primero la media de los valores observados." […] "Cuanto mayor sea la desviación estándar, más dispersos o menos homogénos serán los valores que hemos observado. Si la desviación estándar es pequeña o menor en comparación con otra, se entiende que los datos observados son más homogéneos, puesto que están más concentrados en torno a la media" (Almela *et alii*. 2005: 105-106).

En la actualidad, los grandes corpus del español están recopilando gran cantidad de datos, con cifras muy amplias. Los datos que se están publicando también nos ofrecen información sobre la frecuencia de las palabras y los resultados cuantitativos que nos ofrecen, muestran un panorama similar a las observaciones hechas por otros autores.

El CREA en su última versión (2008) cuenta con ciento sesenta millones de formas. La RAE publica en su página web la lista de las palabras más frecuentes de este corpus, la lista de las 1000, 5000 y 10 000 palabras más frecuentes del español. El CORPES XXI, por el momento no ha presentado la lista de frecuencias, aunque su extensión es mayor, según datos de la RAE consta de 225 millones de formas en su versión de 2016. Según G. Rojo (2008:19) en este corpus a pesar de su extensión se produce el mismo fenómeno que ya se ha señalado:

> suponíamos que un alto número de formas de frecuencia igual a uno con un corpus de tamaño x pasarían a ser formas de frecuencia igual a dos con un corpus de tamaño 2x. Algo de eso sucede, sin duda, pero ocurre también que la entrada de nuevos textos produce la entrada de un número igualmente alto de formas con frecuencia igual a uno (Rojo 2008:19).

La proporción del léxico de las lenguas se ha descrito en un modelo matemático denominado la ley de Zipf, que parece cumplirse en todas las lenguas. Y según Rojo (2008) este funcionamiento está también descrito en la "La llamada *ley de Pareto*, conocida también como *ley del 80 / 20*, [que] establece que en la mayoría de las distribuciones el 80% de los efectos se puede explicar con el 20% de las causas. Se aplica a la distribución de rentas de un país, la cuenta de resultados de una empresa, las ventas de libros, etc.". Es decir, el léxico de las lenguas presenta cuantitativamente hablando ciertas regularidades en su composición.

Otro de los grandes corpus del español, el de Mark Davis, que se puede consultar en línea, contiene, en su primera versión (2001), 100 millones de palabras entre el siglo XIII y el siglo XX. Y en su nueva edición (2016) contiene "casi dos mil millones de palabras de páginas web de 21 diferentes países de habla hispana. Este corpus permite hacer búsquedas en textos en español muy recientes (los textos se recopilaron en 2013 y 2014) y comparar los diferentes dialectos".

En la página en línea de consulta de este corpus se compara con otros corpus del español para resaltar sus virtudes. En comparación con el Corpes XXI, que todavía está en construcción, se resalta el menor número de unidades del corpus de la RAE, entre otros problemas de interfaz.

En relación a otros corpus más extensos, como el *Sketch Engine* (con 9,6 billones de palabras y el *COW* (*Corpora from de Web*) que posee el doble de lemas que el de Davis; este autor señala, sin embargo, que son corpus con problemas de lematización y sin anotar, por lo que no son fácilmente utilizables, entre otras razones.

La lista de frecuencias que presenta Davis contiene 20 000 lemas y está basada en la primera edición de su corpus (https://www.wordfrequency.info/spanish.asp). Como el mismo autor señala, el tamaño no es todo en un corpus, y la anotación y la posibilidad de obtener partes del corpus según el género, el país, etc. deben también permitir obtener la lista de palabras más frecuentes en cada uno de esos campos.

La revisión de estas listas en comparación con otros corpus nos muestra también diferencias en las unidades que aparecen. Estas diferencias se pueden señalar también

en las unidades gramaticales (Ávila Martín 2017) pues la palabra más frecuente en el CREA es *de*, mientras en el corpus de David es *el*, y *de* no aparece hasta el cuarto lugar.

Los primeros puestos de la lista están ocupados, como es sabido, por palabras de carácter gramatical o funcionales, aunque no se hace distinción de las diversas categorías de algunas unidades. Si se consideran las unidades léxicas que ocupan los primeros puestos en el CREA, se observa que tienen que ver con determinadas formas verbales con usos muy variados como *era* (45) y *hay* (57). También encontramos unidades referidas a elementos de tiempo: *años* (47), *tiempo* (70); y muchas de las que aparecen entre las 200 palabras más frecuentes del español están relacionadas con temáticas referidas a la forma de gobierno y de la ordenación político social: *vida* (76), *gobierno* (86), *país* (98), *estado* (104), *nacional* (141), *trabajo* (142), *política* (152), *partido* (171); también destaca entre las unidades léxicas que ocupan las primeras posiciones *España* (134) y *Madrid* (140).

En el corpus de Davis, que hemos consultado parcialmente, *nacional* (131), *político* (141), ocupan puestos relevantes. La forma de tratamiento *vos* aparece en una posición destacada *vos* (741), mientras que en la lista de frecuencias del CREA no aparece hasta la posición 2573. Este dato corrobora el hecho de que la lista de palabras resultante de un corpus está muy condicionada por la naturaleza del corpus utilizado.

La recopilación de unidades léxicas presenta además otros problemas en función de la utilidad que se le quiera dar al uso de esas listas. Tanto si se utilizan para la enseñanza de lenguas como si se utiliza en estudios de Psicolingüística, existen ciertos aspectos del funcionamiento léxico que deben ser tenidos en cuenta. Entre esos aspectos se encuentra el componente morfológico, la combinación de unidades y sus características sintácticas. Las investigaciones recientes (Nation 2017) ponen de manifiesto que las listas de palabras para usos didácticos deben considerar la frecuencia como un factor determinante, pero deben también tener en cuenta la morfología, la inclusión de prefijos y sufijos, las familias de palabras, la representación de la hominimia y la polisemia, o la inclusión de la fraseología.

La elaboración de listas de palabras con diferentes finalidades tendrá que tener en cuenta diversos factores para poder ser eficaces. El estudio y análisis de los listados de frecuencias nos ofrecerá información sobre el uso del lenguaje con datos mucho más precisos que hasta el momento actual.

3. CONCLUSIONES

Las diferencias en las listas de palabras tienen que ver con el material recogido, los criterios de corrección que se introducen, la naturaleza de los textos seleccionados y el objetivo para el que se ha realizado el corpus. El desarrollo de los métodos informáticos está propiciando que se puedan realizar grandes corpus pero también es necesaria la realización de corpus más pequeños que se acerquen a determinadas variantes para analizar un dato específico. En estos corpus empieza a ser más interesante que sean etiquetados, homogéneos y de fácil acceso, lo que presenta todavía dificultades de diverso tipo.

En cuanto al funcionamiento del léxico las listas de frecuencias y su comparación nos están mostrando que usamos un número de unidades limitado. A pesar de que el léxico de las lenguas está constituido por un número muy extenso de unidades, las que utilizamos para la comunicación están formadas por un número más reducido.

Las listas de palabras ordenadas por su frecuencia están directamente relacionadas con el corpus que se utilice, y el resultado depende del tema que se esté tratando, especialmente en lo que se refiere a las palabras menos frecuentes, que están

relacionadas con la temática del corpus. De este modo, las palabras menos frecuentes solo aparecen una vez y responden a los temas tratados en el corpus. La comparación de los listados de frecuencias muestran diferencias según los métodos y los materiales utilizados. También se han señalado constantes en el funcionamiento cuantitativo del léxico.

Finalmente, el funcionamiento léxico incluye también aspectos morfológicos y sintácticos que deben ser tenidos en cuenta en la selección de las unidades que se presentan en las listas, según la utilidad que se les quiera dar.

4. REFERENCIAS BIBLIOGRÁFICAS

4.1.Bibliografía general

Almeida, M. (2013): "La frecuencia de las palabras en los procesos de variación y cambio". *RSEL* 43-2, 37-62.

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R. & Almela, M. (2005): *Frecuencias del español*: *diccionarios y estudios léxicos y morfológicos*. Madrid: Universitas.

Alvar Ezquerra, M. (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, 19-33.

Ávila Martín, M. C. (2010): "Estadística y Lingüística de Corpus: implicaciones pedagógicas en la enseñanza y aprendizaje del léxico". *Cauce* 33, 163-165.

Ávila Martín, M. C. (2017): "Las listas de palabras: validez de la frecuencia en la enseñanza ELE". En *Actas del XXVIII Congreso Internacional de ASELE. Léxico y cultura en LE/L2: corpus y diccionarios*. Tarragona: Universitat Rovira i Virgili.

Briz, A. & Albelda, M. (2009): "Estado actual de los corpus de lengua española hablada y escrita: I+D". En *El español en el mundo*. *Anuario del Instituto Cervantes* 2009. Madrid: Instituto Cervantes, 165-226.

Cuetos, F., Glez-Nosti, M., Barbón, A. & Brysbaert, M. (2011): "SUBTLEX-ESP: Spanish word frequencies based on film subtitles". *Psicológica* 32, 133-143.

Davis, Mark (2005): "Vocabulary Range and Text Coverage: Insights from the Forthcoming Routledge Frequency Dictionary of Spanish". En *Selected Proceedings of the 7thHispanic Linguistics Symposium*. Ed. David Eddington. Somerville, MA, Cascadilla Proceedings Project, 106-115.

Ezquerra, R. (1974): "Los diccionarios de frecuencia en español". En *Boletín de la Asociación Europea de Profesores de Español* 10, 3-27.

Ezquerra, R. (1977): "Los diccionarios de frecuencia en español II". En *Boletín de la Asociación Europea de Profesores de Español* 16, 43-52.

Hidalgo Guerrero, F.J. (2017): *Edición crítica y estudio de la Arquitectura de las lenguas* (1887) *de Eduardo Benot*, Tesis doctoral de la Universidad de Almería.

Lara, L. F. (2006): *Curso de lexicología*. *México*: El Colegio de México.

Llisterri, J. (2018): "Los corpus de lengua oral". [En línea] <http://liceu.uab.cat/~joaquim/language_resources/spoken_res/Corpus_lengua_oral.html> [10/04/2018].

Nation, I. S. P. (2016): *Making and using Word Lists for Langage Learnign and Testing*. John Benjamins.

Rojo, G. (2008): "Lingüística de corpus y lingüística del español". *Actas del XV Congreso de la ALFAL*. Montevideo.

Rojo, G. (2016): "Los corpus textuales del español", en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia Lingüística Hispánica*. Oxon: Routledge, 285-296.

4.2. Listados con indicación de las frecuencias

Alameda, J.R. & Cuetos Vega, F. (1995): *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Universidad de Oviedo: Oviedo.

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R. Almela, M. (2005): *Frecuencias del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Ávila Muñoz, A. M. (1999): *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Buchanan, M. A. (1927): *A Graded Spanish Word Book*. Toronto: University Press.

Davis, M. (2006): *A Frecuency Dictionary of Spanish*. New York: Routledge.

Díaz Castañón, C. (1977): *Vocabulario básico del español y sus aplicaciones a la enseñanza*. Oviedo: Universidad de Oviedo.

García Hoz, V. (1953): *Vocabulario común, vocabulario usual y vocabulario fundamental. Determinación y análisis de sus factores*. Madrid: Instituto San José de Calasanz.

García Hoz, V. (1976): *El vocabulario general de orientación científica y sus estratos*. Madrid: C.S.I.C.

Juilland, A. & Chang Rodríguez, E. (1964): *Frecuency Dictionary of Spanish Words*. La Haya: Mouton

Justicia, F. (1995): *El desarrollo del vocabulario: diccionario de frecuencias*. Universidad de Granada: Granada.

Keniston, H. (1920): "Common words in Spanish". *Hispania* 3, 85-108.

Keniston, H. (1929): Spanish Idiom List. N. York: The Macmillan Company.

Márquez Villegas, L. (1975): Vocabulario del español hablado. Madrid: Sociedad General Española de Librería.

Martínez Martín, J. A. & García Pérez, E. (2004): Diccionario frecuencias del castellano escrito en niños de 6 a 12 años. Universidad Pontificia de Salamanca: Salamanca.

Ministerio de Educación y Ciencia (MEC) (1989): Vocabulario básico de la E.G.B. Madrid: Espasa Calpe.

Morales, A. (1986): Léxico básico del español de Puerto Rico. San Juan de Puerto Rico: Academia Puertorriqueña de la Lengua Española.

Real Academia Española: "Lista de frecuencias". En Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.*<http://www.rae.es> [10/01/2018].

Rodríguez Bou, I. (1952): *Recuento de Vocabulario Español*. Universidad de Puerto Rico.

Sebastián, N. & Martí, M. A., Carreiras, M. F., Cuetos, F. (2000): *LEXESP. Léxico informatizado del español*. Universidad de Barcelona: Barcelona.

Terráez Gurrea, M. (2001): *Frecuencias léxicas del español coloquial: análisis cuantitativo y cualitativo*. Valencia: Universidad de Valencia.