LÉXICOS BÁSICOS DE ESPAÑA (*LEBAES*) Y DE CANARIAS (*LEBAICan*). PROYECTOS DE INVESTIGACIÓN

PEDRO BENÍTEZ PÉREZ
Universidad de Alcalá de Henares
CLARA EUGENIA HERNÁNDEZ
Universidad de Las Palmas
JOSÉ ANTONIO SAMPER
Universidad de Las Palmas

1. Como es sabido, el léxico fundamental de una comunidad lingüística está formado por la suma del léxico básico y del léxico disponible. Este último, según la definición de Humberto López Morales (1986: 62), se configura como el caudal léxico utilizable en una situación comunicativa dada, el conjunto de términos que solo se actualizan cuando se necesita aportar una información muy precisa y específica; el léxico básico, por el contrario, recoge las palabras más frecuentes, aquellas que aparecen continuamente en los discursos, con independencia del tema que se trate. Como es lógico, en los primeros lugares de los léxicos básicos aparecen siempre las palabras gramaticales; los sustantivos, dada su menor estabilidad estadística (como corresponde a su mayor concreción semántica), son las palabras menos frecuentes. Ambos léxicos -el básico y el disponible- son absolutamente complementarios para llegar a saber la realidad léxica de una determinada comunidad de habla.

Los dos tipos de diccionarios son hoy imprescindibles para una planificación de la adquisición y aprendizaje del léxico, tanto materno como extranjero, pues constituyen una base estadística seria sobre la que proceder a seleccionar y graduar las unidades. Aunque se trate de recuentos que necesiten de ponderación cuidadosa -por reflejar el lenguaje adulto- no se concibe ya proceder a la operación de determinar el vocabulario que debe pasar a la competencia léxica de los alumnos sin disponer de unos materiales objetivos y rigurosos.

REALE, 3, 1995, pp. 9-17.

Desde luego que por encima de tales objetivos *prácticos*, se encuentra la determinación de la estructura léxica de una comunidad lingüística dada.

En nuestro país se lleva a cabo desde hace algunos años una serie de investigaciones sobre léxico disponible, coordinada por Humberto López Morales: se trata de la que realizan Pedro Benítez Pérez en Madrid, José Antonio Samper *et alii* en Gran Canaria, Francisco García Marcos y María Victoria Mateo en Andalucía, José Ramón Gómez Molina en Valencia, Maitena Etxebarría en el País Vasco, y José Antonio Bartol y Julio Borrego en Salamanca. En cuanto al léxico básico, contamos con el trabajo de Alphonse Juilland y Eugenio Chang Rodríguez 1964, *Frecuency Dictionary of Spanish Words* (FDSW) que, como es bien conocido, recoge el léxico básico de España en el periodo comprendido entre las dos guerras mundiales (1920-1940); no ha habido posteriormente ningún intento de recoger el léxico básico de años más cercanos (por ejemplo, el de la segunda mitad del siglo), cuando tan próximos estamos ya al final del XX. Por otro lado, no se cuenta hasta el momento con ningún léxico básico regional[1].

En estas páginas queremos presentar las bases teóricas y los principios metodológicos del *Léxico Básico de España (LEBAES)* y del *Léxico Básico de las Islas Canarias (LEBAICan)*, en los que ya estamos trabajando los autores de este informe. El que se pretenda elaborar un léxico básico canario no quiere decir que Canarias no esté representada, como el resto de las regiones españolas, en la muestra general de todo el país[2]. Pero pensamos que, además del listado básico para todo el territorio nacional, tiene interés también conocer el léxico básico de una zona como Canarias, de tanta importancia en la formación del español atlántico y con una relación histórica tan estrecha con el continente americano a través de los sucesivos flujos migratorios. El disponer de ambos léxicos nos permitirá llevar a cabo, entre otras cosas, fructíferos paralelos.

---

[1] En lo que concierne a Canarias, no fue propósito de los autores del *Léxico del español usual en Canarias (Ciclo inicial)* la elaboración de un léxico básico en el sentido estricto del término. Con esa recopilación se quería "ofrecer al profesor un conjunto de palabras que se entiende deben formar parte del Vocabulario Básico activo y pasivo de los niños canarios que cursan el Ciclo Inicial" (p. 7). La metodología seguida para seleccionar las aproximadamente dos mil palabras de que consta no concuerda con la requerida para la elaboración de un léxico básico del español de las Islas (se recopilan las palabras "de más alta frecuencia de uso en el *castellano común*, según los recuentos de los distintos diccionarios de frecuencia" y se añade "un par de cientos de palabras registradas o no por el DRAE y de uso generalizado entre los hablantes de las islas"). El *Léxico del español usual en Canarias* fue publicado por la Consejería de Educación del Gobierno de Canarias en 1986.

[2] A diferencia del *FDSW*, que solo recoge textos del español peninsular: "only Spanish peninsular sources were admitted" (p. 14).

Una tarea de este tipo no tendría el mismo valor si no pudiera permitir el establecimiento de comparaciones con los trabajos ya realizados, siempre que éstos ofrezcan pautas metodológicas coincidentes para garantizar la fiabilidad del cotejo. Ya sabemos que la lexicometría ha recorrido un largo camino, que ha supuesto la elaboración de criterios cada vez más rigurosos y, por tanto, más fidedignos. En esa historia los trabajos de Juilland[3] sobre el léxico de las lenguas románicas constituyen un hito indiscutido. En nuestro ámbito lingüístico, contamos también con el ejemplar trabajo de Amparo Morales, *Léxico básico del español de Puerto Rico (LEPR)*, publicado en 1986. La autora, que ha adoptado muchas de las innovaciones presentes en los diversas contribuciones de Juilland, nos aporta unas bases metodológicas y unos planteamientos teóricos que, por su rigor, se convierten en valiosas guías de las presentes investigaciones.

En esta breve exposición nos referiremos, por un lado, a los criterios que seguimos para seleccionar el material básico, con la finalidad de conseguir una muestra realmente representativa, y, por otro, a los problemas relacionados con la segmentación de la unidad de análisis, sin olvidar la información acerca de las fórmulas matemáticas que se usarán para los cálculos de los índices de dispersión compleja y de uso. Parece apropiado también que abordemos cuál ha sido la evolución de los diccionarios de frecuencia.

2. Salvo el citado diccionario de Juilland, que ha quedado sumamente anticuado en sus resultados, nada se ha hecho en el país que se asemeje a un léxico básico. Es verdad que en varias ocasiones se han emprendido trabajos de recuento léxico, pero ni sus bases teóricas ni la metodología empleada permiten hablar de este tipo de vocabulario. El *LEBAES* es, por lo tanto, una necesidad imperiosa[4].

3. La selección de la muestra se ajusta a los principios siguientes:

(a) carácter sincrónico. Se trata de una de las condiciones generalmente aceptadas en los trabajos actuales, aunque de todos es conocida la existencia de diccionarios de frecuencia que han incluido en sus materiales de base tanto textos contemporáneos como obras de siglos pasados. De acuerdo con la metodología más comúnmente seguida hoy, la etapa de la

---

[3] Además del citado *FDSW*, *Vid.* A. Juilland, P.M.A. Edwards e I. Juilland, 1965; A. Juilland, D. Brodin y C. Davidovitch, 1971; y A. Juilland y V. Traversa, 1973.

[4] Para una amplia información sobre los trabajos de léxicoestadística aplicada a la enseñanza de la lengua materna, *Vid.* Romera Castillo 1991.

lengua a la que debe corresponder el material léxico ha de abarcar un conjunto determinado de años consecutivos, que, para asegurar la vigencia del léxico recopilado, deben ser cercanos al momento de la investigación. En el *LEBAES* y en el *LEBAICan*, hemos seleccionado un periodo de 25 años, que empieza en 1970 y termina en 1994, ambos inclusive.

(b) carácter escrito. El material oral tiene la gran desventaja de que no permite una fácil clasificación en *géneros* (ensayo, narrativa, etc.), como ocurre con los textos escritos[5]. Por ello resulta inusual la utilización de textos orales en los diccionarios de frecuencia[6]. A pesar de que pueda parecer una aclaración innecesaria, dejamos constancia de que en nuestros materiales tampoco se incluyen escritos no impresos (como cartas, redacciones escolares, etc.), que alguna vez, sobre todo por la finalidad docente que perseguían los trabajos, han sido incorporados en ciertos diccionarios de frecuencia.

(c) carácter homogéneo, que implica la eliminación de todo el material que suponga el uso de un lenguaje estereotipado o la aparición de metalenguaje, es decir, lo que no corresponda a la lengua general. Por ello, no entran en los listados publicaciones como devocionarios, libros de poesía, manuales de gramática, glosarios y listas de palabras aisladas, literatura infantil, etc. Las mismas razones sirven también para descartar los textos de carácter dialectal, por lo que no se incluye toda la literatura costumbrista que, pretendidamente, *refleja* los rasgos regionales.

(d) carácter exhaustivo. La calidad de los escritos no constituye un criterio para la selección de los títulos publicados entre 1970 y 1994; por el contrario, todos los títulos de un género determinado que han visto la luz en los años indicados son recogidos en el listado básico y, por consiguiente, tienen las mismas posibilidades de ser elegidos.

(e) clasificación. El conjunto general de obras se divide en los cinco mundos que han venido considerándose en los diccionarios que utilizan las fórmulas de dispersión de Juilland: drama, narrativa, ensayo, literatura técnica y periodismo. Para cada mundo se recopilará un listado constituido por 100.000 palabras, con lo cual nos encontraremos ante cinco *corpora* distintos. Trabajaremos, por lo tanto, con una muestra total de 500.000 palabras (como hicieron Juilland y Morales): las características metodológicas que hemos venido exponiendo permiten -según el criterio de Juilland (1964: 22)- no tener

[5] *Vid.* las clasificaciones que se hicieron en el *Vocabulaire du français fondamental* como prueba de la dificultad que entraña la división de los mensajes orales según géneros, a causa de la superposición de temas en la misma conversación.

[6] En este sentido es excepcional el DEM (*Diccionario del español de México*), donde se tienen en cuenta materiales orales, incluso de carácter dialectal.

que recurrir a listados superiores, hasta de varios millones de palabras, como contemplaron algunos recuentos antiguos.

El listado de base con las obras publicadas en España en el periodo comprendido entre 1970 y 1994 ha sido extraído (o se puede extraer) de la *Bibliografía Española*[7]. Para el caso de Canarias, los P.I.C., centros de información dependientes del Ministerio de Cultura, nos aportan una relación completa de lo editado en las Islas, relación que hemos confrontado con el *Catálogo de ediciones canarias*, publicado por la Viceconsejería de Cultura y Deportes del Gobierno Autónomo, que recoge las ediciones isleñas desde los años cincuenta hasta 1988. Lógicamente, de estos listados deben desaparecer todas aquellas obras escritas por autores que no sean españoles (en el *LEBAES*) o, en el caso del *LEBAICan*, canarios (para el léxico regional tampoco se tienen en cuenta las publicaciones de aquellos canarios que han vivido, o viven, permanentemente fuera del territorio insular y cuyas publicaciones han visto la luz regularmente en editoriales extrainsulares). De las consideraciones precedentes se deduce que también las traducciones han de quedar fuera de los listados básicos.

De los listados, ya *depurados* de acuerdo con lo dicho en el párrafo anterior, se elegirá, completamente al azar el 15% del total de obras publicadas.

Partiendo de un cálculo que nos proporcione el número total de páginas de las obras seleccionadas en el sorteo, por cada mundo, la media de palabras por oración y de oraciones por página en cada género, sabremos el número de oraciones que debe aportarnos cada texto para alcanzar las 100.000 palabras de cada mundo. Tras ello, procederemos a la selección al azar de las páginas y de las oraciones concretas que constituirán las unidades de análisis.

El mundo periodístico, por sus características específicas, requiere un tratamiento peculiar, del que ahora no podemos ocuparnos.

4. Antes de pasar a exponer los problemas relativos a la presentación de los materiales, conviene hacer referencia a las fórmulas mediante las que se seleccionarán alrededor de 5.000 palabras básicas en ambas investigaciones.

---

[7] Desde el año 1958 hasta 1970, la *Bibliografía Española* ha sido publicada por el Servicio Nacional de Información Bibliográfica, Madrid, Dirección General de Archivos y Bibliotecas. A partir de 1971 y hasta 1975 el responsable de la elaboración es el Instituto Bibliográfico. Finalmente, el periodo comprendido entre 1976 y 1994 ha sido recogido en CDROM por la Biblioteca Nacional.

Ya están lejos los tiempos en que la medida de selección de palabras se basaba única y exclusivamente en la frecuencia. Los diccionarios de Juilland emplean índices de *dispersión compleja* y *de uso*. En nuestro caso, utilizaremos la misma fórmula de dispersión que empleó Amparo Morales 1986, elaborada sobre la propuesta antes por Juilland 1973, con significativos refinamientos:

$$D = 1 - \frac{\sqrt{n\, x_i^2 - T^2}}{2T}$$

donde $T^2 = (x_i)^2$ es la suma elevada al cuadrado de la frecuencia total de cada palabra y nx^2 es el producto del número de mundos o categorías por la suma de los cuadrados reales de las frecuencias de la palabra en cada mundo.

Por otro lado, el índice de uso, de acuerdo con su fórmula ($U = F\,x\,D$), nos va a permitir ponderar la frecuencia de cada palabra por su dispersión y, consecuentemente, nos señalará cuáles son las palabras que deben aparecer en los listados finales.

5. La elaboración de estos diccionarios supone dar respuesta previa a una serie de problemas lexicométricos, relacionados con la segmentación de la unidad de análisis, para los que hay propuestas divergentes. Esta es la parte del trabajo en que es más importante y decisiva la intervención de los investigadores. En nuestro caso, contamos con la ventaja de que partimos de los mismos presupuestos teóricos y los mismos principios metodológicos (muy cercanos a los de Morales 1986) porque han demostrado su eficacia de manera incuestionable. Independientemente de las soluciones que tomemos ante determinadas situaciones concretas y que expondremos en las líneas siguientes, hay dos postulados básicos:

- la unidad de segmentación es la *palabra*. En esto coincide la gran mayoría de diccionarios, en los que ha pesado, sin duda, la facilidad de identificación de esa unidad.

- la *lematización*. Hoy no se concibe un diccionario básico que no incorpore listas de palabras lematizadas, puesto que, como ha señalado Juilland (1964: 24-25), el propósito de estos trabajos no es tanto el de aportar listas jerarquizadas de palabras como el de describir la estructura de la lengua. Pero la lematización nos lleva a enfrentarnos a un conjunto de problemas prácticos, sin resolver o resueltos de forma dispar en trabajos previos:

(a) Polisemia y homografía. Nos encontramos aquí sin pautas teóricas que permitan establecer una *jerarquía de relevancia* de los rasgos semánticos con un alcance general; esta situación hace muy difícil la decisión acerca de

qué rasgos son realmente significativos y deben suponer entradas distintas en el diccionario. Por eso no debe extrañar que en el *FDSW* los autores, como indica Morales (1986: 19), hayan prestado más atención a la distinción gramatical que a la de tipo semántico. La falta de pautas generales nos obliga, pues, a aquilatar las diferencias significativas con el fin de adoptar siempre soluciones coherentes, sin olvidar, claro está, la ayuda que supone el conocimiento de las distinciones establecidas en el *FDSW* y en el *LEPR*.

(b) Asignación de palabras a clases. Se trata de un aspecto de suma importancia, a pesar de la dificultad de alcanzar soluciones plenamente satisfactorias. El objetivo fundamental es el de establecer criterios uniformes que permitan dar las mismas soluciones a problemas idénticos. Son dos los niveles que debemos diferenciar:

(b_1) Desambiguación de la homonimia sintáctica, mediante la distinción, como entradas separadas, de los homógrafos que representan clases gramaticales diferentes (por ejemplo, pronombres y adjetivos demostrativos). Si bien es cierto que hay formas claramente diferenciadas en la teoría gramatical, se presentan muchos problemas con aquellas que no tienen un deslinde teórico tan definido o que ofrecen dificultades para identificar sus valores en el texto (valga como ejemplo relevante la forma *se*). En estos últimos casos, se hace preciso un acuerdo previo de los investigadores de ambos diccionarios para precisar los límites de la especificación de clases gramaticales.

(b_2) Desambiguación de la homonimia morfológica, aportando los totales de cada categoría dentro de la entrada correspondiente. De nuevo aquí nos encontramos ante soluciones seguidas tradicionalmente (por ejemplo, el que el género suponga entradas diferentes en los sustantivos, pero no en los adjetivos) y ante otras que han recibido distinto tratamiento en los diferentes diccionarios (como ocurre con la especificación del género de las formas ambiguas *mi*, *tu*, *su*, *le*, *se*). No hace falta insistir en la necesidad de marcar unas pautas conjuntas para dar la misma solución a los ejemplos que presentan problemas.

(c) División y agrupación de cadenas gráficas. Tampoco estos problemas, que Charles Müller 1963 sitúa en el *nivel de texto*, han recibido respuestas homogéneas, si bien hay dos casos de segmentación que destacan por lo contrario: nos referimos a las amalgamas de preposición y artículo (*al*, *del*) y a las de verbos y enclíticos; en ambos casos se prefiere casi unánimemente el recuento separado. Más dificultades presentan algunas agrupaciones -como las perífrasis verbales o las locuciones de carácter preposicional, adverbial y conjuntivo-, en las que resulta muy delicado precisar el momento en que se lexicalizan completamente. Dado el estado de

los estudios gramaticales, consideramos que la mejor solución es la que consiste en prescindir, como hizo Amparo Morales en el *LEPR*, de la *unidad de función* en aras de la coherencia metodológica.

CUADRO 1

hallazgo *n.*		5.68	9	0.63	
hallazgo	1	2	4	0	2
hambre *n.*	22.72	31		0.73	
	8	11	3	2	7
hambre	8	11	8	2	6
hambres					1
harto *av.*		5.00	7	0.71	
harto	1	2	2	2	0
hasta *p.*		657.03	726	0.90	
hasta	177	176	135	132	106
hato *n.*		7.20	18	0.40	
	0	2	12	3	1
hato		1	12		1
hatos		1		3	
hazaña n.	4.44	7		0.63	
	1	3	0	1	2
hazaña	1	2		1	1
hazañas		1			1

6. Una vez que el ordenador haya terminado con todos los recuentos, los resultados se presentarán en forma de diccionario, según la siguiente estructura:

En primer lugar aparecerá el lema (en negrita), con su marca de clase de palabra, y a renglón seguido tres cifras, las correspondientes a 1) índice de uso, 2) frecuencia absoluta y 3) dispersión. Inmediatamente después se podrán leer las frecuencias absolutas del lema en cuestión en cada uno de los cinco mundos estudiados. Continuará la columna del lema con las apariciones léxicas específicas, seguidas de sus frecuencias absolutas particulares por mundo [CUADRO 1].

7. El panorama que hemos descrito en las líneas precedentes nos sitúa ante dificultades, sobre todo en lo que refiere al segundo nivel de análisis, a las que hemos debido enfrentarnos desde los momentos iniciales de este trabajo. Sin embargo, la esperanza de que nuestro esfuerzo resulte útil en ese camino que conduce al mejor conocimiento de la estructura de nuestra lengua nos ha animado a emprender esta tarea con los mejores ánimos. Los primeros pasos están dados.

BIBLIOGRAFÍA

JUILLAND, Alphonse y Eugenio CHANG RODRÍGUEZ, 1964, *Frequency Dictionary of Spanish Words*, The Hague, Mouton.

JUILLAND, Alphonse, P.M.A. EDWARDS e I. JUILLAND, 1965, *Frequency Dictionary of Rumanian Words*, The Hague, Mouton

JUILLAND, Alphonse, D. BRODIN y C. DAVIDOVITCH, 1971, *Frequency Dictionary of French Words*, The Hague, Mouton.

JUILLAND, Alphonse y V. TRAVERSA, 1973, *Frequency Dictionary of Italian Words*, The Hague, Mouton.

LÓPEZ MORALES, Humberto, 1986, *Enseñanza de la lengua materna. Lingüística para maestros de español*, Madrid, Playor.

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

MÜLLER, Charles, 1963, "Le mot unité de texte et unité de lexique", *Travaux de Linguistique et Littérature*, 1, pp. 165-175.

ROMERA CASTILLO, José, 1991, "Hacia una bibliografía sobre didáctica del léxico", *Lenguaje y Textos*, 1, pp. 43-51.