Estadística y Vocabularios: Perspectivas actuales

Por Julia VALERA

A partir del siglo XIX, y sobre todo el XX, surgen una serie de trabajos que intentan renovar y completar el estudio del vocabulario mediante la aplicación de índices estadísticos. Estas investigaciones responden en su mayoría a preocupaciones pedagógicas relacionadas con el aprendizaje del lenguaje. La lingüística cuantitativa comienza por el registro de letras, morfemas, fonemas, etc., considerados como unidades. Pero será sobre todo a partir de mediados de este siglo cuando los vocabularios cobren nuevo impulso al poner de relieve la teoría de la información el carácter sistemático del lenguaje.

Los vocabularios han pasado a formar parte de la lingüística cuantitativa, y la interpretación de datos obtenidos mediante la técnica del recuento de términos ha permitido establecer leyes y formular hipótesis sobre la estructura de los mismos y del léxico que contienen.

INDICES ESTADISTICOS

Sin entrar en el análisis de los primeros estudios estadísticos realizados en fonología y en morfología (Kading, Zipf, Hennon) nos centramos en los trabajos directamente relacionados con la elaboración de vocabularios con finalidades pedagógicas. El objeto es conocer las dificultades y soluciones encontradas en este terreno concernientes a la utilización de índices estadísticos.

Contamos, en principio, una cierta confusión terminológica, tanto respecto a la definición de vocabulario y léxico, términos indistintamente utilizados, como a las múltiples acepciones de los vocabularios de base —elemental, general, fundamental— y a las distintas significaciones que se le confieren a un mismo índice, entre ellos, al de Valencia. Matoré refiriéndose a estos problemas de terminología considera que pueden ser la resultante de la falta de explicación de presupuestos lingüísticos que a su vez dificultaría una definición precisa de los objetos de estudio (1).

Thorndike construye el «Teacher's Word Book» en 1921. Es el primer vocabulario, destinado a la enseñanza de una lengua, que utiliza el método estadístico para determinar «científicamente» los términos más usuales. Su influjo será enorme. Podemos pensar que lo que le conduce a realizar un trabajo de este tipo es su conexión con la corriente empirista, que le permite considerar cada una de las palabras como estímulo con entidad propia, del mismo modo que sus preocupaciones por los problemas del aprendizaje le llevan a interesarse por el lenguaje en su dimensión pedagógica. Los autores del «Francés Fundamental» opinan que existe un corte entre el vocabulario de Thorn-

(1) G. MATORE: «Histoire des dictionnaires français». Ed. Larousse, París, 1968, pág. 20.

dike y otros vocabularios, tales como el «Basic English» precisamente en función de la utilización de criterios estadísticos. Pero de hecho la concepción lingüística implícita sigue siendo la misma: la concepción taxonómica del lenguaje propia de la época clásica.

Habrá que preguntarse hasta qué punto la ausencia de una elaboración teórica acerca de un fenómeno tan complejo como el lenguaje en la mayor parte de los autores de los vocabularios no está en gran medida determinada por un punto de partida pragmático en el que priman los criterios de eficacia, aprendizaje rápido, etc. Considerar la usualidad del lenguaje como un hecho natural significa ignorar que en cada sociedad se favorecen ciertos tipos de discursos y se excluyen otros. La usualidad es la resultante de una depuración discursiva en la que intervienen instituciones, aparatos de censura, imposición de normas, sistemas de dominación. Como señala Foucault «en toda sociedad la producción del discurso es a la vez controlada, organizada y redistribuida por una serie de procedimientos que tienen por función conjurar los poderes y los peligros, dominar el suceso aleatorio, esquivar la pesada, la temible materialidad» (2).

A partir del «Vocabulario Usual, Común y Fundamental», del Dr. García Hoz y de la elaboración un poco posterior del «Francés Fundamental», la utilización de índices estadísticos se generalizará como uno de los criterios objetivos para determinar los términos más usuales.

El «Vocabulario Usual, Común y Fundamental» (3) es el primer vocabulario de base realizado en España. Es un vocabulario restringido, elaborado con fines psicopedagógicos, y que tiene por objeto determinar cuál es el vocabulario utilizado por el hombre de la calle que sería de algún modo el exigible al nivel de la E.G.B. Se utiliza la palabra como unidad de registro, siendo consciente su autor de las limitaciones que esto presenta y explicitando que no se trata de un estudio semántico, sino de constatación de formas comúnmente utilizadas. Es de destacar además la variedad de las fuentes empleadas, así como su representatividad. Se señalan tres estratos diferentes, poniéndose así de relieve el distinto valor de uso del vocabulario obtenido.

El «Francés Fundamental» (4) presenta la novedad de registrar términos de la lengua hablada y de acotar la noción de disponibilidad. Una vez realizada la primera fase del trabajo sus autores se dan cuenta de la insuficiencia del criterio de frecuencia para estimar la probabilidad de uso de un término y tendrán en cuenta su distribución en los distintos textos en que aparece. Surge así el «índice de repartición» que había sido denominado por Vander Beke «rango» (5). Utilizan los índices de frecuencia y de repartición sin operar una fusión de ambos y dando prioridad al de frecuencia que fija el umbral de aceptación de un término determinando además su clasificación. Obtienen una lista de palabras que no incluye, o lo hace en muy pequeña medida, términos que se refieren a cosas concretas.

Con el fin de obtener el vocabulario concreto realizan una nueva investigación que determine el grado de «disponibilidad» de un término concreto. Para ello, esquemáticamente, parten de 16 centros de interés y constatan las palabras que una muestra de sujetos asocia a cada uno de dichos centros de interés. Totalizando los rangos en las distintas listas obtienen el índice general de disponibilidad de cada uno de los términos.

P. Guiraud plantea dos objecciones al «Francés Fundamental»:

— La comprensión de un texto es imposible a partir de un vocabulario de base, por muy amplio que sea, ya que cuando se habla es en circunstancias concretas y con fines determinados, haciéndose imprescindible el conocimiento de palabras específicas en relación a una situación concreta.

— Existe el problema de fijar el nivel a partir del cual el rango de un término corresponde al rango real. Considerando que los términos de las listas obtenidas se

(2) M. Focault: «L'ordre du discours». Ed. Gallimard, París, 1971, págs. 10-11.
(3) V. GARCIA HOZ: «Vocabulario Usual, Común y Fundamental». C.S.I.C., Madrid, 1952.
(4) G. GOUGENHEIM, R. MICHEA y otros: «L'élaboration du Français Fondamental» (1er. degré). Ed. Didier, París, 1964.
(5) Citado por CH. MULLER: «Fréquence, dispersion et usage à propos des dictionnaires de fréquence». Rev. Cahiers de Lexicologie, núm. 2, 1965, pág. 33.

distribuyen según la fórmula r 1.35 (constante), se puede calcular aproximadamente el rango real de un término. Con 300.000 términos no hay suficiente garantía a partir del término número 500 (6).

Matoré por su parte, refiriéndose también al método estadístico aplicado en el «Francés Fundamental», afirma que las rectificaciones hechas por sus autores a la clasificación por frecuencias son insuficientes, presenta lagunas manifiestas, y figuran por el contrario en él, términos que no deberían figurar. Añade que el error cometido es el haber utilizado el criterio de frecuencias creyendo que es posible registrar los términos de un vocabulario igual que se pueden clasificar objetos materiales (7).

Mackey, en el Centro Internacional de Información e Investigación sobre bilingüismo de la Universidad de Laval, utiliza en la construcción del «Vocabulario disponible del Francés» los índices de frecuencia, repartición y disponibilidad ya citados, introduciendo un nuevo índice: el «índice de la valencia». Entiende por tal el modo de ordenar los términos partiendo de cuatro parámetros: definición, combinación, inclusión y extensión. El problema se le plantea en el momento en que intenta combinar estos cuatro índices. Confiesa que no puede llegar a su utilización conjunta para fijar distintos estratos o niveles de un vocabulario, dado que los índices no tienen todos igual valor y no ha llegado a definir el peso específico de cada uno (8).

La aparición de un nuevo índice —«índice de uso»— definido como la resultante aritmética del índice de frecuencia y del de repartición, tiene lugar en el «Frequency Dictionary of Spanish words», dirigido por M. A. Juilland y por E. Chang Rodríguez en la Universidad de Stanford (9). Uno de los obpjetivos de este trabajo era elaborar un vocabulario de base formado por 5.000 términos. El corpus del que se obtendrían dichos términos estaba formado por 500.000 palabras correspondiente a cinco áreas distintas, 100.000 de cada una de ellas.

La elaboración de los resultados siguió estas etapas:

— Cada término registrado aparece con cinco subfrecuencias (áreas). La suma de las subfrecuencias es igual a F. La Frecuencia media es igual, en este caso, a $\frac{F}{5}$. Se eliminan los términos de frecuencia inferior a 5.

— Interviene entonces la noción de «dispersión», equivalente a la de «repartición», cuya función es clasificar a los términos no eliminados, pero además determinar la inclusión de los términos de frecuencia media igual o superior a 5 que no están o están muy desigualmente repartidos en las cinco áreas.

— Por ser pequeño el número de áreas para calcular el índice de dispersión (D) tienen en cuenta no sólo el número de áreas en que aparece el término, sino también las cinco subfrecuencias comparadas a la frecuencia teórica o media. Calculan la varianza —media de los cuadros de las desviaciones a la frecuencia media— y obtienen el coeficiente de variación (V) que es igual al cociente de la desviación típica y la frecuencia teórica. Dividiendo este coeficiente de variación por VN-1 obtienen un índice que variará de 0 a 1. Invirtiendo el sentido de la variación —toman el complemento de este índice a la unidad— 1 aparece cuando las frecuencias están repartidas por igual en todas las áreas y 0 cuando las frecuencias de un término pertenecen a una sola área. La fórmula del índice de dispersión sería $D = 1 - \sqrt{\frac{V}{N-1}}$, siendo N el número de áreas. Cada término estaría representado por dos índices numéricos: (F) y (D) (frecuencia y dispersión).

— Obtienen un tercer índice: el «índice de uso» U, multiplicando la frecuencia por la dispersión: $U = F \times D$.

(6) P. GUIRAUD: «Problèmes et méthodes de la statisque lingüistique». Ed. PUF, París, 1960, pág. 95.
(7) G. MATORE: Op. c., pág. 225.
(8) W. F. MACKEY: «Le Vocabulaire disponible du français». Ed. Didier, París, 1971, pág. 26.
(9) A. JUILLAND y E. CHANG RODRIGUEZ: «Frequency Dictionary of Spanish Words». Ed. Mouton, París, 1961.

Según Muller este índice de uso sigue planteando dificultades desde el punto de vista estadístico, si bien las reduce al mínimo, ya que continúa existiendo una cierta arbitrariedad en este procedimiento de selección y de clasificación de términos. Existe una cierta desigualdad en favor de los términos cuya frecuencia es 5 o múltiplo de cinco: así un término de frecuencia igual a 10 puede teóricamente tener un índice de uso igual a 10; pero otro de frecuencia igual a 9 no puede sobrepasar un índice de uso igual a 8, y lo mismo sucede con un término de frecuencia igual a 11 que tampoco puede superar un índice de uso igual a 10. Se plantea así un sesgo que actúa indudablemente en la clasificación y que en último término remite a cuestiones metodológicas (10).

Phal en su «Vocabulario General de Orientación Científica» (11) utiliza nociones nuevas: «extensión de empleo» y «combinabilidad». La extensión de empleo se refiere a la vez al número de apariciones de un término (frecuencia) y al número de contextos diferentes en los que dicho término aparece (repartición). Phal dice que lo ideal sería combinar frecuencia y repartición mediante un procedimiento estadístico aceptable tal como lo hicieron Juilland y Chang Rodríguez, pero dado que las áreas de que él parte no contienen igual número de términos, igualar el corpus supondría realizar reajustes laboriosos y muy complejos (12).

En consecuencia, este autor procede estableciendo una relación simple entre frecuencia y repartición ($F \times R$) para obtener el «número de orden» de cada término. Añade que en consecuencia esta relación no tiene valor estadístico real y que la utilización del número de orden hay que hacerla con reservas.

Por lo que respecta a la combinabilidad dice que es necesario tener en cuenta el contexto, es decir, constatar la asociación de los términos con los inmediatamente vecinos.

Extensión de empleo y combinabilidad equivaldría para Phal a la «valencia» de un término.

Vemos pues que tanto el índice de frecuencia como el de uso plantean problemas estadísticos que no están todavía resueltos. Además de estas críticas concretas los lingüistas plantean una serie de objeciones epistemológicas al empleo de la cuantificación referida al lenguaje.

M. Coyaud dice que el formalismo y la cuantificación en lingüística pueden ser objeto de las mismas objeciones que Köhler hace al behaviorismo en su «Psicología de la forma», que pueden resumirse del modo siguiente: el prestar más atención a los problemas de la medida que a los interrogantes que plantea el fenómeno a estudiar da lugar a la utilización de métodos matemáticos empleados en las ciencias físicas, sin preguntarse previamente si una transposición de campos es posible y bajo qué condiciones» (13).

Ch. Muller, refiriéndose también a la estadística aplicada al lenguaje reafirma esta opinión al decir que los resultados no son satisfactorios desde el punto de vista científico, dado que en alguna medida están fundamentalmente hasta el momento en criterios empíricos (14). Wagner y Matoré afirman que la frecuencia de uso de un término no tiene significación en sí misma y que debe ser interpretada en función de la naturaleza del signo y de los contextos (15). Es decir, que los índices estadísticos no fundamentan en último término más que un estudio descriptivo del lenguaje, no pudiendo dar explicación de su funcionamiento.

La cuestión de la utilización de criterios de cuantificación en la elaboración de los vocabularios, pensamos que no es un problema cerrado, sino que se inscribe en una

(10) CH. MULLER: Op. c., pág. 41.
(11) A. PHAL: «Vocabulaire Général d'Orientation Scientifique». CREDIF, París, 1971.
(12) A. PHAL: «La rechesche en lexicologie au CREDIF. La part de lexique commun dans les vocabulaires scientifiques et techniques». Rev. Langue Française, núm. 2, mayo 1969.
(13) M. COYAUD: «Transformations lingüistiques et classifications lexicales». Rev. Cachiers de Lexicologie. número 2, 1965, pág. 27.
(14) CH. MULLER: «Initiation aux méthodes de la statistique lingüistique». Ed. Hachette, París, 1972, segunda ed., pág. 9.
(15) R. L. WAGNER: «Les Vocabulaires françaises». Ed. Didier, París, 1976, pág. 159.
G. MATORE: «La méthode en lexicologie». Ed. Didier, París, 1953, pág. 62.

cuestión más amplia y más compleja: la utilización de los modelos matemáticos en las ciencias humanas. Es probable que fueran los psicólogos experimentales los que al interesarse por el aprendizaje de la lengua hayan considerado necesario establecer vocabularios siguiendo sus criterios de cientificidad. De ahí la relación que en la actualidad tienen los vocabularios con la psicología del aprendizaje.

PERPECTIVAS ACTUALES

Algunos de los autores ya citados —Gougenheim, Quemada, Phal— después de enfrentarse con su propia práctica en el campo de los vocabularios y con las críticas, provenientes fundamentalmente de los lingüistas, adoptan una nueva perspectiva y realizan actualmente nuevos trabajos, que si bien siguen siendo de lingüística aplicada y con carácter eminentemente pedagógico, intentan tener en cuenta las adquisiciones recientes de la lingüística al mismo tiempo que reexaminan las dificultades de la aplicación de los procedimientos estadísticos.

Exponemos a continuación los estudios actualmente en marcha que consideramos más relevantes:

a) «Tresor de la langue française»

Gougenheim, Wagner, Dubois, Matoré y un equipo de colaboradores están realizando desde 1964 en la Universidad Nancy y en colaboración con el CNRS (Centro Nacional de Investigación Científica) el «Tresor de la langue française» utilizando ordenadores en la automatización de datos. Puede servirnos como punto de referencia de su planteamiento teórico lo que escribe Wagner (16): Los términos funcionan en segmentos de enunciados y en frases que determinan sintácticamente las condiciones de sus usos... No deben pues extraerse artificialmente correlaciones. Continúa diciendo que han rehusado utilizar el procedimiento lento, costoso y de una eficacia dudosa empleado en la elaboración del Francés Fundamental. Plantea la necesidad de llegar a la construcción de un sistema morfo-sintáctico en el que cada signo esté representado diacrónica y sincrónicamente.

Como primer paso reúnen un fondo completo de obra de lexicografía y lexicología. Pasan luego a la selección de los textos que faciliten una información exhaustiva sobre las diferentes épocas del vocabulario francés.

Como resultado de esta investigación piensan obtener en lo que a cada término se refiere: bibliografía, descripción fonética y morfológica, descripción ortográfica y gramatical, etimología, contenido semántico, frecuencia de empleo, lista de expresiones o frases en las que aparece, interrelaciones semánticas y lexicológicas.

Una de las múltiples posibilidades que el material recogido ofrece según sus autores, es el poder determinar los grupos binarios —asociaciones secuenciales de dos términos semánticos que se presentan juntos con cierta frecuencia. El problema está en determinar cuándo estas asociaciones no se deben al azar, es decir, determinar un umbral que permita seleccionar los grupos binarios. A partir de un corpus de seis o siete millones de términos determinarán estadísticamente los grupos binarios cuya asociación se explica por motivaciones semánticas. Estos grupos les permitirán situar los términos en el interior de estructuras sintagmáticas en las que realmente funcionan con frecuencia.

El primer volumen del Tresor se publica en 1972 (17). Está compuesto por numerosos textos de los siglos XIX y XX, todos ellos literarios. Contiene 75.415 términos, correspondientes a 71 millones de frecuencias.

b) Centro de francés moderno y contemporáneo de Besançon

Este centro, órgano de CNRS, está dirigido por Quemada, existiendo en la Universidad de Montreal y de la Sarre observatorios asociados funcionalmente a él.

(16) R. L. WAGNER: Op. c., T. II, pág. 91.
(17) Varios: «Dictionnaire alphabetique des fréquences». Centre des Recherches pour un Trésor de la Langue Française. Ed. Didier, París, 1972.

Según Quemada, este trabajo —poner al día un inventario del francés contemporáneo— trata de estudiar no la distribución de los términos, sino la distribución de las variantes léxico-semánticas de los mismos. Esta distribución, que tiene en cuenta la significación, permitirá obtener:

— El modelo estructural en el que el término se emplea.

— La fórmula generalizada de la aptitud combinatoria de un término con un significado preciso.

— Definir campos semánticos mediante la reagrupación de unidades que presentan la misma significación.

Constatamos pues que este autor al igual que los anteriores intenta realizar un estudio de la significación de los términos partiendo de sus relaciones sintagmáticas (18).

c) CREDIF

A. Phal dentro del marco institucional del CREDIF y después de elaborar el VGOS plantea la necesidad de un estudio morfo-sintáctico del vocabulario que tenga como base la frase (19). Comienza un nuevo trabajo con Descamps sobre la Geología, con el fin de analizar el funcionamiento de los términos en contexto.

Una primer etapa del trabajo consiste en reagrupar en sintagmas todos los empleos de un término dado, lo que le permitirá descubrir una serie de microsistemas morfo-sintácticos más o menos relacionados entre sí y que funcionan en el interior del discuros geológico. Esta investigación deberá además conducir a establecer leyes que podrán aplicarse no sólo a la geología sino también a la sintaxis de la frase científica en general. Sus autores aducen para ello dos razones: un corpus amplio y bien seleccionado puede reflejar distintos registros de la prosa científica, además la geología se sitúa en la confluencia de distintas ciencias y técnicas relacionadas entre sí.

En consecuencia deberán aparecer después del registro de las palabras en contexto los distintos tipos de frases más aptos para la expresión científica. A partir de aquí y reagrupando los tipos privilegiados de frases se podrá constatar la predominancia de empleo de una serie de ellas —frases modelo— en la prosa científica.

Además de la comprobación de estas hipótesis teóricas estos autores explicitan algunas de las aplicaciones prácticas que se derivarán de su investigación: estudio de derivaciones, paso del término a la frase, presentación de ejercicios de fijación que permitirán aprender los términos y sus significados tal como aparecen en las frases o sigtagmas que realmente usa el discurso geológico.

Estas investigaciones recientes parecen pues utilizar como vía de superación de las limitaciones de los vocabularios basados en el registro de la palabra, el tener en cuenta la significación mediante el establecimiento de familias parafrásticas. Esto no significa sin embargo que las dificultades que se presentan en la elaboración de los vocabularios queden totalmente resueltas. A medida que avanzamos en la exposición encontramos una complejidad creciente en los planteamientos: tener en cuenta la significación supone hacer referencia a formaciones dircursivas que tienen solamente una autonomía relativa reenviando a procesos de carácter social. Los vocabularios y las cuestiones que plantean aparecen así estrechamente vinculadas no sólo a determinaciones lingüísticas y psicopedagógicas, sino también a una política de la lengua determinada.

---

(18) Citado por D. SLAKTA: «Les problèmes du lexique à la lumière des thèses et des tavaux récentes». Rev. Langue Française, núm. 2, mayo 1969, pág. 94 y ss.

(19) A. PHAL: «Analyse lingüistique: de la langue quotidienne à la langue des sciences et des techniques». Rev La Français, dans le momde, núm. 61, diciembre 1968, págs. 7-11.