

## DE LA FRECUENCIA A LA DISPONIBILIDAD LÉXICA

El hecho de que las lenguas estén integradas por elementos discretos permite diversos ejercicios de identificación y segmentación mediante la aplicación de determinados criterios de clasificación. Estas operaciones facilitan los procesos para cuantificar las células que las componen, fijar su organización y establecer su estructura jerárquica. Resultan, por tanto, esenciales en la búsqueda de lo característico y diferenciador de cada una de ellas, en la comprensión de los principios o leyes sobre los que asientan sus mecanismos de comunicación y en la explicación de sus modelos relacionales.

La necesidad de alcanzar conclusiones más precisas, objetivas y científicamente irrefutables, abre la puerta de la lingüística al tratamiento de los números que resultan de ese tipo de análisis mediante la aplicación de métodos estadísticos adecuados.

El más antiguo indicador utilizado en el campo léxico, como se ha relatado en las páginas anteriores, se fundamenta en el recuento de las *frecuencias absolutas y relativas* de empleo de unidades léxicas. Nacen de aquí los primeros listados de palabras ordenadas en razón de la *frecuencia* de aparición en textos y conversaciones que serán el soporte de los *diccionarios de vocabulario básico*.

López Morales [LÓPEZ MORALES. 1986] llega a diferenciar cuatro tipos distintos de diccionarios<sup>1</sup>. Habla de *diccionarios normativos* que catalogan un gran número de términos e incluyen arcaísmos, regionalismos, tecnicismos y cultismos. Se constituyen en el inventario del legado léxico de una comunidad de habla. Junto a ellos distingue

---

<sup>1</sup> Este paréntesis queda justificado porque cada una de las categorías reseñadas vienen a propiciar el desarrollo de una vía, o etapa, hasta desembocar en la *disponibilidad léxica*.

las *listas de frecuencia* que recogen sólo los vocablos vivos, los más comunes en el seno de la colectividad, aquellos que son generalmente utilizados y mayoritariamente comprendidos. Estos vocablos se ordenan estableciendo una jerarquía de acuerdo a la cantidad de veces que se repiten en un texto. Ponen de manifiesto la limitación de los hablantes a la hora de manejar el caudal léxico de su propia lengua. El grupo tercero lo constituyen los diccionarios de *léxico básico* que resultan de recuentos selectivos. En este caso se contabilizan las palabras en *mundos* diferentes delimitados en función de la clase de texto generado o de las condiciones en que se produce. Se matiza aquí la *frecuencia* con factores nuevos como la *dispersión* o el *uso*. El último modelo que enumera López Morales, la *disponibilidad léxica*, nace con la intención de organizar las palabras atendiendo a su facilidad para aparecer ligadas a temas de comunicación y situaciones particulares muy concretas. La *frecuencia*, en ellos, deja de ser, al menos en teoría, el criterio dominante para la ordenación.

Al mismo tiempo que la lingüística se interesa por el recuento de *las frecuencias*, la psicología descubre que la *frecuencia de uso* está en íntima relación con fenómenos de orden psicológico como la rapidez y exactitud en la *percepción* de las palabras y el *tiempo de reacción* verbal. Miller [MILLER. 1956:273] establece experimentalmente una correlación de 0.70 entre *frecuencia de uso* y *percepción* y destaca lo sorprendente que resulta una “correlación tan íntima cuando se observa que las dos evaluaciones - *frecuencia* y *umbral* - provienen de fuentes distintas”.

El *umbral de percepción* es una medida que conecta *sensación* y *percepción*. La *sensación* se refiere a las experiencias inmediatas básicas generadas por los estímulos simples [MARTÍN y FOLEY. 1996]. Es la respuesta de los órganos de los sentidos frente a un estímulo [FELDMAN, R. 1999]. La *percepción* se encarga de descifrar esas sensaciones dándoles significado y organización [FELDMAN, R. 1999]. Implica, en consecuencia, la actividad del cerebro. Se denomina *umbral de percepción* a la cantidad mínima de estímulo necesario para que sea detectado por los sentidos y enviado al cerebro para su interpretación y respuesta. Por ejemplo, el *umbral de percepción* de un sonido está establecido, para el individuo medio y en condiciones de silencio, en la detección del *tic-tac* de un reloj de pulso a siete metros de distancia.

El concepto de *tiempo de reacción* se considera, en Psicología, acuñado por Gustav Theodor Fechner, también conocido bajo el seudónimo de Dr. Mises [FECHNER, G. 1860; edición inglesa de HOWES, D. 1961. Traducción de ADLER, H.]. Fechner es el autor de la teoría de la *inferencia inconsciente de la percepción* que estima que las *sensaciones*

no permiten el acceso directo a fenómenos ni a objetos, sino que sirven a la mente como *señales de realidad*.

La *percepción* requiere un proceso lógico y activo por parte del perceptor que utiliza la información suministrada por la *sensación* para inferir las propiedades de los fenómenos u objetos y responder en consecuencia. Formula al respecto, en lo que es su principio más famoso, que la intensidad de una sensación se incrementa a lo largo del estímulo ( $S = k \log R$ ) para caracterizar las relaciones psicofísicas externas.

Al hacer esto, creyó haber alcanzado el camino para demostrar una verdad filosófica fundamental: que la mente y la materia son simples modos diferentes de concebir una misma realidad. Su pensamiento es adoptado por la Psicología Cognitiva que lo incorpora a sus experimentos.

Franciscus Cornelis Donders hace pública en 1865 una comunicación preliminar sobre el *tiempo de reacción* en la que informa del trabajo que realiza con un estudiante, Johan Jacob de Jaager, que concretará su disertación doctoral en ese mismo año. Poco después, 1868, [DONDERS, F. 1868; en inglés 1969] defiende su generalización como procedimiento experimental y lo aplica en el estudio del tiempo invertido por las operaciones mentales. Concluye que el *tiempo de reacción* es aditivo y evalúa por separado los tiempos necesarios para responder a estímulos bajo *condiciones de decisión* y de *simple no decisión*. Hoy son muchas las ramas de la ciencia que utilizan esta medida, aunque sin alcanzar la importancia que tuvo en los laboratorios psicológicos de la época.

El cálculo de la *frecuencia de uso* de la palabra dentro de un idioma pasa, de esta manera, a ser considerado por lingüistas y psicólogos como un dato de primer orden en sus estudios y aplicaciones.

Se suele iniciar la relación histórica de esta clase de trabajo con J. W. Kärding quien, en 1897, [KÄRDING. 1897] establece las palabras más usuales de la lengua alemana. Siguen Thorndike [THORNDIKE. 1921], que en 1921 hará algo parecido en lengua inglesa, Henmon, en 1924 [HENMON. 1924], en Estados Unidos, Buchanan en 1927 [BUCHANAN. 1941], Vander Beke [VANDER BEKE. 1935]...<sup>2</sup>; pero, sin duda, la referencia mayoritariamente aceptada como esencial para el desarrollo de esta rama de

---

<sup>2</sup> En este apartado sólo se hace mención de los aspectos metodológicos más destacables de unos pocos trabajos que, por su trascendencia, han marcado un punto de inflexión en esta línea de investigación y, por ello, son lugar común, referencia y fundamento de la mayoría de los estudios actuales. El relato completo de su devenir histórico está muy bien tratado y comentado en María Victoria Mateo [1998:31-47]. Es una buena fuente para llenar lagunas y omisiones.

la léxico-estadística es la investigación, mediado el siglo XX, de Gougenheim y sus colaboradores en Francia. Ellos publican, en 1956, unos listados de términos que resultan de valorar, mediante la *frecuencia*, un extenso *corpus* de conversaciones obtenidas por registro directo.

Después de la Segunda Guerra Mundial el inglés empieza a imponerse como el idioma de las comunicaciones internacionales y el francés se siente amenazado. Francia necesita reforzar su posición en las colonias para restaurar su prestigio exterior, lo que hace cuestión de estado la enseñanza del francés como lengua extranjera. El Ministerio Nacional de Educación establece una comisión encargada de fijar los contenidos del “francés elemental”, rebautizado posteriormente “francés fundamental”. Gougenheim, el pedagogo Rivenc y Michéa asumen la máxima responsabilidad en la empresa de encontrar métodos que faciliten el aprendizaje de esta lengua y favorezcan su difusión. El diccionario final del *francés elemental/fundamental* recopila las palabras más frecuentemente usadas y su conocimiento pasa a ser la base para la adquisición progresiva y racional del idioma.

Este enfoque es seductor en su principio, pero pronto aparecen voces críticas. Ya había advertido Michéa [MICHÉA. 1950:188-189] sobre la existencia de un vocabulario *temático* y otro *atemático*.

La lectura atenta de cualquier serie léxica basada en el recuento exclusivo de *frecuencias* evidencia que, al margen de las palabras llamadas *gramaticales* (artículos, preposiciones, pronombres...), las más habituales en cualquier discurso y ocasión, existen otras cuyo uso es más o menos relevante según el contexto, tema o materia de la conversación. Suelen ser vocablos que nombran objetos y actos inmediatos o muy cercanos al binomio hablante-oyente.

”Los lingüistas admiten que el cálculo de la *frecuencia de uso* está realmente fundamentado cuando se enumeran las palabras-herramienta, los verbos, los nombres susceptibles de aparecer en un texto o una conversación cualesquiera”. [FRAISSE, P. 1977a:187].

El mismo Gougenheim [GOUGENHEIM. 1956] reconoce que el sujeto hablante dispone de dos vocabularios: el vocabulario de la *frecuencia*, que le suministra el marco de su discurso, y un inmenso vocabulario de *disponibilidad* en el que las palabras concretas se organizan con relación a sus necesidades.

Surgen intentos de encuadrar las *frecuencias* en estadios delimitados. Buchanan

[BUCHANAN. 1941], por ejemplo, discrimina según siete áreas temáticas: obras dramáticas, novela, poesía, folklore, prosa, textos técnicos y textos periodísticos. Matiza la *frecuencia* con el concepto, ya explicado, de *rango*. Vander Beke [VANDER BEKE. 1935] estima especialmente la aparición de una palabra en varios textos,...

Pero será el hecho de avanzar en el conocimiento de la metodología estadística, que revela vicios y omisiones en la recogida de muestras, junto a un mayor grado de exigencia y rigor en la formulación de hipótesis, lo que aconseje la búsqueda de otros indicadores y medidas capaces de completar los *recuentos de frecuencias*. Afirma, en este sentido, Juilland [JUILLAND, A. 1964:V]:

“Originally we planned to add depth to our investigations by relying on data provided by available frequency dictionaries and graded word lists, those of Henmon, vander Beke, and Gougenheim for French; of Brown, Carr, and Shane for Portuguese; of Macrea for Rumanian; and of Buchanan, Rodríguez Bou, and García Hoz for Spanish. Unfortunately, we soon realized that these studies, mainly because of their limited pedagogical objective, were not suited to our purpose. Even frequency dictionaries and graded word lists compiled with discrimination and care were of little help, because differences in the sampling techniques, in the scanning procedures, in the computing methods, and in the weighting formulas disqualified the results as a sound basis for meaningful comparison and generalization”.

“Planeamos agregar profundidad a nuestras investigaciones confiando en los datos proporcionados por los diccionarios disponibles de la frecuencia y calificamos originalmente las listas de la palabra, las de Henmon, Vander Beke y Gougenheim para el francés, de Brown, de Carr, y de Shane para el portugués, de Macrea para el rumano y de Buchanan, de Rodríguez Bou, y García Hoz para el español. Desafortunadamente, pronto descubrimos que estos estudios, principalmente debido a su objetivo pedagógico limitado, no servían a nuestro propósito. Incluso los diccionarios de la frecuencia y las listas clasificadas de palabras compiladas con discriminación y cuidado eran de poca ayuda, porque las diferencias en las técnicas de muestreo, en los procedimientos de la exploración, en los métodos que computaban, y en las fórmulas que aplicaban descalificaron los resultados como base adecuada para la comparación y la generalización significativas”.

La constatación de esta realidad abre otras líneas de investigación que se concretan en nuevos postulados. Se llega así al concepto de *dispersión* para corregir posibles desviaciones debidas al azar u otras circunstancias con incidencia en la muestra

analizada. Es el equipo de Juilland el que dará el giro definitivo a los estudios frecuentistas. Y lo hacen en un proyecto que él mismo explica [JUILLAND, A. 1964:V] y justifica así:

“This volume inaugurates our Collection THE ROMANCE LANGUAGES AND THEIR STRUCTURES, devoted to the publication of the results of nearly a decade of investigations into the structure of the Romance languages undertaken with the help of electronic computers at the University of Pennsylvania from 1956 to 1961, and at Stanford University since 1961.

The general purpose of these investigations was to prepare structural frameworks, formal procedures, and programming routines designed to facilitate the use of computers in the descriptive, comparative, and historical study of the various aspects of natural languages; the more specific purpose is to use these frameworks, procedures, and routines in an exhaustive study of the phonological, grammatical, and lexical structuration of the five major Romance languages, French, Italian, Portuguese, Rumanian, and Spanish.

If the study of the paradigmatic relations which hold between invariants or classes of invariants in the System adds a second dimension to the "linear" study of syntagmatic relations that hold between variants in the Text, statistical data add the depth of a third dimension to the universe represented by the textual manifestations of a natural language. Though the importance of quantitative data for gaining a genuine insight into linguistic structuration has never been questioned, the gathering of statistical information by visual-manual techniques was such a time-consuming task in precomputer days, that it usually discouraged this type of inquiry”.

“Este volumen inaugura nuestra colección los IDIOMAS ROMANCE Y SUS ESTRUCTURAS, dedicada a la publicación de los resultados de casi una década de investigaciones en la estructura de los idiomas romance emprendidas con la ayuda de computadores electrónicos en la universidad de Pensilvania a partir de 1956 y hasta 1961, y en la universidad de Stanford desde 1961.

Los fines generales de estas investigaciones eran preparar armazones estructurales, procedimientos formales, y las rutinas de programación diseñadas para facilitar el uso de computadoras en el estudio descriptivo, comparativo, e histórico de los varios aspectos de idiomas naturales; el propósito más específico es utilizar estos armazones, procedimientos, y rutinas en un estudio exhaustivo de la estructuración fonológica, gramatical, y léxica de los cinco idiomas, francés, italiano, portugués, rumano, y español, principales lenguas romance.

Si el estudio de las relaciones paradigmáticas que actúan entre los invariantes o las clases de invariantes en el sistema agrega una segunda

dimensión al estudio 'lineal' de las relaciones sintagmáticas que celebran entre las variantes en el texto, los datos estadísticos agregan la profundidad de una tercera dimensión al universo representado por las manifestaciones textuales de una lengua natural. La importancia de los datos cuantitativos para ganar una penetración genuina en la estructuración lingüística nunca ha sido cuestionada, sin embargo, la reunión de la información estadística por técnicas visual-manual era una tarea tan desperdiciadora de tiempo en los días del pre computador, que generalmente desalentó este tipo de investigación".

Se trabaja, además, en encontrar fórmulas para determinar el *coeficiente de uso* que relacione *frecuencia* y *dispersión*. Se procede, para ello, a estratificar el conjunto léxico en cinco mundos o universos, constituidos por 100 000 palabras, delimitados por el contenido o las condiciones formales de los textos y seleccionados por procedimientos aleatorios. El *índice de uso* de cada término decidirá su inclusión en los diccionarios básicos de la lengua<sup>3</sup>.

"On the basis of sampling techniques established by modern statistics, each language was reduced to a body of about 20 000 to 25 000 sentences totalling about 500 000 words: samples of this size are large enough to be genuinely representative of the language under investigation, and small enough to be economically processable by electronic means. To insure the representative character of a "standard contemporary universe", each was divided into five equal "worlds" of about 100 000 words: 1) plays, consisting exclusively of dialogue and approximating best the spoken language; 2) fiction, consisting of

---

<sup>3</sup> La significación de este trabajo es tan importante para el devenir metodológico posterior de la investigación lingüística que justifica el abuso a la referencia textual que apoya la argumentación en este punto. Pocas veces se han aunado los esfuerzos de tantas personalidades y expertos de campos diversos en aras un objetivo común. Así lo hace constar el propio Juilland, cuando manifiesta:

"Cae de su peso que un programa de estudios de tal alcance no se pudiera haber emprendido sin la ayuda y sin el ánimo de muchos colegas y colaboradores. En la Universidad de Pensilvania yo estoy grandemente endeudado con George O. Seiver, anterior Presidente del Departamento de Idiomas de Lenguas Romance y Literatura; al Profesor Saul Gorn, Director del Centro de Computadores; a mis colegas anteriores Eugenio Chang-Rodríguez, coautor de este Diccionario, a Arnold Reichenberger, a Bodo L. Richter; estoy también agradecido a mis colaborador, Dr. Nicolas Morcovesco, que asumió la tarea de Programador Principal; a las Doctoras Catherine Davidovitch, Dorothy Brodin, y Lilian Szklarczyk, que colaboraron en el proyecto francés; a Sebastián diBlasi, que colaboró en el proyecto italiano; a P. M. H. Edwards que contribuyó a las etapas preliminares del proyecto Rumano; y a William y Elaine Polin. En la Universidad de Stanford, yo deseo reconocer mi deuda con los Decanos Robert Sears y Virgil Whitaker, con los Decanos anteriores Albert Bowker y Philip Rhineland, al Profesor N. Forsythe, Director del Centro de Computación, al Dr. Wade Cole, y al Profesor Herbert Solomon, Cabeza Ejecutiva del Departamento de la Estadística; entre mis colaboradores, yo he recibido ayuda preciosa de Eric Liu, Hiroshi Miyaji, Elisabeth Popov, Randal Whitman, y especialmente de Mrs. Jean Beeson". [JUILLAND, A.1964:VII]

novels and short stories; 3) prose, consisting of essays, memoirs, correspondence, etc.; 4) periodicals, consisting of dailies, weeklies, and monthlies; and 5) technical literature, consisting of writings on medicine, engineering, physics, botany, etc. For each of these genres, words of about 5 000 sentences totalling out 100 000 words were selected by random selection”.

“Tomando como base las técnicas de muestreo establecidas por la estadística moderna, cada lengua fue reducida a un cuerpo entre 20 000 y 25 000 oraciones que sumaban cerca de 500 000 palabras: las muestras de este tamaño son lo suficientemente grandes para ser genuino representante de la lengua bajo investigación y bastante pequeñas para ser económicamente procesables por medios electrónicos. Para asegurar el carácter representativo de un ‘universo’ contemporáneo estándar cada uno fue dividido en cinco ‘mundos’ iguales de cerca de 100.000 palabras cada uno: 1) juegos, consistiendo exclusivamente en diálogos lo más aproximado posible a la lengua hablada; 2) ficción, novelas que consisten en historias cortas; 3) prosa, ensayos que consisten en memorias, correspondencia, etc.; 4) periódicos, diarios, semanarios y publicaciones mensuales; y 5) literatura técnica, escritos sobre medicina, ingeniería, física, botánica, etc. Las cerca de 5 000 oraciones que sumaban más de 100 000 palabras, para cada uno de los géneros que integraban estos mundos fueron designadas por selección al azar”. [JUILLAND, A. 1964:VI]

En relación con el algoritmo metodológico que adopta, expone [JUILLAND, A. 1973:XXXV] la necesidad de refinar las técnicas de cuantificación al demostrarse que fallan las primeras intuiciones de lexicógrafos y estadísticos cuando identifican *uso* en el idioma con *frecuencia* y ordenan, en consecuencia, las palabras según coeficientes iguales, o proporcionales, al número de sus ocurrencias en la muestra. Propone como correctivo varias técnicas basadas en el concepto de *dispersión*. El cálculo que utiliza el equipo de Juilland es la *desviación estándar* respecto a la *media* estadística [JUILLAND, A. 1973:XXXVI-XXXVIII].

$$D = 1 - \frac{1}{2} * \frac{\sigma}{\bar{X}} = 1 - \frac{\sqrt{n \sum X_i^2 - T^2}}{2T} \quad [1]$$

El coeficiente de variación es  $\frac{\sigma}{\bar{x}}$

- n Representa el número de géneros, aquí 5.
- T Es la suma de las *frecuencias* en el género.

*T*, por tanto, es:

$$T = X_1 + X_2 + X_3 + \dots + X_n$$

$$T = \sum X_i$$

luego:

$$\sum X_i^2 = X_1^2 + X_2^2 + X_3^2 + \dots + X_n^2 = T^2 \quad [2]$$

La media de los valores de las *frecuencias* es

$$\bar{X} = \frac{1}{n} \sum X_i = \frac{T}{n} \quad \gg \quad T = n \bar{X} \quad \gg \quad n = \frac{T}{\bar{X}} \quad [3]$$

Entonces el  $\sigma^2$  de variación, que es el promedio de las diferencias de las medias, puede quedar recogido en la expresión:

$$\sigma^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 \quad [4]$$

La raíz cuadrada de la ecuación [4] es conocida en Estadística como *desviación típica*. También se puede escribir así:

$$\sigma^2 = \frac{\sum X_i^2 - n\bar{X}^2}{n} \quad [5]$$

Multiplicando [5] por  $n^2$  se llega a

$$n^2 \sigma^2 = n \sum X_i^2 - (n\bar{X})^2 \quad [6]$$

Acudiendo a [3] se puede sustituir parte del segundo miembro de la igualdad:

$$n^2\sigma^2 = n \sum X_i^2 - T^2 \quad [7]$$

El segundo miembro de [7] es, precisamente, el numerador bajo radical de la ecuación [1] que definía, para Juilland, la *dispersión*. Se puede, en consecuencia, describir el cálculo de la dispersión de la forma [8] considerando, además, el contenido de las igualdades recogidas en [3].

$$D = 1 - \frac{n\sigma}{2T} = 1 - \frac{\sigma}{2\bar{X}} \quad [8]$$

Es importante advertir que en *Frecuncy Dictionary of Italian Words* [JUILLAND, A. 1973:XXXVII y XXXVIII] se aprecian dos errores tipográficos<sup>4</sup> que en modo alguno permitirían llegar, como hace el autor, a la ecuación [8]. En efecto, en la fórmula [1] aparece

$$1 - \sqrt{\frac{n \sum X_i^2 - T^2}{2T}} \quad (DICE) \quad [1]$$

en lugar de

$$1 - \frac{\sqrt{n \sum X_i^2 - T^2}}{2T} \quad (DEBE DECIR) \quad [1]$$

En [8] dice:

$$1 - \frac{\sigma}{2\bar{X}} \quad (DICE) \quad [8]$$

---

<sup>4</sup> Aparecen, al menos, en la edición manejada por el doctorando en la Biblioteca Nacional de Madrid ejemplar de referencia 1/146668.

cuando quiere decir

$$1 - \frac{\sigma}{2\bar{X}} \quad (\text{DEBE DECIR})$$

[8]

Estas erratas no se observan, sin embargo, en *Frequency Dictionary of Spanish Words* [JUILLAND, A. 1964:XIV-LXXVIII] donde se explica con todo detalle el origen del *corpus* manejado y se demuestra la idoneidad de la fórmula de dispersión, la cual se erige a partir de las propiedades inherentes a la *curva normal de distribución o campana de Gauss* [JUILLAND, A. 1964:LIV].

El cálculo de la *dispersión* es sencillo. Se determina la *desviación estándar*, se divide entre el doble de la *media* obtenida y se resta este cociente de la unidad. Se consigue con ello un índice que oscila entre 0 y 1. Un valor  $D = 0$  significará una pésima *dispersión* y un valor  $D = 1$  la mejor *dispersión* posible. Una palabra cuyas ocurrencias se arracimen en una sola categoría tendrá una *dispersión* de 0 sin importar su *frecuencia*.

Por el contrario si las ocurrencias están igualmente distribuidas entre los géneros tiene un 1 de *dispersión* sin importar la *frecuencia*.

Pero ni la *frecuencia*, ni la *dispersión* de la ocurrencia de palabras en textos representativos o conversaciones cotidianas, son capaces de predecir su uso real en el conjunto del idioma [JUILLAND, A. 1973:XLI]. Por eso se intenta combinar los dos conceptos para precisar más la situación. El equipo de Juilland propone una fórmula que combina ambos indicadores buscando su reflejo proporcional en el guarismo resultante. Se espera para igual *frecuencia* un más alto o bajo *coeficiente de uso* de acuerdo con la *dispersión*, mientras que con la misma *dispersión* el *coeficiente de uso* pasa *depender* de la *frecuencia*.

$$USO = \frac{FRECUENCIA * DISPERSIÓN}{100}$$

“This means that for  $D = 1$ , which is the case when occurrences are equally dispersed among the categories of the sample,  $U = F$ ; for  $D = 0$ , which is the case when occurrences cluster in one single category,  $U = 0$ ; for  $D = 0.5$ ,  $U = F/2$ , etc. In other words,  $U$  varies between a high equal to  $F$  and a low

equal to 0, depending on the magnitude of  $D$ . Consequently,  $U$  is a dispersion percentage of  $F$ , which means that a less frequent word can be ranked ahead of a more frequent one, provided that its occurrences are more evenly dispersed in the genres of the sample”.

“Esto significa que para  $D = 1$ , que es el caso cuando las ocurrencias se dispersan igualmente entre las categorías de la muestra,  $U = F$ ; para  $D = 0$ , que es el caso cuando las ocurrencias arraciman en una sola categoría,  $U = 0$ ; para  $D = 0.5$ ,  $U = F/2$ , etc., es decir,  $U$  varía entre un alto igual a  $F$  y un nivel bajo igual a 0, dependiendo de la magnitud de  $D$ . Consecuentemente,  $U$  es un porcentaje de la dispersión de  $F$ , que significa que una palabra menos frecuente se puede alinear delante de otra más, con la condición de que sus ocurrencias se dispersen más uniformemente en los géneros de la muestra”. [JUILLAND, A. 1973:XLII].

De forma paralela a este proceso se acuña en Psicología la idea de *familiaridad léxica* que viene a tender un puente entre los planos lingüístico y psicológico. La *familiaridad léxica* se explica así:

“No es, como la *frecuencia de uso*, un índice estadístico de la lengua, sino un rasgo que caracteriza, para cada sujeto, su lenguaje” [FRAISSE, P. 1977a:180].

La base de estos trabajos se sustenta en la hipótesis que atribuye al fenómeno de la *familiaridad* un alto grado de correlación con la *frecuencia de uso* en virtud de la actuación de mecanismos de orden psicológico con presencia activa en los factores lingüísticos. Toman en consideración que la *frecuencia de uso* se encuentra en íntima relación con la rapidez y exactitud de la percepción de las palabras y el *tiempo de reacción verbal*.

La *familiaridad léxica* se puede definir operacionalmente como una *escala de medición* obtenida por procedimientos clásicos, esto es, a partir de la noción intuitiva que los sujetos tienen de la proximidad o lejanía de cada palabra. Experiencias de laboratorio basadas en la medición del *umbral* de reconocimiento de palabras [HOWES, D. y SOLOMON, R. 1951:401-410] han mostrado que ese indicador es tanto más bajo cuanto más grande es la *frecuencia* alcanzándose correlaciones de *Spearman* entre 0.57 y 0.87 [HOWES, D. 1954:106-112].

Es Francia, también, quien subvenciona los trabajos iniciales. Los más importantes los llevan a término Fraisse en París y Noizet en Marsella. Los experimentos responden básicamente a dos tipologías:

1. Se solicita a un grupo de sujetos que puntúen entre 1 y 5 cada una de las palabras incluidas en un listado, generalmente extraído de las escalas de Gougenheim y seleccionadas de acuerdo a su *frecuencia* y longitud, tomando como criterio su percepción subjetiva acerca de lo cercanas y familiares que les resultan.
2. Se mide el *tiempo de reacción* del sujeto a la hora de identificar y reconocer esas palabras presentadas a manera de estímulo.

Los resultados demuestran inequívocamente la existencia de importantes diferencias entre la *frecuencia* de uso y los *coeficientes de familiaridad*, algo ya advertido por los lingüistas, y que se explica porque el hablante utiliza corrientemente infinidad de palabras, además de las catalogadas como *gramaticales*, sin tomar conciencia de ello por incluirlas dentro de locuciones adverbiales o frases más o menos hechas (de un momento a otro, del mismo género, con gran pena...). Estos términos no son reconocidos como próximos cuando son presentados de forma aislada y con valor exclusivo en sí mismos. Fraise afirma que:

“Las palabras están ligadas a los objetos. Unas y otros están, por así decir, adosados de forma mutua, aun cuando la frecuentación de los objetos provoca un reforzamiento del vínculo entre el objeto significado y la palabra que lo designa, y por eso mismo aumenta la familiaridad de esa palabra. De ello resulta que ciertas palabras concretas, aunque de cierta *frecuencia de uso*, son, no obstante, muy familiares, porque los objetos que ellas significan se encuentran a diario. En este sentido, tal vez podríamos decir que si la *frecuencia de uso* atañe a la lengua, es decir, a una realidad sociológica, la familiaridad atañe al habla, vale decir, al empleo implícito o explícito de la lengua en el comportamiento”. [FRAISSE, P. 1977a:188].

La declaración anterior significa una crítica para la *frecuencia de uso* por su actuación sobre los signos aislándolos de las realidades que designan. Subraya, además, la dificultad para que un sujeto perciba como suya una palabra haciendo total abstracción del objeto por ella señalado. Este reconocimiento depende más de la cercanía de la cosa nombrada que de lo habitual del término dentro del habla cotidiana.

La *familiaridad* así planteada se coloca a medio camino entre la *frecuencia* de las situaciones y la *frecuencia* de utilización de las palabras correspondientes en el lenguaje explícito.

Cercano al concepto de *familiaridad* se encuentra el modelo de análisis léxico

conocido como *disponibilidad*. Según Gloria Butrón [BUTRÓN, G. 1987:15] es Michéa [MICHÉA. 1953:338-344] el primer lingüista en realizar una encuesta sobre *disponibilidad léxica*. Michéa define así el nuevo concepto:

“En présence d’une situation donnée, les mots que viennent le premiers à l’esprit sont ceux qui sont liés tout spécialement cette situation et la caractérisent. [...] Un mot disponible est un mot qui, sans être particulièrement fréquent, est cependant toujours prêt à être employé, et se présente immédiatement à l’esprit o ù moment ou l’on en a besoin”.

“En presencia de una situación dada, las palabras que vienen las primeras al espíritu son las que están especialmente vinculadas a esta situación y las caracterizan. [...] Una palabra disponible es una palabra que, sin ser especialmente frecuente, debe, sin embargo, estar siempre dispuesta a emplearse, y se presenta inmediatamente al espíritu en el momento en que es necesaria”. [MICHÉA, R. 1953:342].

Ya se ha comentado como los primeros vocabularios basados en el cálculo de la *frecuencia léxica* hacen surgir voces que recuerdan la existencia de términos de *frecuencia* muy baja y que, sin embargo, son habitualmente empleados en conversaciones cotidianas.

Hay palabras, caso de los artículos o las preposiciones sin las cuales sería imposible la comunicación, que tienen índices muy altos de *frecuencia* y aparecen de forma recurrente en cualquier tipo de relación comunicativa [MACKEY 1971]. Este léxico constituye aproximadamente un tercio de los elementos que componen la frase.

Por el contrario, hay otros términos muy relacionados con el tema a debate cuya *frecuencia* depende de la naturaleza del texto, del grupo social o de circunstancias coyunturales. Se impone, por tanto, la diferenciación entre léxico *frecuente* y léxico utilizado en situaciones frecuentes, entre *vocabulario frecuente* y *vocabulario disponible*.

Michéa [MICHÉA. 1953] también plantea la incorporación del concepto de *centro de interés* a los trabajos lexicológicos. Los *centros de interés* [DECROLY, O. 1922] son una propuesta pedagógica clásica que, desde un enfoque globalizador, intenta acercarse a los intereses naturales de los educandos. Van a resumir las “ideas-fuerza” que mueven o motivan a los alumnos y tiene su origen en las necesidades más elementales tanto físicas como intelectuales o sociales. Deben girar, dice Decroly, alrededor de cuatro grandes temas: alimentarse para conservar y desarrollar la vida, protegerse contra la

intemperie, defenderse contra el peligro y actuar y trabajar solidariamente, recrearse y mejorar.

La opción de Michéa consiste en computar los nombres temáticos o concretos a los que accede la memoria del hablante cuando afronta una situación específica que le sitúa en una determinada dirección semántica. La sugerencia es recogida por el propio Gougenheim que, en las sucesivas ediciones de su diccionario, clasifica el *corpus* que recoge en dieciséis centros de interés.

- “1. Les parties du corps.
2. Les vêtements.
3. La maison.
4. Les meubles de la maison.
5. Les aliments et boissons des repas.
6. Les objets placés sur la table et dont on se sert à tous les repas de la journée.
7. La cuisine, ses meubles et les utensiles qui s’y trouvent.
8. L’école, ses meubles et son matériel scolaire.
10. La ville.
11. Le village ou le bourg.
12. Les moyens de transport.
13. Les travaux des champs et du jardinage.
14. Les animaux.
15. Les jeux et distractions.
16. Les metieres”. [BUTRÓN, G. 1987:85].

Con la nueva propuesta metodológica se entra en el debate de cuántos y cuáles deben ser los *centros de interés* a considerar desde el punto de vista léxico. Opinan de forma y manera activa, entre otros, Njock [NJOCK. 1979], Galisson [GALISSON. 1979], Azurmendi [AZURMENDI. 1983], Justo Hernández [JUSTO HERNÁNDEZ. 1986], Canizal Arévalo [CANIZAL ARÉVALO. 1987], Max Echeverría [ECHEVERÍA, M. 1987]...

López Morales, a partir de 1983, define la *disponibilidad léxica* como “el caudal léxico utilizable en una situación comunicativa dada” [LÓPEZ MORALES, H. 1983a:213], vinculando definitivamente las palabras con la situación o contexto comunicativo en que son empleadas. Además, hace notar que el *índice de uso* de Juilland, en su oscilación entre 0 y 1, crece en la medida en que la *dispersión* no se reparte equitativamente entre todos los *mundos* considerados, lo que, en su opinión, significa una descompensación de la *frecuencia* a favor de la *dispersión* [LÓPEZ MORALES, H. 1983a:7].

El procedimiento habitual para determinar el nuevo indicador consiste en registrar de manera ordenada los vocablos que cada informante actualiza, en un tiempo dado, asociados a un *centro de interés* que se presenta como estímulo. Se supone que serán los más *disponibles* para el hablante en ese entorno conversacional. En un primer momento se puntúan las palabras teniendo en cuenta sólo su *frecuencia*, pero pronto se plantea la necesidad de cuantificar también el lugar de aparición. La demanda de ponderar el factor *posición* nace del modo de operar de la memoria que recupera antes los conceptos que le son más familiares. López Morales elabora, junto a Lorán [LÓPEZ MORALES, H. 1983b], una estrategia de cálculo del *índice* o *coeficiente de disponibilidad* que considera por primera vez las *frecuencias* en cada *posición*.

El índice de *disponibilidad léxica* se definió, en palabras de Butrón [BUTRÓN, G. 1987:21], “antes de proponerse ninguna forma concreta de calcularlo”. Ella misma ensaya modelos que complementan los de Lorán. En 1987 López Chávez y Strassburger Frías [LÓPEZ CHÁVEZ. 1987] presentan un algoritmo diferente.

Estimar la *disponibilidad* plantea varios problemas. Están, por ejemplo, la atribución de *pesos* a las *posiciones*, la recogida de datos mediante listas abiertas o listas cerradas y el tamaño de las muestras, que implican tratamientos numéricos adecuados y distintos. Además hay que hacer realmente comparables los indicadores resultantes de investigaciones diferentes.

El estudio de la *disponibilidad léxica* tiene hoy plena vigencia por su interés lingüístico y pedagógico, y no entra, en modo alguno, en contradicción con los léxicos básicos. Por el contrario, complementa los catálogos que especifican la proporción de uso de las palabras en el conjunto de la lengua, el *léxico básico*, con las nóminas concretas de términos, esencialmente nombres, que el hablante utiliza en contextos comunicativos específicos, el *léxico disponible*.

*De la frecuencia a la disponibilidad léxica*

