Lexicometría de corpus

Jordi Porta Zamorano y **Rafael J. Ureña Ruiz**
Departamento de Lingüística Computacional
Real Academia Española
c/ Academia 1. 28071 Madrid
{porta,rafa}@rae.es

1. *El sistema de consulta de corpus*

El *DLC* está desarrollando un sistema de consulta de corpus para finalidades lexicométricas consistente en un indexador y un lenguaje de interrogación. El indexador está basado en **ficheros invertidos** y **árboles de intervalos** y es capaz de representar la información contenida en textos anotados lingüísticamente con marcación estructural de tipo SGML. El lenguaje de interrogación permite la recuperación de la concordancia y la distribución de su frecuencia haciendo referencia literal o mediante expresiones regulares a cualquiera de las unidades de información indexadas o una combinación de ellas mediante operadores booleanos o de secuencia, con posibilidad de restringir estructuralmente el ámbito de búsqueda.

Las unidades indexadas del corpus *CREA* son: palabras (incluidas unidades textuales como fechas, números y entidades con nombre), lemas, categorías, rasgos morfosintácticos y otras abstracciones léxicas.

Un cliente *web* permite la consulta lexicométrica y presenta esta información mediante tablas y gráficos que ayudan a la interpretación de los resultados.

2. *Lexicometría*

Una de las medidas más usadas en lexicografía basada en corpus y en diccionarios de frecuencias es la **dispersión léxica** (D). La dispersión refleja la distribución de la frecuencia a lo largo de las particiones de un corpus. Existen otras medidas lexicométricas interesantes derivables de la dispersión: la **concentración** (C), que es la medida complementaria a la dispersión, formulada como $C = 1 - D$, y el **uso** (U), que permite la caracterización de léxicos nucleares, cuya formulación más simple es $U = F \cdot D$, donde F es la frecuencia.

Para que un índice de dispersión pueda ser interpretable, las particiones deben ser **coherentes**, esto es, las divisiones de un corpus deben realizarse con criterios lingüísticamente relevantes. Los ejes clásicos para la caracterización del léxico son el diacrónico, el diatópico y el diastrático. En el caso del *CREA* los criterios de diseño inducen distintas particiones coherentes: cronológica (año de publicación), geográfica (área, zona y país), temática (área y subárea) y por formato de publicación.

2.1. El índice de dispersión de *Carroll*

La modelización más conocida de dispersión es la de Juilland y Chang-Rodriguez (1964). La restricción más importante para la aplicación de este índice es que las particiones deben ser del mismo tamaño. Aunque ha habido intentos de reformulación del índice de *Juilland* para particiones de tamaño desigual (Rafel i Fontanals, 1996), este índice ha sido usado recientemente, a costa de la coherencia en el particionado, por Leech, Rayson, y Wilson (2001).

Carroll (1970) propone una medida alternativa a la dispersión de *Juilland*, basada en el concepto de **entropía**, que evita el sesgo producido por el diferente tamaño de las particiones. El corpus se subdivide en n particiones en las que se calcula, para cada partición i, la subfrecuencia F_i del elemento cuya dispersión queremos medir. Las frecuencias relativas f_i de cada partición se obtienen dividiendo cada F_i por el tamaño n_i de la partición. La probabilidad de aparición en cada partición i es $p_i = f_i / \sum f_i$, siendo la formulación final de la dispersión $D(p) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i / \log_2 n$. Esta modelización mide la incertidumbre o heterogeneidad de la distribución de probabilidad de las particiones. El denominador $\log_2 n$ normaliza la entropía para obtener un índice en el intervalo $[0, 1]$.

3. Ejemplos del CREA

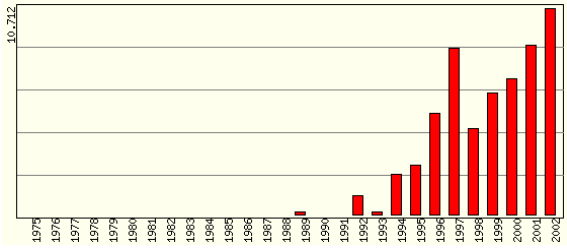
Como ejemplos de la información proporcionada por el sistema presentado, podemos observar que los lemas *portavoz* y *vocero* tienen una distribución geográfica complementaria: mientras que la frecuencia del primero se concentra en España ($D_H = 16{,}1\,\%$; $D_E = 83{,}9\,\%$), el segundo lo hace en Hispanoamérica ($D_H = 97{,}6\,\%$; $D_E = 2{,}4\,\%$). La dispersión por área temática nos muestra que ambos lemas concentran sus apariciones en textos informativos (*portavoz* 75,7 %; *vocero* 74,4 %), en concreto en las subáreas temáticas de política y finanzas (57,2 %; 52,1 %, respectivamente). Por otro lado, la dispersión según el formato de publicación nos muestra que *portavoz* y *vocero* pertenecen al vocabulario de los medios de comunicación:

	Prensa	Revista	Libro	Misc.
portavoz	86,1 %	7,3 %	3,5 %	3,0 %
vocero	51,9 %	44,2 %	3,2 %	0,7 %

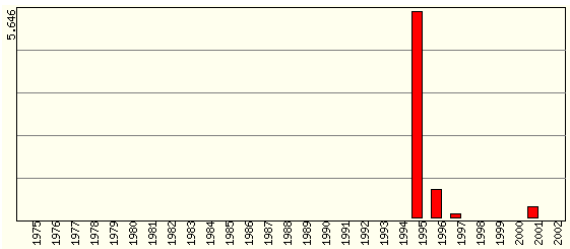
También muestran distribuciones geográficas contrarias *teléfono celular* y *teléfono móvil*:

	América	España
teléfono móvil	11,3 %	88,7 %
teléfono celular	94,0 %	6,0 %

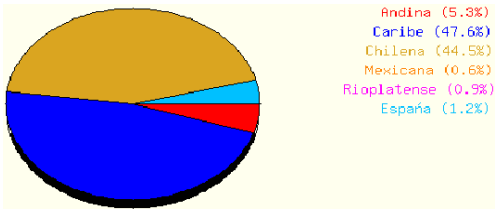
El uso acumulado de ambos va en aumento como puede apreciarse en el siguiente histograma de frecuencias relativas:



y su dispersión cronológica (D_C) es de 0,644. Compárese con la dispersión en el tiempo de *de* cuya frecuencia relativa se mantiene a lo largo del tiempo entre un 4–5 % ($D_C = 0{,}989$) y con el lema *fletán* ($D_C = 0{,}174$), de aparición casi exclusiva en el año 1995 en prensa (98,6 %), como se muestra en el histograma siguiente:



Tomemos ahora como ejemplo *guagua*. En el *DRAE* hay dos entradas que se corresponden principalmente con *guagua*[1] (*autobús*) y *guagua*[2] (*niño* o *pan con forma de niño*). La consulta del lema *guagua* al corpus muestra la siguiente distribución por área geográfica: $D_H = 90{,}1\,\%$; $D_E = 9{,}9\,\%$, lo cual coincide con las diferentes marcas geográficas del *DRAE*, mayoritariamente americanas (excepto Canarias). Además, la distribución por zonas y países dentro de América no es homogénea, como se muestra en el siguiente gráfico de distribución de frecuencias relativas:



Las concordancias documentan la doble entrada del *DRAE*: *guagua*[1] en la zona Caribe y *guagua*[2] en las zonas Chilena y Andina. En el resto de zonas su presencia no es significativa desde el punto de vista estadístico.

4. Conclusiones

El sistema presentado permite mostrar la distribución léxica. La información proporcionada permite, entre otras, la validación de abreviaturas y marcas en diccionarios, la detección de arcaísmos y neologismos, la confección de recursos de anotación especializados y léxicos nucleares para alguna combinación de los ejes de variación contemplados.

Bibliografía

Carroll, J. B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behaviour*, (3):61–65.

Juilland, A. y E. Chang-Rodriguez. 1964. *Frequency Dictionary of Spanish Words*, volumen S1. Mouton & Co.

Leech, G., P. Rayson, y A. Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman.

Rafel i Fontanals, Joaquim. 1996. *Diccionari de freqüències: Llengua no literària*, volumen 1. Institut d'Estudis Catalans.