Ibérica
ISSN: 1139-7241
iberica@aelfe.org
Asociación Europea de Lenguas para Fines Específicos
España

Campillos Llanos, Leonardo; Ueda, Hiroto
Frecuencia y dispersión léxicas en textos médicos divulgativos en español
Ibérica, núm. 30, 2015, pp. 61-84
Asociación Europea de Lenguas para Fines Específicos
Cádiz, España

Frecuencia y dispersión léxicas en textos médicos divulgativos en español

Leonardo Campillos Llanos y Hiroto Ueda
Universidad Autónoma de Madrid (Spain) y University of Tokyo (Japón)
leonardo.campillos@uam.es & hiroto.ueda.tokio@gmail.com

Resumen

Se analiza la frecuencia, dispersión y uso de las categorías léxicas (sustantivos, adjetivos, verbos, epónimos y adverbios en *–mente*) en 363 textos médicos divulgativos en español (334.690 palabras). Se propone el concepto de dispersión lineal, adecuado para analizar corpus homogéneos que carecen de secciones diferenciadas. Los resultados muestran que las palabras más usadas no portan significados asociados al dominio médico, excepto los sustantivos. Asimismo, las palabras en los rangos más bajos de uso están fuertemente determinadas por los contenidos de cada texto. Los sustantivos y los adjetivos son las categorías que presentan mayor riqueza léxica (más lemas diferentes), lo cual se asocia al carácter expositivo del corpus. El artículo ofrece un listado de lemas más usados, de utilidad para los profesionales que trabajan en este dominio.

Palabras clave: frecuencia, dispersión lineal, uso, textos divulgativos médicos, español.

Abstract

Frequency and lexical dispersion in medicine popularisation texts in Spanish

We analyse frequency, dispersion and usage of lexical categories (nouns, adjectives, verbs, eponyms and adverbs ending in *–mente*) in 363 medicine popularisation texts in Spanish (334690 words). We propose the concept of linear dispersion, which is suitable for analysing homogeneous corpora without different subsections. Results indicate that the most commonly used words do not have medical meanings, excepting nouns. In addition, words in the lowest ranks are heavily dependent on the contents of each text. Nouns and adjectives show the highest lexical richness (more different lemmas), which is related to the informative nature of the corpus. The article includes an appendix with the most used lemmas that may be useful for professionals working in this domain.

Keywords: frequency, linear dispersion, usage, medicine popularisation texts, Spanish.

1. Introducción

Desde principios del siglo XX, los estudios estadísticos del léxico (o lexicométricos) suscitaron interés por conocer y mejorar la adquisición de la lengua materna (L1) y/o la lengua segunda (L2) (Alvar Ezquerra, 2005; véase un panorama completo en Ávila Martín, 2010). La mayor parte de los trabajos obtuvo datos de corpus generales (no restringidos a una temática especializada). Keniston (1920), por ejemplo, emprendió uno de los primeros trabajos para el español recopilando las frecuencias de las palabras de uso cotidiano a partir de obras teatrales, novelas, prensa y narrativa. A dicho estudio le siguieron los de Buchanan (1927) a partir de textos escritos, así como el de Rodríguez Bou et al. (1952), que abordaron el vocabulario de los preescolares en Puerto Rico. En España no fue hasta mediados de los años 50 cuando García Hoz (1953) recogió un listado de vocabulario con fines pedagógicos. Su corpus ascendía a 400.000 palabras y abordaba cuatro planos de uso del lenguaje: cartas privadas, documentos oficiales, periódicos y literatura. Por su parte, Juilland y Chang Rodríguez (1964) emplearon un corpus de novelas, ensayos, obras teatrales, textos de dominio especializado y periódicos (alrededor de 500.000 palabras). Dicho proyecto constituyó el modelo para otras variedades del español (por ejemplo, de Puerto Rico; Morales, 1986). Con finalidad pedagógica, Ueda (1987, 1989-1990) recogió listados de frecuencia analizando textos de 10 campos diferentes.

Los avances informáticos han permitido obtener cómputos manejando más datos, aunque los corpus han mantenido un esquema similar. Alameda y Cuetos (1995) recopilaron un banco de datos de 2 millones de palabras; Sebastián, Martí y Carreiras (2000) alcanzaron los 5 millones y trabajos más recientes han aumentado tal cifra. Así, el volumen de Almela, Cantos, Sánchez y Almela (2005) o el diccionario de frecuencias del español (Davies, 2006) han utilizado corpus de aproximadamente 20 millones de palabras.[1]

Los estudios lexicométricos han permitido observar fenómenos cuantitativos del lenguaje. Por ejemplo, la regularidad estadística del léxico debida a la abundancia de vocablos de uso general frente a la baja aparición de palabras específicas del contenido, como formula la ley de Zipf (1949). Respecto a la riqueza léxica, también está probado empíricamente que el

aumento de palabras de un corpus no conlleva un crecimiento continuo del número de lemas diferentes, sino que progresivamente se alcanza una fase de estancamiento. Así lo predice la ley de Herdan (1964) o se verifica midiendo la proporción entre tipos y casos (*type/token ratio*) (véanse detalles en Cantos, 2013).

Los conceptos de frecuencia de aparición de una palabra, su dispersión (o distribución) en distintos textos (Juilland y Chang Rodríguez, 1964; Carroll, 1970; Gries, 2008) o su valor de uso ofrecen perspectivas complementarias para lograr una visión global del comportamiento estadístico de la lengua (Porta Zamorano y Ureña Ruiz, 2003). Por ejemplo, la mayor frecuencia de aparición de un vocablo en un documento no implica mayor peso en toda la colección de textos, si dicha voz se registra solo en una porción del corpus.

En el ámbito de la lingüística, el estudio de lenguas de especialidad (Cabré, 2004) surgió, entre otros, por las necesidades lexicográficas de elaborar diccionarios y vocabularios de tecnicismos (Rodríguez Díez, 1979). Paralelamente, en el área del procesamiento del lenguaje natural nació el interés por conocer los patrones lingüísticos de las sublenguas debido a sus implicaciones ingenieriles (Temnikova y Cohen, 2013). Dicho conocimiento ha contribuido a mejorar las aplicaciones de traducción automática (Sommers, 2000), minería de texto, o extracción y recuperación de información (Friedman, Anderson, Austin, Cimino y Johnson, 1994). En cuanto al español, existen análisis del lenguaje periodístico (Hernando, 1990), pero escasean los estudios en el dominio médico realizados con una metodología complementaria a los trabajos de Navarro (1997, 2005). Esto contrasta con el volumen de estudios sobre el inglés médico (Skelton y Whetstone, 2012; Mungra y Canziani, 2013; Verdaguer, Laso y Salazar, 2013; Vila Barbosa, 2013).

El objetivo de este trabajo es aportar un estudio lexicométrico de documentos médicos divulgativos en español. El artículo presenta el análisis de frecuencia, dispersión y uso del vocabulario usado en textos extraídos del corpus MultiMedica (Moreno Sandoval y Campillos Llanos, 2013). Se ofrecen solo los datos de las categorías léxicas más usadas (por frecuencia y dispersión): adjetivos, verbos, sustantivos y adverbios acabados en *–mente*. Asimismo, indicamos en las listas las palabras recogidas en el *Diccionario de términos médicos* (en adelante, *DTM*) de la Real Academia Nacional de Medicina (en adelante, RANM, 2011). Los datos aquí reunidos son novedosos y de interés para la elección del léxico en la enseñanza del español

con fines específicos, la lexicografía especializada, o el procesamiento del lenguaje natural en este dominio.

2. Metodología

Esta sección aborda, en primer lugar, el corpus lingüístico empleado para el análisis; en segundo lugar, el proceso de etiquetado morfológico y la obtención de lemas del corpus; y por último, los conceptos estadísticos considerados en nuestro estudio.

2.1. Corpus

Los datos se han obtenido de los textos divulgativos en español del corpus MultiMedica (Moreno Sandoval y Campillos Llanos, 2013). En total, reúne 363 textos (334.690 palabras): 297 de la revista *OCU-Salud*, que publica textos escritos por médicos y editados por periodistas; y 66 textos de *Tu otro médico*, sitio web que reúne artículos enciclopédicos escritos por médicos para el público no especializado. El corpus carece de subsecciones o conjuntos definidos en que agrupar textos de tipología homogénea, lo cual nos llevó a proponer un concepto distinto de dispersión (sección 2.2). Los textos fueron publicados entre los años 1997 y 2008, de modo que el corpus recoge vocablos de actualidad, pero no impide que se soslayen términos y conceptos incorporados posteriormente en la jerga (nuevos procedimientos diagnósticos, técnicas, fármacos y enfermedades).

2.2. Etiquetado morfológico y obtención de lemas

La anotación morfológica se llevó a cabo mediante GRAMPAL (Moreno Sandoval y Guirao Miras, 2006), un analizador para el español que consta de un vocabulario de más de 50.000 lemas, entre los cuales se recogen unidades léxicas complejas como locuciones (*atención primaria*) o marcadores discursivos (*es decir*). GRAMPAL segmenta unidades léxicas contraídas: por ejemplo, *al* o *del* reciben cada una dos anotaciones (una correspondiente a la preposición, y otra al artículo). De igual modo anota la información de cada categoría en los verbos con clíticos: por ejemplo, *dándoselo* se segmenta en tres anotaciones: verbo (*dar*), pronombre (*se*) y pronombre (*lo*).[2] El procesamiento es automático, pero requiere revisión posterior para corregir errores. Dos lingüistas revisaron y corrigieron los datos durante 4 meses.

Los errores se debieron, en primer lugar, a la incorrecta segmentación de unidades: por ejemplo, el analizador marcó *vamos a ver* como marcador discursivo, en un contexto en que era verbo, preposición e infinitivo: *Vamos a ver los siguientes síndromes*. En segundo lugar, multitud de palabras no recibieron una categoría o lema porque no fueron reconocidas. Este problema fue recurrente entre los términos médicos, dado que GRAMPAL está preparado únicamente para el lenguaje general. En tercer lugar, aparecieron errores por asignación incorrecta de categorías. Estos fueron reiterados entre vocablos homónimos (por ejemplo, *vino*, sustantivo o verbo) y entre palabras policategoriales, especialmente, sustantivos y adjetivos (por ejemplo, *químico* puede ser tanto adjetivo como nombre). Mención especial merecen las ambigüedades en los participios con función adjetiva, que plantean problemas de categorización (véase su tratamiento en RAE, 2009: 2214-ss). Generalmente, dichas palabras se anotaron como adjetivos (por ejemplo, *acatarrado*), siempre que no se acompañaran de complementos regidos, en cuyo caso se anotaron como verbos.

A pesar de las ambigüedades de categorización y los errores de etiquetado, un análisis del acuerdo entre anotadores de un 5% de los textos (18) mostró que el acuerdo ronda el 98% de medida F,[3] así que se estima una consistencia notable en la anotación.

Tras anotar y revisar el etiquetado, se extrajeron de los textos los lemas y las formas, que fueron analizados mediante nuestro programa "Frecuencia y dispersión lineal".

2.3. Conceptos estadísticos

Las medidas computadas son: frecuencia (absoluta y normalizada por mil), rango de frecuencia, dispersión lineal media, rango de dispersión lineal media, valor de uso y rango de uso. A fin de esclarecer la comprensión de dichos valores, explicaremos cada concepto tomando como ejemplo el lema *aditivo* (Tabla 1; Frec.: frecuencia absoluta; Frec./mil: frecuencia normalizada por mil; RF: rango de frecuencia; DLM: dispersión lineal media; RDLM: rango de dispersión lineal media; U: valor de uso; RU: rango de uso).

Lema	Frec.	Frec./mil	RF	DLM	RDLM	U	RU
aditivo	83	0,226	5	0,690	7	57,285	4

Tabla 1. Ejemplo de valores estadísticos correspondientes al lema "aditivo".

Para obtener la frecuencia absoluta, calculamos el recuento final de cada vocablo. La frecuencia normalizada por mil es la cifra más adecuada para poner en relación los resultados con otros corpus. El cálculo se obtiene dividiendo la ocurrencia de cada palabra entre el total de palabras y multiplicando el resultado por mil. Así, la frecuencia absoluta para el lema aditivo (83 ocurrencias) se corresponde con una frecuencia normalizada por mil de 0,226, calculada del siguiente modo: 83/367.482 x 1000 = 0,226.

El rango de frecuencia facilita la adscripción de las palabras a subconjuntos del corpus establecidos conforme a la cifra de ocurrencias. En el volumen de Almela et al. (2005), por ejemplo, se dividieron los resultados en cinco bandas de frecuencia (baja, moderada, notable, alta y muy alta). En nuestro trabajo, calculamos diez (n=10) rangos de frecuencia basados en la definición siguiente. En primer lugar, en vez de utilizar la frecuencia (F) misma, utilizamos el logaritmo para obtener una distribución lineal: Log(F). El logaritmo expresa adecuadamente el comportamiento del léxico conforme a su frecuencia de aparición. Su representación gráfica corresponde a una curva en que los rangos de mayor frecuencia presentan una pendiente más pronunciada, mientras que los rangos de frecuencia más baja se aproximan a la horizontal (Baayen, 2001: 14; Ueda, 2013). La base del logaritmo se define con la siguiente fórmula:

$$\text{Base del logaritmo} = Mx^{1/10}$$

donde Mx es la frecuencia máxima recogida en nuestros datos (esto es, Mx = 35.536 ocurrencias correspondientes al artículo *el*). Así, la base del logaritmo es $35.536^{1/10}$ = 2,851. Con esta base, la máxima frecuencia se convierte en 10 (es decir, el logaritmo en base 2,851 de 35.536 es 10):

$$\text{Log } Mx^{1/10}\,(Mx) = 10 \quad \Longrightarrow \quad \text{Log}_{2,851}\,(35536) = 10$$

A continuación, para obtener los rangos de 1 a 10, calculamos el logaritmo en base 2,851 de las ocurrencias de cada palabra, con cifras redondeadas en el rango superior a partir del dígito 0. Por ejemplo, si *aditivo* presenta una frecuencia de 83 ocurrencias, entonces el rango se calcula así: $\text{Log}_{2,851}$ (83) = 4,21, y el rango de frecuencia se redondea a 5 por rebasar la cifra de 4. Se trata de un vocablo de ocurrencia moderada.

En cuanto a la dispersión lineal media, proponemos un concepto distinto respecto a trabajos previos. Juilland y Chang-Rodríguez (1964) tomaron la

desviación estándar observada entre frecuencias de un vocablo en distintos textos, empleando subcorpus de tamaño semejante. Posteriormente, Carroll (1970) propuso un concepto de dispersión adecuado a un corpus con particiones de diferente tamaño, aunque Gries (2008) revisó su propuesta con una formulación posterior. En este trabajo, a diferencia de dichos investigadores, consideramos el conjunto de textos como un continuo. El motivo es la ausencia de secciones o subcorpus en nuestra colección. Considerando el texto como un continuo, también se observan distintos grados de dispersión de los vocablos, variando desde una dispersión ideal máximamente equitativa (las palabras se distribuyen en el texto de manera uniforme) hasta una dispersión extremadamente sesgada (las palabras se concentran solo en ciertos puntos del continuo textual). El valor de dispersión de un vocablo es una cifra comprendida entre 0 y 1, siendo 0 el valor de dispersión mínima, y 1 el de la dispersión máxima. Nuestro propósito es calcular el grado de lo sesgada que está la dispersión real (a) con respecto a la dispersión ideal equitativa (b) (figura 1).

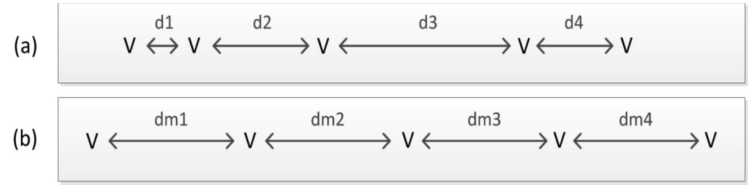

Figura 1: Dispersión real (a) y dispersión lineal media (b) de un vocablo V.

La figura representa un vocablo V con cinco apariciones en el corpus, entre las cuales hay cuatro intervalos de distancia (d1, d2, d3 y d4). La dispersión lineal media (b) representa en el continuo textual la distancia media (dm, siendo dm = dm1 = dm2 = dm3 = dm4). La distancia media se calcula dividiendo la cifra total de palabras del corpus (T) entre la frecuencia de un determinado vocablo (F), según esta fórmula:

$$dm = \frac{T}{F}$$

Para calcular la dispersión lineal normalizada (DLN), son precisos varios cómputos intermedios. Primero, se obtiene la distancia acumulada al cuadrado, que es el cuadrado de la suma acumulada de las diferencias entre la distancia real de aparición de cada vocablo (d_i) y la distancia ideal media (dm):

$$\sum_i (d_i - dm)^2$$

Como explicamos antes, el valor de la dispersión lineal normalizada es una cifra entre 0 (dispersión mínima) y 1 (dispersión máxima). Así, las categorías que aparecen con la distribución más alta en el corpus presentan un valor de dispersión próximo a 1: por ejemplo, el artículo *el* (con un valor de 0,985). En cambio, lemas con una concentración alta (restringida solo a ciertas secciones del corpus) presentan un valor próximo a 0: por ejemplo, *quiropráctica* (0,063). El cálculo de la dispersión lineal de cada ocurrencia se obtiene a partir de la desviación estándar (DE):

$$DE = \sqrt{\frac{\sum_i (d_i - dm)^2}{F}}$$

Por su parte, la desviación estándar normalizada (DEN) requiere calcular el valor medio de distancia (M), que es el cociente entre la suma de distancias y la frecuencia:

$$M = \frac{\sum_i d_i}{F}$$

La desviación estándar normalizada (DEN) resulta de dividir la desviación estándar (DE) entre el producto del valor medio de distancia (M) y la raíz cuadrada de F - 1:

$$DEN = \frac{DE}{M\sqrt{F-1}}$$

Finalmente, la dispersión lineal normalizada (DLN) resulta de restar de 1 la desviación estándar normalizada (DEN):

$$DLN = 1 - DEN$$

Intentaremos aclarar cómo se calcula la dispersión lineal normalizada tomando el ejemplo de *aditivo*. Para comprender el concepto, se puede observar este fragmento:

> ¡Su comida, sin *aditivos*! Los *aditivos* desaconsejables o susceptibles de provocar efectos indeseables. Los alimentos que contienen más cantidad de *aditivos* (…)

La distancia entre la primera aparición del lema *aditivo* (posición 7) y la segunda (posición 10) es de 3 (10 – 7 = 3). La distancia entre la segunda aparición (posición 10) y la tercera (posición 25) es de 15 (25 – 10 = 15). Por otra parte, la distancia media (dm) es el resultado de dividir el número total de elementos del corpus (367.482) entre el número total de intervalos entre las ocurrencias de la palabra correspondiente (en el caso de *aditivo*, 83 intervalos). Así, la distancia media (ideal) entre cada aparición de *aditivo* en nuestro corpus es 367.482 / 83 = 4.427,49 palabras. A continuación, se calcula la diferencia entre la distancia media ideal y la distancia real de cada una de las ocurrencias. Entre la primera ocurrencia y la segunda, la diferencia de la distancia ideal y la real es de 4.427,49 – 3 = 4.424,49 palabras y su valor cuadrado es $4.424{,}49^2$ = 19.576.111,76. Si se aplica este procedimiento a las 83 apariciones de dicho lema, se obtiene la suma de diferencias (o diferencia acumulada) entre la distancia real de aparición del vocablo (d_i) y la distancia media (dm) o ideal. Estos valores permiten calcular la desviación estándar (DE), la desviación estándar normalizada (DEN) y la dispersión lineal normalizada (DLN) mediante las fórmulas anteriormente expuestas. El resultado de la dispersión lineal normalizada de *aditivo* es 0,690.

Seguidamente, para calcular los rangos de dispersión de 1 a 10, redondeamos el resultado de DLN multiplicada por 10 en el rango superior a partir del dígito 0. Así, la dispersión lineal normalizada de *aditivo* (0,690) se sitúa en un rango de dispersión 7.

Finalmente, el valor de uso (U) es el producto de frecuencia (F) por dispersión (D) (Juilland y Chang Rodríguez, 1964; Porta Zamorano y Ureña Ruiz, 2003):

$$U = F \times D$$

Así, el valor de uso de *aditivo* es 83 x 0,690 = 57,285. Por último, se calcula el rango de uso de igual modo que el rango de frecuencia o dispersión. En este caso, *aditivo* aparece en un rango de uso de 4.

Los cálculos se realizaron con el programa "Frecuencia y dispersión lineal". La figura 2 muestra los resultados durante el procesamiento. En las columnas se recogen los siguientes valores: *Freq.*: Frecuencia; *Prev. pos.*: Posición previa; *Distance*: Distancia; *Mean Dist.*: Distancia media; *Difference*: Diferencia; y *Acum. Diff*: Diferencia acumulada. Los valores aparecen a partir de la segunda ocurrencia de un vocablo (pues son necesarias la posición previa y la distancia entre palabras).

N	LEMA_CAT	Freq.	Prev.pos.	Distance	Mean.Dist.	Difference	Accum.Diff.
1	¡_B						
2	SU_T						
3	COMIDA_S						
4	,_B						
5	SIN_P						
6	ADITIVO_S						
7	!_B						
8	EL_T						
9	ADITIVO_S	83	7	3	4427.4940	4424.4940	4424.494
10	DESACONSEJABLE_A						
11	O_C						
12	SUSCEPTIBLE_A						
13	DE_P						
14	PROVOCAR_V						
15	EFECTO_S						
16	INDESEABLE_A						
17	EL_T	35481	9	9	10.3571	1.3571	1.357
18	ALIMENTO_S						
19	QUE_R						
20	CONTENER_V						
21	MÁS_T						
22	CANTIDAD_S						
23	DE_P	20620	14	10	17.8216	7.8216	7.821
24	ADITIVO_S	83	10	15	4427.4940	4412.4940	8836.988

Figura 2. Resultados de la fase intermedia del procesamiento (se usa punto para decimales).

La figura 3 muestra los resultados finales. Las columnas recogen los siguientes valores: lema (*Lemma*), categoría gramatical (*Cat.*), frecuencia absoluta (*Freq.*), rango de frecuencia (*F. Rank*), frecuencia normalizada por mil (*F. per mil*), dispersión lineal media (*L. Disp.*), rango de dispersión lineal (*L. D. Rank*), valor de uso (*Usage*) y rango de uso (*U. Rank*).

Lemma	Cat.	Freq.	F.Rank	F.Permil	L.Disp.	L.D.Rank	Usage	U.Rank
¡	B	46	4	.125	.812	9	37.330	4
SU	T	2148	8	5.845	.973	10	2089.562	8
COMIDA	S	162	5	.441	.839	9	135.926	5
,	B	20226	10	55.039	.993	10	20085.227	10
SIN	P	426	6	1.159	.943	10	401.573	6
ADITIVO	S	83	5	.226	.690	7	57.285	4
!	B	46	4	.125	.812	9	37.332	4
EL	T	35481	10	96.552	.995	10	35307.620	10
DESACONSEJABLE	A	9	3	.024	.566	6	5.094	2
O	C	2521	8	6.860	.976	10	2461.729	8
SUSCEPTIBLE	A	21	3	.057	.733	8	15.396	3
DE	P	20622	10	56.117	.993	10	20486.983	10
PROVOCAR	V	287	6	.781	.908	10	260.631	6
EFECTO	S	528	6	1.437	.914	10	482.558	6

Figura 3. Resultados finales tras el procesamiento de los datos (se usa punto para decimales).

3. Resultados

Solamente ofrecemos recuentos de palabras con una frecuencia mínima de 2 ocurrencias, excluyendo las que solo se registran una vez en nuestro corpus (*hápax legómenon*). El reparto muestra que más de tres cuartos (82,89%) corresponde a categorías léxicas, a diferencia del resto de categorías

(17,11%). El desglose por tipo de categoría léxica revela que un poco menos de la mitad son sustantivos (45,03%); un poco más de la quinta parte (21,16%), adjetivos; y les siguen verbos (14,33%) y adverbios en *–mente* (2,21%). Los epónimos (términos que reciben el nombre propio de investigadores: por ejemplo, *Alzheimer*; Alcaraz Ariza, 2002) resultan prácticamente esporádicos (0,15%). Cabe plantearse si un análisis de textos técnicos del mismo dominio obtendría resultados semejantes. La tabla 2 y la figura 4 resumen los datos.

		Frecuencia de lemas	
		Absoluta	**Relativa**
Categorías léxicas	Sustantivos	3.526	45,03%
	Adjetivos	1.657	21,16%
	Verbos	1.122	14,33%
	Adverbios en *-mente*	173	2,21%
	Epónimos	12	0,15%
	Total	6.490	82,89%
Otras categorías		1.340	17,11%
Total de lemas		7.830	100%

Tabla 2. Frecuencia absoluta y relativa de lemas por categorías léxicas.

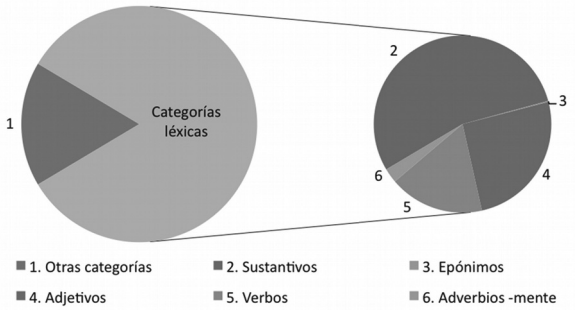

Figura 4. Proporción de lemas obtenidos por categorías.

Resulta interesante contrastar nuestras cifras con las del corpus general CUMBRE (Almela et al., 2005: 55). En ambos las categorías más frecuentes coinciden, lo cual parece revelar que su frecuencia es independiente del tipo de texto (especializado o misceláneo). No obstante, en nuestros datos los adjetivos muestran una frecuencia más alta que los verbos, siendo a la inversa en CUMBRE. Esto sí puede resultar distintivo del tipo de género, pues el uso del adjetivo (especialmente, relacional) se explica por el carácter descriptivo o expositivo de nuestros textos (recuérdese que son divulgativos, pero están escritos por médicos y editados por periodistas, y prima la función

informativa). En cambio, en un corpus general los textos pueden ser de género variado (por ejemplo, narrativo) y el uso del verbo aporta dinamismo. Nuestros resultados coinciden con los de Hernando (1990) y Sabaj (2004), quien contabilizó menos verbos en el corpus científico técnico al compararlo con otros corpus distintos.

A continuación presentamos los resultados de cada tipo diferente (*type*) de lema por categoría léxica. Primero, exponemos los lemas léxicos más frecuentes, indicando el porcentaje que representan respecto a la totalidad

Orden	Lema	Cat.	Frec./mil	% del corpus	RF	RD	RU
1	ser	V	18,874	1,89%	9	10	9
2	haber	V	7,260	0,73%	8	10	8
3	poder	V	7,121	0,71%	8	10	8
4	tener	V	4,308	0,43%	8	10	8
5	estar	V	3,758	0,38%	7	10	7
6	deber	V	3,176	0,32%	7	10	7
7	hacer	V	2,640	0,26%	7	10	7
8	**tratamiento**	S	2,449	0,24%	7	10	7
9	**enfermedad**	S	2,231	0,22%	7	10	7
10	**medicamento**	S	1,992	0,20%	7	10	7
11	**caso**	S	1,899	0,19%	7	10	7
12	problema	S	1,829	0,18%	7	10	7
13	**paciente**	S	1,826	0,18%	7	9	7
14	**persona**	S	1,788	0,18%	7	10	7
15	**riesgo**	S	1,769	0,18%	7	10	7
16	**médico**	S	1,641	0,16%	7	10	7
17	**producto**	S	1,614	0,16%	7	9	6
18	**estudio**	S	1,516	0,15%	7	10	6
19	**efecto**	S	1,437	0,14%	6	10	6
20	tratar	V	1,410	0,14%	6	10	6
21	producir	V	1,350	0,13%	6	10	6
22	tomar	V	1,333	0,13%	6	10	6
23	**síntoma**	S	1,314	0,13%	6	9	6
24	**tipo**	S	1,282	0,13%	6	10	6
25	**año**	S	1,238	0,12%	6	10	6
26	**ver**	V	1,230	0,12%	6	10	6
27	realizar	V	1,118	0,11%	6	10	6
28	**niño**	S	1,102	0,11%	6	9	6
29	mismo	A	1,088	0,11%	6	10	6
30	**alimento**	S	1,080	0,11%	6	9	6

Tabla 3. Los 30 lemas con mayor rango de uso en el corpus.

de palabras del corpus (367.482). Nótese que los resultados no son comparables a los obtenidos por Juilland y Chang Rodríguez (1967) o Patterson y Urrutibéheity (1975), ya que ellos calcularon las frecuencias incluyendo las categorías gramaticales. Las 50 palabras léxicas más frecuentes suponen más del 10% del tamaño del corpus (en concreto, 10,18%) y su dispersión también es máxima. De ellas solo las 10 más frecuentes suponen más del 5% del corpus. La tabla 3 lista las 30 palabras léxicas más frecuentes, con la siguiente información: orden de frecuencia (Orden), lema (Lema), categoría (Cat.; V, verbo; S, sustantivo; y A, adjetivo), frecuencia normalizada por mil (Frec./mil), porcentaje de cada ítem respecto a la totalidad del corpus (% del corpus), rango de frecuencia (RF), rango de dispersión lineal (RD) y rango de uso (RU). Además, marcamos en negrita los términos registrados en el *DTM* (RANM, 2011).

En segundo lugar, se abordan los resultados de las ocurrencias por rangos de frecuencia (tabla 4) y dispersión (tabla 5) de los lemas de cada categoría.

Rango frecuencia	1	2	3	4	5	6	7	8	9	Total
Sustantivo	**723**	**1365**	**755**	**431**	196	45	11			3.526
Adjetivo	**398**	**675**	**330**	172	70	12				1.657
Verbo	**175**	**371**	**260**	**188**	93	28	3	**3**	**1**	1.122
Adverbios *-mente*	**39**	**78**	**36**	18	2					173
Epónimos	5	5	2							12
Total	**1.340**	**2.494**	**1.383**	**809**	361	85	14	3	1	6.490

Tabla 4. Tipos de lemas (types) correspondientes a cada categoría léxica en cada rango de frecuencia.

Rango dispersión	1	2	3	4	5	6	7	8	9	10	Total
Sustantivo	**609**	142	168	280	**408**	**459**	**595**	**505**	284	76	3.526
Adjetivo	**233**	60	91	121	177	**222**	**270**	**287**	147	49	1.657
Verbo	70	33	35	59	96	140	184	**233**	**209**	63	1.122
Adverbios *-mente*	5	9	9	11	22	24	**42**	35	16		173
Epónimos	7		1		3	1					12
Total	**924**	244	304	471	**706**	**846**	**1.091**	**1.060**	**656**	188	6.490

Tabla 5. Tipos de lemas (types) correspondientes a cada categoría léxica en cada rango de dispersión.

Se puede observar que solo aparece un subconjunto restringido de sustantivos y verbos en los rangos de frecuencia más altos (7, 8 y 9). En estos, los lemas verbales corresponden a vocablos médicos, sino al léxico general (por ejemplo, *ser, poder* o *tener*). Los adjetivos y los adverbios en *–mente* restringen su aparición a rangos intermedios de frecuencia (aunque abundan

en los más bajos) y de alta dispersión. Respecto a la dispersión, la mayoría de las categorías léxicas posee valores altos, a excepción de los epónimos, con una escasa frecuencia y dispersión. Esto es, existe un grueso del vocabulario médico compartido por la mayoría de los textos. Con todo, la mayor riqueza y variedad de lemas se registra en los rangos de frecuencia bajos (1, 2, 3 y 4). Por otro lado, abundan los lemas en el rango más bajo de dispersión (lo que indica que multitud de vocablos solo aparecen en ciertas secciones), y en los rangos intermedios altos. La tabla 6 relaciona la cifra de lemas en cada rango de frecuencia y de dispersión. Como se explicó, dicha relación queda sintetizada con el valor del uso (U = F x D).

Disp. / Frec.	D:1	D:2	D:3	D:4	D:5	D:6	D:7	D:8	D:9	D:10	Total
F:1	506	108	92	92	98	81	116	88	70	89	**1.340**
F:2	367	122	180	290	463	470	375	168	49	10	**2.494**
F:3	45	13	26	75	127	250	448	373	26		**1.383**
F:4	6	1	5	13	15	41	137	363	228		809
F:5			1	1	3	4	15	67	249	21	361
F:6								1	32	52	85
F:7									2	12	14
F:8										3	3
F:9										1	1
Total	**924**	244	304	471	706	**846**	**1.091**	**1.060**	656	188	6.490

Tabla 6. Tipos de lemas diferentes (*types*) en cada rango de frecuencia por cada rango de dispersión.

Los resultados muestran las tendencias esbozadas. Por un lado, la mayoría de los lemas presentan un rango de frecuencia baja (1, 2 y 3) y una dispersión intermedia alta (6, 7, y 8). Estos casos son tanto términos médicos (*enrojecimiento* [F:3; D:8] o *epilepsia* [F:3; D:7]), como palabras ajenas al dominio (*leyenda* [F:2; D:7] o *adherir* [F:3; D:8]). Se trata de vocablos determinados por los contenidos de cada texto. Por otro lado, muy pocos lemas presentan rangos altos de frecuencia y dispersión. En los rangos extremos (frecuencia 8 y 9, y dispersión 10) aparecen los verbos *tener* [F:8; D:10], *haber* [F:8; D:10], *poder* [F:8; D:10] y *ser* [F:9; D:10], que no son palabras del dominio. En cambio, en los rangos de frecuencia 7 y 8, y 10 de dispersión, ya aparecen los términos médicos (*tratamiento* [F:7; D:10] o *enfermedad* [F:7; D:10]). A continuación explicamos por categorías los resultados, que se desglosan en el apéndice.[4]

4. Discusión

En nuestro análisis, observamos que no existe correlación alta entre elementos muy frecuentes y muy dispersos (coeficiente de correlación = 0.03). Creemos que este resultado es esperable y no se relaciona con el tipo de textos del corpus, pero se debería confirmar esta intuición con otro corpus médico más heterogéneo (compuesto por textos técnicos, divulgativos y de otra tipología). Por este motivo, es importante ofrecer, como en este trabajo, el valor sintético del uso a partir de la frecuencia y la dispersión. No obstante, dado que el uso es producto de la frecuencia por la dispersión, sí existe relación entre ambos valores. Esto es, las palabras más frecuentes presentan alto uso, y viceversa, las palabras menos frecuentes presentan bajo uso.

Respecto a los sustantivos, destacan los propios del dominio médico en los rangos de frecuencia y dispersión más altos: por ejemplo, *tratamiento* (F:7; D:10), *enfermedad* (F:7; D:10), *medicamento* (F:7; D:10), *paciente* (F:7; D:9) o *médico* (F:7; D:10). En estos rangos, asimismo, se registran vocablos polisémicos que solo pertenecen al dominio en ciertas acepciones: por ejemplo, *caso* (F:7; D:10), *problema* (F:7; D:10) o *riesgo* (F:7; D:10). Entre los vocablos de frecuencia media y dispersión alta, se registran términos que designan entidades de actualidad médica o de amplia presencia social (de ahí el hecho de que aparezcan repartidos por todo el corpus): por ejemplo, *colesterol* (F:6; D:9), *cáncer* (F:6; D:8), *estrés* (F:5; D:8), *hipertensión* (F:5; D:8), *depresión* (F:4; D:8) o *diabetes* (F:4; D:9). No obstante, la mayor diversidad de lemas nominales se registra en los rangos de frecuencia más bajos y en todos los rangos de dispersión (especialmente, los más altos). Un conjunto de estos lemas expresa especialidades médicas y entidades propias de disciplinas cuyos límites se solapan con la medicina: *ozono* (F:4; D:4), *pediatra* (F:3; D:7), *tabaquismo* (F:3; D:7), *logopedia* (F:1; D:1). El resto no se asocian necesariamente al dominio y dependen del tema de cada texto; por ejemplo, *cuenta* (F:5; D:10) o *atención* (F:5; D:9). La variedad de lemas nominales que poseen una frecuencia y dispersión baja podría variar en otro corpus médico formado por textos de diferente tipología, pero suponemos que los sustantivos que aparecerían en los rangos más altos serían semejantes a los aquí obtenidos.

Los verbos que aparecen en los puestos de frecuencia y dispersión más alta, sin embargo, carecen de contenido léxico del dominio. La razón es que se trata de verbos auxiliares o cuyas propiedades semánticas son más próximas

a las del léxico instrumental. *Ser* (F:9; D:10) ocupa los rangos más altos, y le siguen *haber* (F:8; D:10), *poder* (F:8; D:10), *tener* (F:8; D:10), *estar* (F:7; D:10), *deber* (F:7; D:10) y *hacer* (F:7; D:10). Los verbos específicos o distintivos del dominio, denominados verbos término (Lorente Casafont, 2002), se empiezan a registrar puntualmente en el rango de frecuencia 5, y proliferan en los rangos inferiores (si bien aparecen en todos los rangos de dispersión): por ejemplo, *vacunar* (F:4; D:6) o *diagnosticar* (F:5; D:9). El resto de verbos que no se relacionan con el dominio médico se registra en todos los rangos de frecuencia y dispersión: por ejemplo, *fluctuar* (F:1; D:4) o *imaginar* (F:2; D:7).

El caso de los adjetivos es semejante al de los verbos. En los rangos intermedios de frecuencia (5 y 6) y de dispersión alta (8, 9 y 10) aparecen lemas no estrictamente relacionados con el dominio médico, que podrían también registrarse en estudios semejantes con otros textos: *mismo* (F:6; D:10), *posible* (F:6; D:10) o *bueno* (F:6; D:10). En estos mismos rangos, aparecen adjetivos relacionales que no son propiamente del dominio médico, pero presentan una alta frecuencia de aparición en términos médicos compuestos: *secundario* (F:5; D:9) forma parte de términos como *efecto secundario* o *cáncer secundario*; *general* (F:5; D:9) aparece en *anestesia general* o *cirugía general*; y *común* (F:5; D:9) se incluye en *acné común* o *migraña común*. Igualmente, en estos rangos también figuran adjetivos del dominio (*médico* [F:6; D:10], *sanitario* [F:5; D:9]) y otros polisémicos (*grave* (F:6; D:9]). Sin embargo, la mayor parte de adjetivos cuya carga léxica está directamente relacionada con la medicina se concentra en rangos de frecuencia inferior a 5, y en todos los rangos de dispersión. Por ejemplo, en los rangos de frecuencia 4 y 5, aparecen adjetivos relacionales que se asocian a las partes de la anatomía o las funciones fisiológicas generales como *cardíaco* (F:5; D:8), *respiratorio* (F:5; D:9) o *renal* (F:5; D:8). En los rangos más bajos (1, 2 y 3) también empiezan a aparecer los adjetivos que designan sentidos anatómicos específicos o propios de ciertas especialidades como *isquémico* (F:2; D:6) o *hematopoyético* (F:3; D:1). Por último, los adjetivos no asociados con la medicina abundan en todos los rangos de frecuencia y dispersión, por ejemplo, *concluyente* (F:3; D:8) o *ruidoso* (F:2; D:2).

Los adverbios en *–mente* se concentran en los rangos más bajos de frecuencia y en los intermedios o altos de dispersión. Casi ninguno posee una carga léxica relacionada con la medicina (por ello no los incluimos en las listas del apéndice). Las excepciones son algunos adverbios de baja frecuencia como *físicamente* (F:2; D:7), *genéticamente* (F:2; D:4), *científicamente* (F:2; D:7), *quirúrgicamente* (F:2; D:7), *médicamente* (F:1; D:6) y *sexualmente* (F:2; D:1). En

estos casos el adverbio expresa el asunto o la perspectiva desde la que se aborda la información. Un estudio que replique este análisis con textos médicos de otras especialidades posiblemente obtendría resultados similares.

Por último, los epónimos muestran una frecuencia baja y una dispersión intermedia o baja en el corpus. Los contenidos de los textos determinan aquellos que aparecen con más frecuencia y dispersión (*Alzheimer* [F:3; D:5], *Crohn* [F:3; D:5]) frente a los de aparición puntual (*Norwalk* [F:1; D:1], *Kegel* [F:1; D:1]).

A modo de síntesis, se pueden distinguir dos tipos de léxico de contenido (por oposición al léxico gramatical; Tabla 7) según sus polaridades de frecuencia (Ueda, 2013).

	Alta frecuencia	Baja frecuencia
Léxico de función	Vocablos gramaticales *a*, *de*, *el*, *un*	Vocablos instrumentales *yo*, *cualquiera*
Léxico de contenido	Vocablos comunes *tener*, *grave*, *tratamiento*, *sanitario*, *enfermedad*	Vocablos específicos *Alzheimer*, *infarto*, *extirpar*, *hematopoyético*

Tabla 7. Tipo de léxico y polaridades de frecuencia.

En el léxico de contenido se pueden diferenciar los vocablos comunes, de alta frecuencia de aparición, y los vocablos instrumentales, de baja frecuencia. A efectos del análisis de textos médicos, los vocablos comunes no poseen sentidos del dominio en cuestión, a excepción de los sustantivos. En cambio, los vocablos específicos sostienen la carga semántica propia del dominio y su variedad y riqueza de lemas es complementaria a su baja frecuencia de aparición. Estas observaciones están en consonancia con los resultados obtenidos por investigadores del procesamiento del lenguaje natural (Jiménez-Salazar, Pinto y Rosso, 2005), quienes han observado que los vocablos situados en un punto de transición entre los términos de alta y baja frecuencia son los que mejor caracterizan los contenidos textuales (Luhn, 1958).

La gráfica de dispersión incluida más abajo (figura 5) sintetiza visualmente lo expuesto.[5] En el eje horizontal se marca con una línea cada una de las ocurrencias de determinados tipos de lemas en el corpus. El artículo *el* (léxico de función gramatical) y el verbo *tener* (léxico de contenido común) presentan una dispersión altísima en el corpus (*el*, [F:10; D:10]; *tener*, [F:8; D:10]), mientras que el adjetivo *cualquier* (vocablo instrumental) muestra dispersión alta y frecuencia menor (F:5; D:10). Vocablos comunes de

contenido (*tratamiento*, [F:7; D:10]; *enfermedad*, [F:7; D:10]; *sanitario* [F:5; D:9] o *grave*, [F:6; D:9]) tienen también alta frecuencia y dispersión, mientras que vocablos específicos solo aparecen en determinadas secciones, como por ejemplo, *infarto* (F:5; D:7) o *Alzheimer* (F:3; D:5).

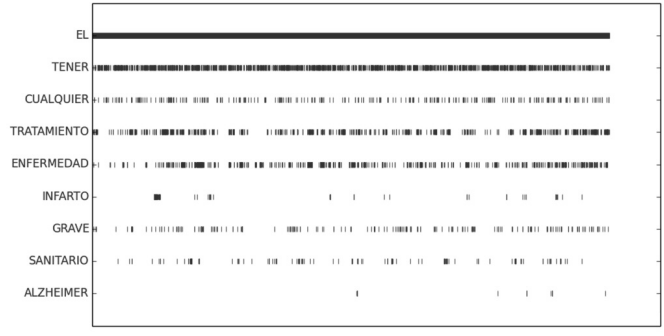

Figura 5. Dispersión léxica de determinados lemas en el corpus.

Las implicaciones de nuestros datos tienen carácter interdisciplinar. Si bien los resultados se restringen a la variedad del español peninsular, y un estudio más completo ha de aglutinar datos de otras variantes, se trata de una aportación a un área huérfana y con potenciales intereses de investigación, creación lexicográfica o aplicación a la ingeniería lingüística. En el apéndice ofrecemos un listado de lemas de las categorías léxicas más usadas (salvo los adverbios en *–mente*, cuyo contenido léxico apenas está relacionado con el dominio) y adjuntamos una selección de los sustantivos, verbos y adjetivos de rango de uso medio (4, 5 y 6). Además de reflejar el uso del léxico en textos divulgativos, esta lista de palabras tiene interés sociolingüístico porque registra la presencia de ciertas enfermedades en la sociedad actual.

5. Conclusiones

Este trabajo ha ofrecido un estudio lexicométrico de las categorías léxicas (sustantivos, adjetivos, verbos, adverbios en *-mente* y epónimos) a partir del análisis de frecuencia y dispersión de 363 textos divulgativos de dominio médico en español. Los resultados están en la línea de otros trabajos y de las aportaciones de otras disciplinas y reafirman la distinción entre un léxico de contenido común y un léxico de contenido instrumental.

La investigación cubre un hueco en los estudios lingüísticos y puede ser de utilidad para los profesionales de la lingüística aplicada. Los listados pueden ayudar a los docentes de español con fines específicos para elaborar el vocabulario médico de su plan de estudios. Los lexicógrafos pueden considerar los resultados de nuestros textos para enriquecer las entradas de los diccionarios especializados. La cuantificación del uso de lemas y categorías también será útil para el procesamiento del lenguaje en este dominio. Como trabajo futuro, sería beneficioso ampliar los datos complementando los textos divulgativos con otros técnicos (por ejemplo, manuales, monografías o artículos científicos). Se abren, pues, vías potenciales para seguir investigando.

Agradecimientos

Este trabajo fue desarrollado en el proyecto "Análisis de datos lingüísticos del español basado en recursos en formato electrónico", financiado por el Banco Santander y el Ministerio de Educación, Deportes y Ciencias de Japón (código de subvención: 20520372). Los autores agradecen la ayuda de Carlos Herrero en la revisión del etiquetado morfológico, y los comentarios de los revisores anónimos para mejorar el artículo.

Historia del artículo:
Recibido 8 abril 2013
Artículo revisado recibido 18 marzo 2014
Aceptado 25 marzo 2014

Bibliografía

Alameda, J.R. y F. Cuetos (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo: Servicio de Publicaciones de la Universidad de Oviedo.

Alcaraz Ariza, Mª.A. (2002). "Los epónimos en medicina". *Ibérica, Journal of the European Association of Languages for Specific Purposes* 4: 55-73.

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

Alvar Ezquerra, M. (2005). "La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera" en Mª. A. Castillo, O. Cruz, J. M. García y J. P. Mora (coords.), *Actas del XV Congreso Internacional de ASELE*, 19-33. Sevilla: Universidad de Sevilla.

Ávila Martín, Mª.C. (2010). "Estadística y lingüística de corpus: Implicaciones pedagógicas en la enseñanza y el aprendizaje del léxico". *Cauce. Revista Internacional de Filología, Comunicación y sus Didácticas* 33: 163-175.

Baayen, H. R. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Bird, S., E. Klein y E. Loper (2009). *Natural Language Processing with Python*. Sebastopol: O'Reilly.

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

Cabré, Mª.T. (2004). "¿Lenguajes especializados o lenguajes para propósitos específicos?" en A. Hooft Comajuncosas (ed.), *Textos y discursos de especialidad: El español de los negocios*, 19-34. Ámsterdam: Rodopi.

Cantos, P. (2013). *Statistical Methods in Language and Linguistic Research*. Oakville: Equinox.

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 Behavior* 3: 61-65.

Davies, M. (2006). *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*. Nueva York: Routledge.

Friedman, C., P.O. Anderson, J.H.M. Austin, J.J. Cimino y S.B. Johnson (1994). "A general natural-language text processor for clinical radiology". *Journal of the American Medical Informatics Association* 1: 161-174.

García Hoz, V. (1953). *Vocabulario usual, vocabulario común y vocabulario fundamental*. Madrid: C.S.I.C.

Gries, S.T. (2008). "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics* 13,4: 403-437.

Herdan, G. (1964). *Quantitative Linguistics*. Londres: Butterworth.

Hernando, B.M. (1990). "Lexicometría del lenguaje periodístico". *Lingüística Española Actual* 12(2): 215-242.

Hripcsak, G. y A.S. Rothschild (2005) "Agreement, the F measure, and reliability in information retrieval". *Journal of the American Medical Association* 12: 296-298.

Jiménez Salazar, H., D. Pinto y P. Rosso (2005). "Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos". *Procesamiento del Lenguaje Natural* 35: 383-390.

Juilland, A. y E. Chang-Rodríguez (1964). *Frequency Dictionary of Spanish Words*. La Haya/París: Mouton.

Lorente Casafont, M. (2002). "Verbos y discurso especializado". *Estudios de Lingüística del Español (ELiEs)* 16. URL: http://elies.rediris.es/elies16/Lorente.html [03/07/2014]

Luhn, H.P. (1958). "The automatic creation of literature abstracts". *IBM Journal* 2,2: 159-165.

Keniston, H. (1920). "Common words in Spanish". *Hispania* 3: 85-118.

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

Moreno Sandoval, A. y J.Mª. Guirao Miras (2006). "Morpho-syntactic tagging of the Spanish C-ORAL-ROM Corpus: Methodology, tools and evaluation" en Y. Kawaguchi, S. Zaima y T. Takagaki (eds.), *Spoken Language Corpus and Linguistic Informatics*, 199-218. Amsterdam: John Benjamins.

Moreno Sandoval, A. y L. Campillos Llanos (2013). "Design and annotation of MultiMedica. A multilingual text corpus of the biomedical domain". *Procedia - Social and Behavioral Sciences* 95: 33-39.

Mungra, P. y T. Canziani (2013). "Lexicographic studies in medicine". *Ibérica, Journal of the European Association of Languages for Specific Purposes* 25: 39-62.

Navarro, F.A. (1997). *Traducción y lenguaje en medicina*. Barcelona: Ediciones Doyma, Fundación Dr. Antonio Esteve.

Navarro, F.A. (2005). *Diccionario crítico de dudas inglés-español de medicina*. 2ª edición revisada. Madrid: McGraw-Hill Interamericana, D.L.

Patterson, W. y H. Urrutibéheity (1975). *The Lexical Structure of Spanish*. La Haya/París: Mouton.

Porta Zamorano, J. y R. Ureña Ruiz (2003). "Lexicometría de corpus". *Procesamiento del Lenguaje Natural* 31: 332-333.

Real Academia Española (RAE) (2009). *Nueva gramática de la lengua española*. 2 volúmenes. Madrid: Espasa.

Real Academia Nacional de Medicina (RANM) (2011). *Diccionario de términos médicos*. Madrid: Panamericana.

Rodríguez Bou, L. et al. (1952). *Recuento de vocabulario español*, 2 volúmenes. San Juan de Puerto Rico: Editorial Universidad de Puerto Rico.

Rodríguez Díez, B. (1979). "Lo específico de los lenguajes científico técnicos". *Archivum* 27-28: 485-521.

Sabaj, O. (2004). "Especificidad, especialización y variabilidad verbal: Una aproximación computacional en estadística léxica". *Revista Signos* 37(56): 75-89.

Sebastián, N., F. Cuetos, M.A. Martí y M.F. Carreiras (2000). *LEXESP: Léxico informatizado*

del español. Barcelona: Ediciones Universidad de Barcelona.

Skelton, J.R. y J. Whetstone (2012). "English for Medical Purposes and Academic Medicine: Looking for common ground". *Ibérica, Journal of the European Association of Languages for Specific Purposes* 24: 87-102.

Sommers, H. (2000). "Machine translation" en R. Dale, H. Moisl y H. Somers (eds.), *Handbook of Natural Language Processing*, 329-346. Nueva York: Marcel Dekker.

Temnikova, I.P. y K.B. Cohen (2013). "Recognizing sublanguages in scientific journal articles through closure properties" en *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*, Sofía, Bulgaria, 4-9 de agosto de 2013, 72 79. Association for Computational Linguistics.

Ueda, H. (1987). *Frecuencia y dispersión del vocabulario español*. Tokio: Universidad de Estudios Extranjeros de Tokio.

Ueda, H. (1989 1990). *Estudio cuantitativo del léxico español*. 2 tomos. Tokio: Publicaciones del Dpto. de Idiomas Extranjeros, Facultad de Artes y Ciencias, Universidad de Tokio.

Ueda, H. (2013). "Analizador lingüístico común con parámetros de gramática, diccionario y cadenas de aplicación" en *V Congreso Internacional de Lingüística de Corpus 2013 (CILC2013)*. Alicante, Universidad de Alicante, 15 de marzo de 2013.

Verdaguer, I., N.J. Laso y D. Salazar (eds.) (2013). *Biomedical English. A Corpus-based Approach*. Ámsterdam: John Benjamins.

Vila Barbosa, Mª.M. (2013). "Corpus especializados como recurso para la traducción". *Onomázein* 27: 78-100.

Zipf, O. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley.

Leonardo Campillos Llanos es investigador asociado al Laboratorio de Lingüística de la Universidad Autónoma de Madrid (España). Ha participado en conferencias internacionales y ha realizado publicaciones relacionadas con la lingüística informática y de corpus, la terminología médica y el análisis de la interlengua oral del español.

Hiroto Ueda es catedrático de la Universidad de Tokio (Japón) y doctor en Filología Española por la Universidad de Alcalá de Henares. Miembro de la Sociedad Japonesa de Hispanistas y de la Academia Norteamericana de la Lengua Española, es especialista en variación léxica del español y autor de numerosas publicaciones, manuales y programas informáticos.

NOTAS

[1] No obstante, Almela et al. (2005) restringieron su estudio a 2 millones de palabras etiquetadas y revisadas a mano, para asegurar la calidad y el rigor.

[2] La segmentación de palabras contraídas o concatenadas explica la variación entre el número de elementos del corpus (334690) y las unidades que anota GRAMPAL (367482, que es la cifra usada en nuestros cálculos).

[3] Hripcsak y Rothschild (2005) ofrecen detalles de la medida F y el acuerdo entre anotadores.

[4] Las listas están disponibles en: http://www.lllf.uam.es/leonardo/pdf/Tablas-resultados-Frec-disp.pdf.

[5] La gráfica se creó con Python Natural Language Toolkit (NLTK) (Bird, Klein y Loper, 2009). URL: www.nltk.org [18/07/14].

Apéndice - Listas de categorías léxicas

Sustantivos* RU			
7	6		
caso	**agua**	**factor**	**prueba**
enfermedad	**alimento**	**forma**	resultado
medicamento	**año**	**grasa**	**salud**
médico	**cáncer**	**grupo**	**sangre**
paciente	**cantidad**	**infección**	**síntoma**
persona	**causa**	información	**sistema**
problema	**colesterol**	**mujer**	**sustancia**
riesgo	consumo	**niño**	**tiempo**
tratamiento	**día**	**nivel**	**tipo**
	dieta	**organismo**	uso
	dolor	país	vez
	efecto	**parte**	**vida**
	estudio	**producto**	**zona**

*En negrita los términos recogidos en el *DTM*

Tabla 8. Lemas (types) de sustantivos en rangos de uso (RU) superiores.

Sustantivos* RU: 5			
ácido	control	hipertensión	peso
alergia	corazón	hombre	piel
análisis	cuerpo	hora	población
anestesia	dato	hospital	presión
animal	diagnóstico	huevo	reacción
antibiótico	diarrea	intervención	relación
atención	dosis	leche	sal
cálculo	edad	lesión	situación
célula	eficacia	medida	sueño
centro	ejercicio	mes	trastorno
comida	enfermo	muestra	vacuna
complicación	estrés	número	vitamina
concentración	familia	ojo	
contacto	fármaco	operación	
contenido	frecuencia	pérdida	

*Todos se recogen en el *DTM*

Tabla 9. Selección de lemas (types) de sustantivos en rangos de uso (RU) superiores.

Adjetivos* RU: 6		
bueno	grande	mismo
distinto	importante	necesario
frecuente	**mayor**	posible

*En negrita los términos recogidos en el *DTM*

Tabla 10. Lemas (types) de adjetivos en rangos de uso (RU) superiores.

Adjetivos*				
RU: 5				
abdominal	clínico	físico	médico	raro
alimentario	crónico	general	mejor	sanitario
alto	diario	grave	menor	secundario
bajo	digestivo	habitual	natural	suficiente
cardíaco	eficaz	largo	normal	último
cardiovascular	específico	malo	preciso	
RU: 4				
activo	corto	humano	oral	sano
afectado	definitivo	infantil	perjudicial	seco
agudo	diagnóstico	inferior	positivo	sensible
alérgico	dietético	intenso	práctico	serio
animal	efectivo	interior	precoz	sexual
anterior	embarazada	intestinal	preventivo	significativo
arterial	esencial	local	protector	solar
autónomo	excesivo	máximo	psicológico	superior
básico	familiar	medio	químico	tóxico
biliar	farmacéutico	mental	quirúrgico	urinario
blanco	fuerte	mínimo	regular	variable
cerebral	gástrico	muscular	renal	vegetal
cervical	genético	negativo	respiratorio	
científico	graso	nervioso	saludable	
corporal	hormonal	nutricional	sanguíneo	

*Todos se recogen en el *DTM*

Tabla 11. Selección de lemas (types) de adjetivos en rangos de uso (RU) superiores.

Verbos*				
RU				
6	5	4		
indicar	afectar	absorber	estimular	operar
provocar	analizar	adelgazar	experimentar	practicar
reducir	causar	agravar	exponer	prescribir
ver	controlar	aliviar	extraer	recetar
	desarrollar	alterar	fijar	reconocer
	detectar	caracterizar	generar	regular
	diagnosticar	conservar	identificar	remitir
	fumar	contaminar	infectar	respirar
	hablar	convertir	introducir	transmitir
	incluir	cuidar	lavar	variar
	ingerir	curar	medir	
	mejorar	dormir	ocasionar	
	pensar			
	prevenir			
	proteger			
	valorar			

*Todos se recogen en el *DTM*

Tabla 12. Selección de lemas (types) de verbos en rangos de uso (RU) superiores.

Epónimos*			
RU			
3	2	1	0
Alzheimer Crohn	Sjögren Parkinson Reye	Papanicolau Wilson	Tourette Addison Gilles de la Tourette Kegel Norwalk

*Todos, excepto Kegel, se recogen en el *DTM*

Tabla 13. Tipos de lemas (types) de epónimos.