Lingüística
Vol. 40, 2024
ISSN 2079-312X en línea
DOI: 10.5935/2079-312X.20240007

LA DISTRIBUCIÓN GEOGRÁFICA DEL LÉXICO EN ESPAÑOL ACTUAL

A DISTRIBUIÇÃO GEOGRÁFICA DO LÉXICO NO ESPANHOL ATUAL

THE GEOGRAPHICAL DISTRIBUTION OF THE CURRENT SPANISH LEXICON

Guillermo Rojo
Universidade de Santiago de Compostela
guillermo.rojo@usc.es
0000-0002-1771-6561

Resumen
El objetivo fundamental de este trabajo consiste en presentar una visión general de la distribución del léxico en los diferentes países hispánicos. Los datos proceden del Diccionario de Frecuencias Léxicas, incorporado a la versión 1.1 del CORPES, construido sobre 180 millones de formas procedentes de los textos de prensa de este corpus. Los resultados obtenidos, que utilizan también en los índices de dispersión, muestran un alto grado de congruencia, relacionada con la frecuencia de uso, compatible con la esperable diferencia entre las variedades, fundamentada en la frecuencia de inventario.

Palabras clave: Léxico; análisis estadístico; frecuencia; dispersión; variabilidad diatópica.

Resumo
O objetivo fundamental deste trabalho é apresentar uma visão geral da distribuição do léxico nos diferentes países hispânicos. Os dados são provenientes do Dicionário de Frecuencias Léxicas, incorporado à versão 1.1 do CORPES, construído a partir dos 180 milhões de formas dos textos de imprensa deste corpus. Os resultados obtidos, que também se baseiam nos índices de dispersão, apresentam um elevado grau de congruência, com base na frequência de uso, compatível com a diferença esperada entre as variedades, com base na frequência de inventário.

Palavras-chave: Léxico; análise estatística; frequência; dispersão; variabilidade diatópica.

Abstract
The primary objective of this work is to present a general vision of the distribution of the lexicon in the different Hispanic countries. The data come from the Diccionario de Frecuencias Léxicas, incorporated into version 1.1 of CORPES, built on 180 million words from the press texts of this corpus. The results obtained, which are also based on the dispersion indices, show a high degree of congruence, based on the frequency of use (type frequency), compatible with the expected difference between the varieties, based on the frequency of inventory (token frequency).

Keywords: Lexicon; statistical analysis; frequency; dispersion; diatopic variability.

Recibido: 16/07/2024
Aceptado: 27/08/2024

1. Introducción: la distribución del léxico del español

Entre los estudios dedicados a las peculiaridades del léxico español en los diferentes países hispánicos podemos identificar tres grandes líneas generales de acceso. Por una parte, los léxicos específicos de algunas de las grandes ciudades del mundo hispánico, insertos en el Proyecto de Estudio Coordinado de la Norma Lingüística Culta del Español Hablado en las Principales Ciudades de Iberoamérica y de la Península Ibérica[1]. Son trabajos hechos sobre cuestionario, al estilo habitual en los atlas lingüísticos, y con los que, por tanto, es posible, aunque muy costoso, comparar, elemento a elemento, las diferentes palabras utilizadas para el mismo significado en las distintas ciudades, como se hace, por ejemplo, en Moreno de Alba (1992: 108 y sigs.). También sobre cuestionario se obtienen las respuestas en el proyecto Variación Léxica en Español del Mundo (Varilex)[2]. En este caso, la publicación agrupa en un cuadro las respuestas a cada pregunta, con lo que es sencillo ver el grado de homogeneidad en las denominaciones de conceptos y objetos en todo el mundo hispánico, pero de nuevo es una visión fragmentaria que se obtiene ítem a ítem. En tercer lugar tenemos las abundantes descripciones de detalle acerca de palabras utilizadas solo en alguno(s) de los países o bien de acepciones específicas de palabras de uso más general, al estilo de lo que se muestra en, por ejemplo, Moreno de Alba (1993: cap. VII), o las secciones correspondientes de los diferentes capítulos de Lipsky (1994) o Haensch (2003).

En los estudios realizados en cualquiera de estas tres líneas prima la búsqueda de lo diferencial, de los usos propios de un determinado país o ciudad que pueden producir extrañeza o confusión en hablantes procedentes

[1] Para un resumen de las características, implicaciones y logros de esta parte del proyecto, *cf*. Caravedo (2000).
[2] Para detalles y resultados, vid. la página web del proyecto: https://h-ueda.sakura.ne.jp/varilex-r/

de otras zonas. No es posible obtener de ellos una visión global acerca de la distribución del léxico español y su grado de homogeneidad o heterogeneidad. Dada la finalidad perseguida, es casi imposible evitar la sensación de que las diferencias léxicas entre las variedades inter-nacionales o intra-nacionales son, además de abundantes, muy marcadas, de modo que la interpretación y comprensión de textos producidos en variedades distintas de la propia parece ser tarea que presenta ciertas dificultades[3].

En sentido contrario apuntan los enfoques que, sin negar el carácter policéntrico del español, insisten en su carácter fuertemente unitario, gracias al cual es posible la comprensión entre hablantes de diferentes variedades con solo algunas dificultades que pueden ser salvadas con bastante facilidad. Los trabajos que tienen esta orientación y pretenden cuantificar el grado de homogeneidad del léxico español dan un resultado muy claro: el léxico español tiene un porcentaje altísimo de palabras de uso general y que, por tanto, son fácilmente entendidas por hablantes de cultura media. En el contexto del proyecto Difusión Internacional del Español por Radio, Televisión y Prensa (DIES-RTP), Raúl Ávila (2000) analizó textos procedentes de varios programas de noticias de radio y televisión dirigidos a un público amplio, no reducido a un único país.

En total, obtuvo un corpus formado por 76 300 palabras gráficas. Contrastó los lemas obtenidos con el Diccionario de la Real Academia Española (DRAE, entonces en su edición de 1992) y algunos otros diccionarios de ámbito general. En los documentados, diferencia entre "a) los de uso general hispánico, no marcados o *sin filiación*; y b) los marcados o *con filiación*" (Ávila, 2000: 44). Caracteriza luego los del segundo grupo (a los que llama "ismos") en extranjerismos, americanismos, latinismos, etc. El resultado es concluyente (Ávila, 2000: 46 y cuadro 2): la suma de "ismos" y no documentados se sitúa en una horquilla que va del 0,51 % (Televisa, México) al 1,24 % (CNN en español).

Muy poco tiempo después, María Antonieta Andión Herrero (2003), interesada por conocer el posible impacto de la prensa digital en la lengua española, reunió un conjunto de 20 artículos procedentes de cinco periódicos digitales de diferentes países y construyó un corpus de un total de unas 12 000 formas gráficas. Entre las 7000 "nocionales" localizadas trató de identificar las que son de "uso distintivo, ya sea porque pueden considerarse desconocidas en un español estándar o porque son poco frecuentes o no se prefieren para expresar un contenido o significado" (Andión Herrero, 2003: 74). El resultado es también muy claro: únicamente "un 1,19 % se considera léxico distintivo usado por las variedades del español que lo producen" (*ibíd*.: 89). El criterio utilizado para la caracterización es, básicamente, el conocimiento lingüístico de la autora del estudio.[4]

[3] Y que, con un planteamiento aplicado a otro terreno, constituye un problema que pretende solucionar la línea del llamado "español neutro", utilizado, por ejemplo, en producciones televisivas destinadas a ser difundidas en varios países a uno y otro lado del Atlántico. Para una visión general del problema y de las cuestiones asociadas a la globalización del español, *vid.* López Morales (2006) y Moreno Fernández (2016).

[4] Razón por la cual de la muestra quedan excluidas noticias procedentes de Cuba (Andión Herrero, 2003: 73).

Por esta misma época, María Josefina Tejera agrupó los envíos de noviembre de 2001 del "resumen noticioso" que la CNN elaboraba a partir de las noticias distribuidas en el canal y enviaba por correo electrónico a los suscriptores de este servicio. Pretendía diferenciar los usos específicamente americanos de los generales. El resultado obtenido es que una proporción muy elevada del léxico utilizado pertenece al conocimiento común y que las peculiaridades aquí recogidas "constituyen un número pequeño que proporcionalmente alcanza sólo un 0,066 %" (Tejera, 2003: 871)[5].

Los resultados obtenidos en estos tres estudios son muy similares, como resalta López Morales (2006: 77 y sigs.), y muestran un idioma que posee un altísimo grado de homogeneidad. Naturalmente, no se puede olvidar que todos ellos han sido elaborados sobre textos periodísticos, lo cual implica un cierto nivel de lengua. Un segundo aspecto que hay que tener en cuenta y que, como veremos, resulta de gran importancia es el escaso tamaño de los corpus manejados. Tejera no da el volumen del suyo, pero indica que ha trabajado con los resúmenes de 30 días, con una media de 20 noticias diarias, lo cual nos lleva a una estimación que difícilmente puede superar las 200 000 palabras. Son tamaños reducidos, admisibles para realizar un análisis detallado de palabras y acepciones como el que practican estos tres autores, pero sin duda insuficientes para tratar de obtener una visión global acerca de la distribución del léxico del español.

Uno de los productos derivados del Corpus del Español del Siglo XXI (CORPES) es el diccionario de frecuencias léxicas que se ha incorporado a la versión 1.1. Se trata de un recurso orientado en la línea de los diccionarios de frecuencias dinámicos (DF) descritos en Rojo (2023) que ha sido construido sobre el subconjunto de los textos de prensa contenidos en la versión 1.0 del CORPES. Esta aparente limitación tiene, a mi modo de ver, varias ventajas para un objetivo como el perseguido. Todo lo que sabemos acerca de la distribución del léxico en los corpus textuales muestra una marcada dependencia con relación a las características de los textos incluidos.

Como se muestra en Rojo (2023) son muchos los lemas que se documentan únicamente en unos pocos textos de un corpus (solo en uno con mucha frecuencia), lo cual favorece una dispersión indesligable del volumen y la variedad de textos utilizados. Es cierto que la tradición de los diccionarios de frecuencia se vincula con mucha frecuencia a la utilización de subcorpus construidos sobre textos de diferentes tipos (los "mundos" del *Frequency Dictionary of Spanish* Words (= FDSW, Juilland y Chang Rodríguez, 1964), por ejemplo), pero no es seguro que esa segmentación proporcione una visión adecuada de lo que sucede. Utilizar solo textos de prensa, en cambio, proporciona un conjunto amplio de textos que presentan unas características semejantes y que, además, tratan la misma variedad de temas, con lo que la medida de la homogeneidad o discrepancia del léxico resulta mucho más próxima a lo que sucede en la lengua cotidiana en textos accesibles a un número importante de lectores.

[5] El recuento se hace sin tomar en consideración "artículos, preposiciones, conjunciones, interjecciones, adverbios, nombres propios, nombres de cargos, numerales, y nombres de meses y días de la semana, además de los nombres de las monedas y los puntos cardinales" (Tejera, 2003: 867).

Decidí, por tanto, utilizar los datos que están en la base del DF del CORPES para tratar de obtener una visión panorámica del léxico español. La gran ventaja de emplear un recurso construido sobre el subcorpus de prensa del CORPES procede del importante volumen de textos que se puede analizar (hasta cierto grado) de forma automática, puesto que el material de entrada está constituido por los elementos que han sido etiquetados y lematizados para su integración en el corpus. Este carácter permite también filtrar los elementos que resultan inadecuados para un propósito de este tipo: han sido eliminados automáticamente nombres propios, cifras, fechas y algunos otros elementos habitualmente excluidos de los estudios generales de frecuencias. Después de todas estas operaciones, el subcorpus obtenido se sitúa en torno a los 180 millones de elementos, cifra que supone un aumento de varios órdenes de magnitud con relación a las utilizadas anteriormente. En este recurso figuran todos los países hispánicos salvo Guinea Ecuatorial y Filipinas, que no tienen en el CORPES textos de prensa con el volumen mínimo requerido para que las comparaciones sean fiables. La distribución, que respeta la utilizada en general para todo el CORPES, puede ser consultada en la *Guía* del recurso[6].

Naturalmente, no todo son ventajas: el empleo de un corpus de este volumen supone que el análisis ha de ser forzosamente automático y, por tanto, tiene que estar reducido al estudio de la distribución de los lemas (con inclusión de la clase de palabras), sin que se pueda entrar en el análisis de las posibles diferencias entre acepciones del mismo lema. En las páginas siguientes me propongo, tras presentar algunos conceptos generales, analizar los resultados más sobresalientes obtenidos de esta primera visión general del léxico hispánico.

2. La frecuencia

2.1. Generalidades

La frecuencia es, simplemente, el número de veces que un elemento lingüístico aparece en un texto. Puede tratarse de una letra, un fonema, una forma gráfica, un morfema, un lema, un elemento gramatical, cierta clase de palabras, una estructura sintáctica, etc. En lo que sigue nos centraremos en lo que sucede con los lemas. Es importante diferenciar entre la frecuencia general o absoluta (FG) y la frecuencia relativa o normalizada (FN). La FG es la que aparece en la definición inicial: el número de ocasiones en que se documenta un determinado elemento en un fragmento, un texto o un corpus. Este concepto, tan simple, resulta útil en las primeras aproximaciones, puesto que nos permite establecer una ordenación entre los elementos considerados según cuál sea la posición que ocupan en una lista ordenada, es decir, su rango[7].

Es fácil ver, sin embargo, que la simple consideración de la FG puede impedirnos captar algunas características que, sin duda, son de interés para la consideración de los elementos léxicos.

---

[6] Véase https://www.rae.es/corpes/assets/rae/files/corpes/guiaDiccionariosFrecuenciasLex.pdf

[7] "Rango" puede remitir también al número de subcorpus en que se documenta una expresión determinada.

La versión 1.1 del CORPES proporciona los datos que figuran en el cuadro 1 con respecto a cuatro lemas (sin indicación de clase) próximos en cuanto a su FG:

	FG	FN (cpm)	Núm. docs.
Jugo	8240	20,18	3507
Ilustrar	8226	20,15	5410
Quieto	8370	20,50	3420
Traslado	8260	20,23	5292

Cuadro 1: Frecuencia general, normalizada (cpm) y número de documentos de varios lemas
Fuente: CORPES. Elaboración propia.

Los cuatro lemas tienen una FG similar y, lógicamente, también una FN muy parecida, pero la columna en la que se indica el número de documentos en que aparece cada uno de ellos hace sospechar que hay algo en su distribución que los diferencia: los casos de *jugo* y *quieto* están mucho más concentrados, mientras que los de *ilustrar* y *traslado* aparecen en un número considerablemente mayor de documentos. Parece probable, por tanto, que su distribución responda a factores que tienen importancia en la configuración general del corpus. No es difícil saber qué es lo que sucede. Si observamos el cuadro que en la aplicación de consulta del CORPES proporciona las FG y FN de *jugo* por áreas geográficas encontramos lo que indico el cuadro 2 (solo para algunas de ellas):

Zona	FAbs.	FNorm.
Andina	739	23,3
Antillas	889	35,12
Caribe continental	1508	30,34
Chilena	808	33,58
España	970	6,88
Estados Unidos	130	28,13

Cuadro 2: Frecuencia absoluta y normalizada (casos por millón) de algunos lemas
Fuente: CORPES. Elaboración propia

La FN más alta corresponde a las Antillas, luego vienen Chile, Caribe continental, Estados Unidos, Andina y, por último, España. Lo que sucede es muy claro: la FN pone en relación el número de casos con el volumen del conjunto del que se han sido obtenidos. Los totales de estas zonas son diferentes y, por tanto, la FG da una perspectiva real, pero que no tiene en cuenta el peso en cada uno de los subconjuntos con los que se trabaja (las áreas lingüísticas en este caso).

Lo mismo que sucede con los porcentajes, la utilidad fundamental de la FN radica en que permite la comparación directa de lo que se encuentra en dos o más subconjuntos distintos (en el caso de los elementos léxicos, dos o más países distintos, dos o más tipos de texto, etc.). En estadísticas vinculadas a elementos léxicos, la FG es casi siempre baja, lo cual hace que, en lugar de

hablar de número de casos por cada 100, lo que nos llevaría a tener que referirnos a FN como 0,0002 y similares, se trabaje habitualmente con casos por cada millón (cpm), que es lo que se indica en el cuadro 2. Por tanto, lo que se indica ahí es que *jugo* aparece una media de 35,12 cpm en las Antillas, mientras que se encuentra solo una media de 6,88 cpm en textos producidos en España. Estas cifras producen una impresión totalmente distinta y, ahora sí, ajustada a la realidad: *jugo* es mucho más usada en las Antillas o Chile que en España, puesto que en este último país aparece algo menos de 7 veces por cada millón de palabras, mientras que en las otras dos zonas tiene 30 o más cpm.

Hay otro aspecto en el análisis de las frecuencias que necesitamos tener en cuenta. Es la existente entre las que he propuesto llamar frecuencia de uso (FU) y frecuencia de inventario (FI) (Rojo, 2011), próximas, pero no idénticas, a las que Bybee (2007) denominó *type frequency* y *token frequency*. La más básica es la frecuencia de uso, que es, simplemente, la cantidad de veces que aparece un elemento en un texto o un corpus. Si hablamos de un lema, la FU se obtiene mediante el recuento de los casos de ese lema localizados en el corpus en cuestión. Si elevamos el nivel de abstracción y pretendemos obtener la frecuencia de, por ejemplo, los adjetivos, encontramos una doble posibilidad. De una parte, la suma de las FU de todos los adjetivos presentes en el texto (*tokens*). De otra, el número de adjetivos distintos que podemos documentar en el conjunto tomado en consideración (*types*). En esta segunda perspectiva, que es la FI, no se contrasta la frecuencia de un determinado adjetivo con la de otros lemas (de su misma clase o de otra), sino el número de adjetivos distintos con respecto a, por ejemplo, el número de sustantivos o de verbos distintos o bien el número total de lemas diferentes.

La diferencia entre FU y FI se aplica en todos los niveles del análisis gramatical. En el caso del léxico, esta diferencia es la que se maneja cuando, como vamos a ver en el apartado siguiente, se comprueba que un grupo reducido de lemas distintos puede dar cuenta de un número muy elevado de apariciones: con los 10 lemas más frecuentes, que suponen el 0,0086 % de la FI, podemos dar cuenta del 40 % de la FU.

2.2. Distribución de las frecuencias en los elementos léxicos

La distribución general de los elementos léxicos tiene ciertas peculiaridades conocidas desde hace ya bastante tiempo. Como saben todos los hablantes, algunas palabras aparecen mucho y otras se usan poco o muy poco. George Kingsley Zipf [1902-1950] mostró en los años 40 del siglo pasado que esas diferencias responden a una regularidad general que se conoce actualmente como ley de Zipf. Según esa ley, si el elemento (lema, elemento gramatical o palabra gráfica) más frecuente en un texto aparece *x* veces, el segundo en frecuencia aparecerá *x*/2 veces, el tercero *x*/3 y así sucesivamente. En general, la frecuencia de un lema que ocupa la posición *p* en el listado por rango será *x*/p.

Una visión distinta, pero congruente con la anterior, es la que se describe con la llamada distribución de Pareto (por Vilfredo Pareto [1848-1923]) o distribución del 80/20.

La formulación general de este principio establece que el 80 % de los efectos se debe al 20 % de las causas. Por ejemplo, en las sociedades humanas es habitual que el 80 % de la riqueza pertenezca al 20 % de la población. Las cifras fijadas por Pareto resultan bastante cortas cuando las contrastamos con lo que se observa en los elementos lingüísticos, como vamos a ver a continuación.

La proyección más relevante de la ley de Zipf sobre la distribución de los elementos léxicos se manifiesta en tres aspectos diferentes. En primer lugar, existe un número muy reducido de elementos que presentan frecuencias muy altas y suponen en conjunto un porcentaje muy importante de la totalidad del texto o el corpus considerado. Atkins y Rundell (2008: 59-60) calculan que las 100 palabras más frecuentes del BNC cubren aproximadamente el 45 % del corpus.

Como se puede apreciar en el cuadro 3, en español los 10 lemas más frecuentes de los textos de prensa de la versión 1.0 del CORPES acumulan en conjunto un porcentaje próximo al 40 % de las formas contenidas en este (sub)corpus, a pesar de que suponen únicamente (FI) el 0,0086 % del total de los lemas (116 707). Con los 50 lemas más frecuentes se "explica", es decir, se reconoce o identifica formalmente, algo más del 50 % (FU) del corpus considerado y con los 100 primeros se llega al 57,24 %[8].

En segundo término, hay un gran número de formas —la inmensa mayoría de ellas— que tienen frecuencias bajas o muy bajas. De nuevo con los datos de el cuadro 3, los 5000 lemas comprendidos entre el rango 5000 y el 10 000 del total de los lemas supone un incremento de únicamente 3,45 puntos porcentuales en la FU. Y los 25 000 añadidos en el último tramo del cuadro aportan solo 0,45 puntos en el total de las frecuencias de uso.

Por último, entre los que tienen frecuencia muy baja ocupan un lugar especial los que poseen FG=1, que son los llamados *hápax* siguiendo la denominación usada en la lingüística clásica para los elementos que se documentan solo una vez y, por tanto, suscitan dudas acerca de su carácter. En el corpus que estamos analizando, los hápax constituyen el 25 % del total de los lemas. Esto es, el 25 % de los lemas (no formas gramaticales, no formas ortográficas) documentadas en un corpus formado por más de 180 millones de formas aparece únicamente una vez. Son muy abundantes en cualquier corpus y en cualquier elemento relacionado con el léxico. Nation (2016) estima que el 50 % de las formas ortográficas distintas del inglés tiene frecuencia igual a 1. Según Rojo (2008, 2017), el porcentaje de hápax se sitúa alrededor del 40 % de las formas ortográficas distintas del español y, lo que es más importante, parece independiente del tamaño del corpus analizado. Un importante corolario de este último rasgo es que garantiza la entrada de formas (elementos, lemas) nuevas en cualquier ampliación del corpus (*cf.* Rojo, 2008).

[8] Lo cual no equivale a la comprensión plena de su significado en cada caso. *Cf.* Schmitt *et al.* (2011) para el análisis de los factores implicados en este proceso. Téngase en cuenta también lo que se indica *infra* acerca de la conveniencia de tener en cuenta la clase a la que pertenecen los elementos implicados en la estadística.

Los primeros x lemas	Porcentaje acumulado de uso
10	38,82
25	47,81
50	52,58
100	57,24
500	71,36
1000	78,85
2000	86,24
3000	90,01
4000	92,37
5000	93,86
6000	94,95
7000	95,76
8000	96,39
9000	96,90
10 000	97,31
15 000	98,54
20 000	99,12
25 000	99,44
50 000	99,89
116 707	100,00
Total lemas	116 707
Hápax	29 172 (25 %)

Cuadro 3: Frecuencias de los lemas documentados en los textos de prensa del CORPES 1.0
Fuente: CORPES. Elaboración propia

Una de las aplicaciones más habituales de las listas de frecuencia consiste en utilizarlas para organizar y estructurar la introducción del léxico en los cursos de lengua, en especial de lengua extranjera. Asociado a ese empleo está el cálculo de la cobertura del contenido de los textos, ya mencionado en párrafos precedentes. La cuestión fundamental desde esta perspectiva consiste en hacer estimaciones acerca del tamaño del lemario que se necesita dominar para "entender" un porcentaje aceptable de los textos (un 98 % o un 95 % en las más generalizadas, *cf.,* por ejemplo, Nation, 2006: 61). De los datos que figuran en el cuadro 3 se deduce que en español son necesarios unos 6000 lemas para reconocer el 95 % de un texto y entre 10 000 y 15 000 para llegar al 98 % (*cf*. también Rojo, 2023: cuadro 1). No obstante, como se señala en Rojo (en prensa), estas estimaciones están condicionadas por el peso que artículos, preposiciones, conjunciones y otros elementos de significado básicamente gramatical tienen en los recuentos de frecuencia y deben ser reconsideradas.

3. La dispersión

3.1. Distribución por países

El segundo gran problema que plantea el empleo simple de las frecuencias léxicas para diferentes finalidades es su falta de sensibilidad a la distribución de las apariciones de un lema.

Un lema puede tener una FG de 45 (y la FN que le corresponda según el tamaño del corpus), pero, por ir a los extremos, todos esos casos podrían estar en un único documento o bien aparecer en 45 documentos distintos. La FG y la FN serían las mismas en todos los casos, pero parece claro que la "importancia" de ese elemento es distinta en esas dos situaciones hipotéticas. Una distribución así es extraña, pero se pueden encontrar casos relativamente próximos. En el corpus ESLORA, por ejemplo, *tintorería* aparece 19 veces (FN = 26 cpm), 17 de ellas en la misma entrevista, y algo parecido sucede con *notaría*, que tiene 21 casos (FN = 28) concentrados en dos documentos. En el CORPES, *neurotecoma* tiene una FG de 41 y una FN de 0,29 cpm, con todos los casos en el mismo documento. Dejando a un lado estos ejemplos extremos y anecdóticos, en todos los corpus se producen desajustes que proceden de la presencia de documentos en los que determinadas palabras tienen una frecuencia muy superior a la esperada según los criterios generales.

La forma adecuada de tomar en consideración fenómenos de este tipo y, con una visión más general, valorar del modo correcto la forma en que un elemento se distribuye entre los diversos componentes de un corpus es, precisamente, calcular su dispersión, esto es, encontrar una medida que valore la configuración con que un elemento se documenta a lo largo del corpus. Una forma clara de lograrlo puede ser tomar en cuenta el número de documentos en los que aparece, como se describe en el párrafo anterior para algunos casos extremos. Esta opción, que puede tener cierto interés en un corpus reducido, no resulta de utilidad en un corpus formado por cientos o miles de millones, en los que, además, un "documento" es una noticia periodística de 300 palabras y también un libro de 70 000.

Por otro lado, lo mismo que sucede con la FN, el cálculo de la dispersión puede servir para proporcionar una visión del modo en que los elementos lingüísticos se manifiestan en diferentes corpus o subcorpus: textos escritos y orales, procedentes de distintos países, áreas temáticas diversas, épocas, etc.). El paso previo es, por tanto, decidir qué criterio o criterios van a ser utilizados para establecer los diferentes conjuntos textuales. En el ámbito hispánico, la agrupación más tradicional es la utilizada en el FDSW y los que han seguido esta línea de trabajo (Morales, 1986; Castillo Fadić, 2021). En este modelo, los textos se distribuyen en cinco "mundos": ficción, teatro, ensayo, prensa y textos técnicos, cada uno de los cuales está formado por un total de 100 000 palabras. En el extremo opuesto, algunos índices se calculan a partir de agrupaciones que no responden a unos parámetros comparables y tienen volúmenes muy diferentes, como Davies y Davies (2018), y otros segmentan el corpus en bloques de textos configurados mediante procesos puramente aleatorios o de simple linealidad de los textos.

En el caso del CORPES, que es el que se trata en este trabajo, me ha parecido que la segmentación de mayor interés inicial es la basada en los diferentes países del mundo hispánico, que es, además, la utilizada para la confección del diccionario de frecuencias léxicas elaborado sobre los textos de prensa de la versión 1.0 de este corpus. La primera aproximación, la más general, puede consistir en obtener la cifra de los lemas que están documentados en los 21 países considerados. Para valorar el resultado es preciso tener en cuenta varios factores.

En primer lugar, de los listados manejados han sido excluidos, como ya he indicado en varias ocasiones, nombres propios, cifras, fechas, direcciones electrónicas y algunos otros tipos de elementos que son irrelevantes para esta clase de estadísticas. En segundo término, lo que podemos manejar aquí es, simplemente, la existencia del lema en cuestión en textos procedentes de un determinado país.

No es posible tener en cuenta el significado concreto en cada caso, puesto que la anotación no alcanza a indicar la acepción que le corresponde en el diccionario de referencia (que sería el DLE en este caso). En sentido contrario, la consideración de lema incluye la clase de palabras, lo cual implica que pueden producirse discrepancias aparentes que no reflejan exactamente el conocimiento del significado, como sucede con las palabras que pueden aparecer como sustantivos y también como adjetivos. Por último, la anotación es exclusivamente automática, de modo que hay un cierto porcentaje de errores en la asignación que pueden hacer variar las cifras, aunque esos errores deberían ser sistemáticos y, por tanto, afectar a los datos de todos los países por igual.

De los datos manejados se concluye que de los 116 707 lemas (con clase) registrados, solo 13 487 (el 11,56 %) tienen documentación en los 21 países. La cifra es, sin duda, muy llamativa y haría pensar que el español es una lengua muy fragmentada (al menos, en lo que al léxico se refiere: solo 1 de cada 10 lemas se encuentra en textos de todos los países) si no hubiéramos aludido ya a las características generales de la distribución del léxico, a la enorme cantidad de elementos que tienen una frecuencia baja o muy baja y al porcentaje de hápax que aparece en cualquier conjunto de textos. Un modo que confirmar empíricamente que la realidad no resulta tan fragmentaria como podría pensarse consiste en obtener la frecuencia de uso conjunta de esos lemas documentados en todos los países.

El resultado es, por lo menos, tan llamativo como el anterior: los lemas que se documentan en los 21 países suman en conjunto 180 388 681 palabras del total de 184 108 153 que tiene el corpus considerado. Es decir, suponen el 97,98 % del total. Esto implica que los lemas que no están en todos (el 88,44 %) suman únicamente un 2,02 % de los usos. Como se ve, el principio de Pareto queda muy por debajo de lo que sucede en el léxico. Como confirmación de lo anterior, y solo para excluir casos anecdóticos, puede añadirse que los lemas que se documentan en 19 o más países tienen una frecuencia conjunta de 181 828 244, esto es, suponen el 98,76 % del total, lo cual deja un resto del 1,24 % para todos los demás.

Como es lógico, son los lemas con frecuencias más altas los que tienen mayor tendencia a aparecer en todos los subcorpus.

El cuadro 4 proporciona las cifras básicas para los tramos de frecuencia que ya hemos utilizado en las secciones anteriores.

Los primeros	x	*Número de lemas*	Porcentaje sobre los lemas del tramo
1000		1000	100
2000		2000	100
3000		2998	99,93
4000		3995	99,88
5000		4986	99,72
6000		5978	99,63
7000		6945	99,21
8000		7884	98,55
9000		8791	97,68
10 000		9637	96,37
15 000		12 588	83,92
20 000		13 393	66,97
25 000		13 486	53,94
50 000		13 487	26,97
100 000		13 487	13,49
116 707		13 487	11,56

Cuadro 4: Lemas correspondientes a diferentes tramos de frecuencia documentados en los 21
Fuente: CORPES. Diccionario de frecuencias léxicas. Elaboración propia

Los resultados son, aunque no inesperados, realmente llamativos. El 99,72 % de los 5000 lemas más frecuentes se documenta en todos los subcorpus y, en los 10 000 primeros, el porcentaje desciende solo al 96,37 %. Es decir, únicamente 4 de cada 100 lemas de los que forman el léxico básico del español común, usado para hablar de todos los asuntos importantes de una sociedad no se registran en todos los países.

Si se tiene en cuenta que en la prensa se tratan diariamente los temas más variados, la conclusión forzosa es que el léxico del español presenta un altísimo grado de unidad. Es cierto que de los inventarios se han excluido los nombres propios, pero permanecen, en cambio, los gentilicios, que proporcionan una cierta cantidad de términos con vocación de especificidad de cada país.

Los datos anteriores, obtenidos de una muestra de algo más de 184 millones de formas, confirman los procedentes de las muestras mucho más reducidas estudiadas por Ávila, Andión Herrero y Tejera. De todas formas, la realidad es poliédrica y es aconsejable por tanto enfocar el tema también desde la perspectiva contraria. Los lemas que se documentan únicamente en un país muestra la otra cara de la distribución del léxico. Los datos aparecen en el cuadro 5:

	Lemas totales	Lemas exclusivos	%
Argentina	50 591	3394	6,71
Bolivia	32 369	650	2,01
Chile	45 883	2542	5,54
Colombia	49 882	2892	5,80
Costa Rica	28 700	577	2,01
Cuba	40 985	1955	4,77
Ecuador	35 577	1006	2,83
El Salvador	27 208	409	1,50
España	74 766	13 283	17,77
Estados Unidos	30 463	396	1,30
Guatemala	27 636	463	1,68
Honduras	26 454	432	1,63
M*é*xico	59 574	5937	9,97
Nicaragua	29 067	649	2,23
Panam*á*	24 862	313	1,26
Paraguay	33 346	788	2,36
Per*ú*	39 085	1359	3,48
Puerto Rico	29 955	550	1,84
República Dominicana	33 650	896	2,66
Uruguay	35 518	0	0,00
Venezuela	45 471	2124	4,67
Totales		40 615	

Cuadro 5: Lemas exclusivos por países
Fuente: CORPES. Elaboración propia

Hay un total de 40 615 lemas que solo figuran en uno de los países, lo cual supone el 34,8 % del total de lemas documentados. Es algo más de un tercio de los lemas, pero la suma de sus FU alcanza solo 96 052, equivalente a una frecuencia media de 2,36 por lema. La horquilla de porcentajes de los lemas exclusivos de un país con respecto al total de lemas registrados va desde el 17,77 % que se da en España al rotundo 0 % registrado en Uruguay. En términos generales, el número de lemas exclusivos parece estar relacionado directamente con el número total de lemas de cada país, asociado a su vez al volumen total del subcorpus correspondiente. De todas formas, puede observarse también que hay algunos otros factores que actúan: Chile y Venezuela, con un número semejante de lemas, muestran algo más de un punto porcentual de diferencia. Y, por supuesto, el caso especial de Uruguay, que tiene un número muy similar al de Ecuador, pero un porcentaje bastante distinto de lemas exclusivos.

Para cerrar esta sección podemos ver qué sucede en un punto intermedio, con los lemas que se documentan en 11 o más de los países representados en el corpus. Son 31 127, es decir, el 26,67 % del lemario total. Por tanto, solo algo más de un cuarto de los lemas registrados están presentes en 11 o más de los 21 países, lo cual parece caminar en la línea diferenciadora. Si se toma en cuenta el rango, son el 99,91 % de los 10 000 primeros, el 99,58 % de los 15 000 y el 98,8 % de los 20 000 más frecuentes.

De nuevo la discrepancia entre las cifras generales, que sitúa todos los lemas en el mismo nivel, y la que diferencia según los rangos.

Creo que las cifras anteriores muestran con claridad las dos caras del fenómeno: hay una enorme cantidad de lemas que se documentan en unos pocos países (incluso en uno), mientras que solo 1 de cada 10 aparece en todos ellos. Al tiempo, los términos comunes a todos los países suponen casi el 98 % de los usos, de modo que los elementos diferenciales quedan reducidos a unos porcentajes marginales.

3.2. Índices de dispersión

El análisis de la documentación de los lemas entre los diferentes subcorpus (construidos, en el caso del CORPES, sobre los distintos países) supone una importante mejora con respecto al dato crudo de la frecuencia de uso. La vinculación entre la frecuencia general y la presencia de los lemas en un número importante de países es un factor perfectamente explicable por sus características propias: es evidente que elementos con frecuencias de uso bajas tienen una probabilidad escasa de aparecer en todos los países[9], de modo que esa asociación no debe ser considerada como un inconveniente para valorar positivamente su utilidad. Sin embargo, para entender de forma adecuada la distribución del léxico entre los diferentes países hispánicos no podemos quedarnos aquí. Este factor es totalmente insensible a dos aspectos de gran importancia. Por una parte, no tiene en cuenta el número de casos que se detectan en cada país, de modo que un lema que aparece 100 o más veces en algunos subcorpus y solo una vez en otros tiene la misma consideración que los que no muestran diferencias tan marcadas en su distribución. Por otra, vinculada a la anterior, un lema general, bien distribuido, debe presentar en cada subcorpus un número de casos congruente con lo que ese subcorpus representa sobre la totalidad del corpus. Es decir, si el subcorpus representa el 8 % del corpus, la frecuencia de cada lema debería tener ese porcentaje. Si la frecuencia del lema es claramente más alta o más baja que la esperada hay que pensar en la actuación de algún factor especial.

Trabajar en esta línea implica utilizar algún recurso estadístico que nos proporcione un índice de dispersión, es decir, una medida basada en procedimientos matemáticos que nos permitan valorar el grado en que un lema está distribuido de forma homogénea o heterogénea entre los diferentes subcorpus construidos. Dada la complejidad de los factores planteados, hay varias formas distintas de hacer los cálculos correspondientes y obtener una cifra que permita hacer esa valoración. Uno de los más utilizados es el que se empleó en el FDSW, que ha sido criticado desde diferentes puntos de vista. Para los interesados en esta cuestión será de utilidad la consulta de Gries (2008) y Egbert *et al.* (2020), trabajos en los que se pasa revista a los diferentes procedimientos utilizados y se analizan sus ventajas e inconvenientes.

[9] De hecho, solo hay dos lemas con FG igual o inferior a 90 que se documenten en los 21 países: el adjetivo *elogiado,* con 86 casos*,* y el verbo *reinstaurar*, con 83. Exigir una frecuencia de 90 supone más que cuadruplicar el número de países considerados, así que teóricamente podrían darse 4 casos en todos (84 en total) y 5 en 6 de ellos.

En la preparación del DF basado en los textos de prensa de la versión 1.0 del CORPES e incorporado a la versión 1.1 se ha seguido el método de la diferencia de proporciones (DP), propuesto por Gries (2008) y descrito también en Brezina (2018: 52-53). Se trata de un estadístico muy intuitivo, que proporciona resultados válidos para cualquier número de subcorpus y con diferentes tamaños. Es el más adecuado, sin duda, para el subconjunto del CORPES que estamos manejando aquí. De forma muy general, el DP se basa en la comparación entre la frecuencia que obtiene un lema en los diferentes subcorpus (la frecuencia observada) y la esperable en función del volumen del subcorpus sobre el corpus total (frecuencia esperada). Es, pues, una vía no muy diferente de la utilizada para el cálculo del χ^2. Se obtiene calculando la suma de las diferencias absolutas entre la proporción observada y la esperada para cada subcorpus y dividiendo el resultado entre 210. La gran ventaja de este procedimiento, además de la sencillez de los cálculos necesarios, estriba en que los resultados oscilan, en la inmensa mayoría de los casos, entre 0 y 1, con lo que proporcionan una idea clara del grado de dispersión de los elementos considerados: los más próximos a 0 son los que muestran una distribución más homogénea y los más cercanos a 1 son aquellos que están distribuidos de forma más irregular.

El primer resultado de interés que podemos obtener es la distribución general de los lemas según los índices de dispersión obtenidos mediante la desviación de proporciones. Las cifras figuran en el cuadro 6.

DP	F	%	% acum.
dp<0.05	342	0,29	0,29
dp>=0.05 y dp<0.1	2999	2,57	2,86
dp>=0.1 y dp<0.2	9494	8,13	11,00
dp>=0.2 y dp<0.3	8488	7,27	18,27
dp>=0.3 y dp<0.4	8034	6,88	25,15
dp>=0.4 y dp<0.5	8109	6,95	32,10
dp>=0.5 y dp<0.6	9426	8,08	40,18
dp>=0.6 y dp<0.7	12 316	10,55	50,73
dp>=0.7 y dp<0.8	19 039	16,31	67,05
dp>=0.8 y dp<0.9	14 544	12,46	79,51
dp>=0.9	23 916	20,49	100,00
Totales	116 707	100,00	

Cuadro 6: Distribución de los lemas por tramos del índice DP
Fuente: CORPES (Diccionario de frecuencias léxicas). Elaboración propia

El cuadro 6 muestra una configuración en la que predomina claramente la distribución heterogénea entre los diferentes países. Los porcentajes acumulados que figuran en la última columna muestran que para alcanzar el 50 % de los lemas registrados es necesario llegar a un DP de 0,69.

[10] Para detalles más técnicos, *vid*. Brezina (2018: 52-53).
En la *Guía* adjunta al diccionario de frecuencias del CORPES https://www.rae.es/corpes/assets/rae/files/corpes/guiaDiccionariosFrecuenciasLex.pdf puede encontrarse una explicación detallada del procedimiento de cálculo y algunos ejemplos ilustrativos.

Los que tienen distribución básicamente homogénea (digamos, con valores inferiores a 0,30) suman poco más del 18 %, inferior incluso a los que tienen DP iguales o superiores a 0,9. El gráfico 1 resulta realmente ilustrativo de lo que sucede:

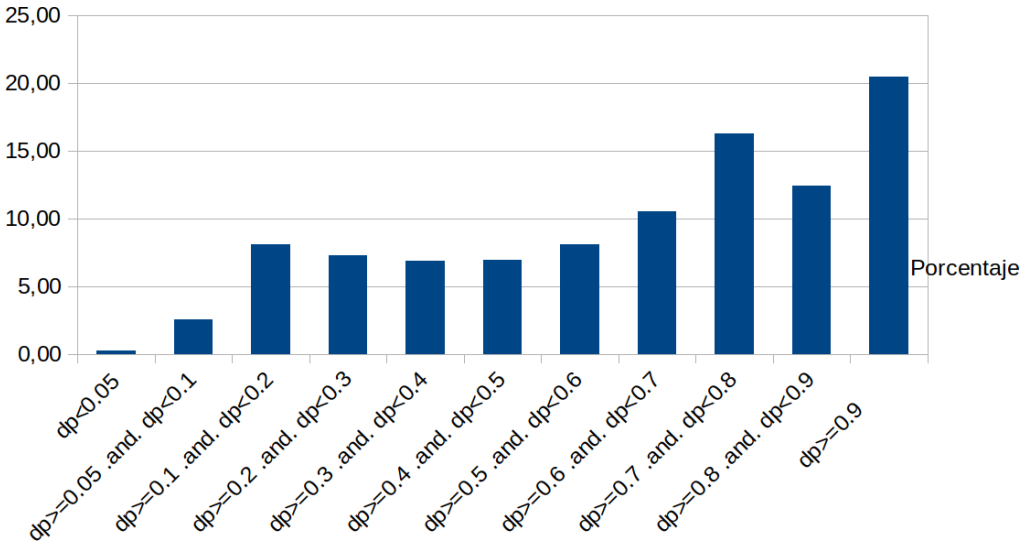

Gráfico 1: Representación de los porcentajes correspondientes a tramos de la DP
Fuente: CORPES (Diccionario de frecuencias léxicas)
Fuente: CORPES (Diccionario de frecuencias léxicas). Elaboración propia

Con datos agrupados en cuatro grandes bloques, la sensación se hace incluso más fuerte, como muestra el cuadro 7, en la que se observa que la suma de porcentajes de los que presentan los índices de dispersión más bajos se queda incluso ligeramente por debajo de lo que suponen los que la tienen más alta.

Carácter del DP	Porcentaje
Dispersión baja	3,1557
Dispersión media-baja	29,2399
Dispersión media-alta	34,943
Dispersión alta	32,9543

Cuadro 7: Porcentajes de elementos según valores de la DP
Fuente: CORPES (Diccionario de frecuencias léxicas). Elaboración propia

La perspectiva cambia radicalmente si, como en el análisis anterior, pasamos de la frecuencia de inventario a la frecuencia de uso, que es lo que se incluye en el cuadro 8. Los lemas con DP<0,05, que son solo el 0,29 % del inventario de lemas, suponen casi el 60 % de la frecuencia de uso. Los que hemos agrupado en el cuadro 7 como de dispersión baja superan el 80 % del total y si consideramos también los que tienen DP inferior a 0,2 alcanzamos el 95 %. En sentido contrario, los que tienen DP igual o superior a 0,8 acumulan únicamente el 0,11 % de los usos.

El gráfico 2 deja ver el enorme contraste que se da entre ambas perspectivas.

	FU	%	% acum.
	108 481		
dp<0.05	598	58,92	58,92
dp>=0.05 y dp<0.1	40 939 564	22,24	81,16
dp>=0.1 y dp<0.2	25 181 944	13,68	94,84
dp>=0.2 y dp<0.3	5 259 029	2,86	97,69
dp>=0.3 y dp<0.4	1 834 355	1,00	98,69
dp>=0.4 y dp<0.5	879 670	0,48	99,17
dp>=0.5 y dp<0.6	693 436	0,38	99,54
dp>=0.6 y dp<0.7	424 599	0,23	99,78
dp>=0.7 y dp<0.8	199 790	0,11	99,88
dp>=0.8 y dp<0.9	134 924	0,07	99,96
dp>=0.9	79 244	0,04	100,00
	184 108		
	153	100	

Cuadro 8: Frecuencia de uso de los lemas según el valor de su DP
Fuente: CORPES (Diccionario de frecuencias léxicas). Elaboración propia

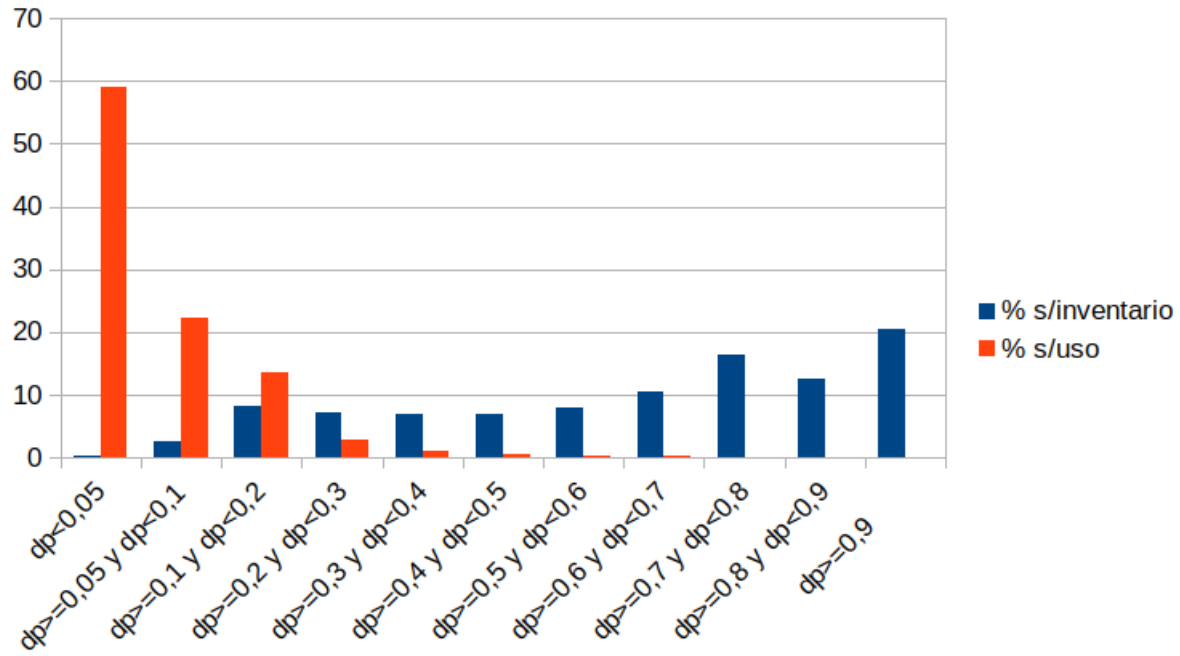

Gráfico 2: Porcentajes de inventario y uso en tramos del índice DP
Fuente: CORPES (Diccionario de frecuencias léxicas). Elaboración propia

4. Conclusiones

Los datos analizados en este trabajo muestran, creo que por primera vez con datos de este volumen, la forma en que se distribuyen los elementos léxicos del español en los diferentes países del mundo hispánico. La tensión entre lo general y lo específico está claramente relacionada con la diferencia entre la frecuencia de uso y la frecuencia de inventario.

Los casi 120 000 lemas (con indicación de clase) resultantes de la anotación de textos de prensa del CORPES aúnan la frecuencia de uso y los índices de dispersión, de modo que el reducido porcentaje de lemas (inventario) que tienen las DP más bajas es el que acumula los mayores porcentajes de la frecuencia de uso. Esta asociación es la que explica finalmente el hecho de que, al menos en estos niveles lingüísticos, los hablantes de cualquiera de las variedades comprendan con facilidad los textos escritos en cualquier otra. Naturalmente, lo anterior no niega la existencia de elementos léxicos que pueden resultar desconocidos para los hablantes de variedades distintas de la originaria del texto, pero establece los límites de esas dificultades.

Para la valoración completa de los datos aquí expuestos no debe olvidarse que el análisis realizado no puede tener en cuenta la existencia de diferentes acepciones de un lema, algunas de las cuales están reducidas a un país. El volumen de datos analizados (unos 180 millones de palabras) explica también las diferencias con los resultados obtenidos en algunas investigaciones previas, realizadas con corpus mucho más reducidos. Con criterios puramente mecánicos, el número de hápax obtenidos en un corpus de 184 millones de palabras y unos 116 000 lemas distintos es 29 172 (el 25 %), con lo que se establece ya un altísimo nivel de partida para los lemas que tienen documentación forzosamente exclusiva de un país. Al tiempo, y yendo al otro extremo, entre los 10 000 lemas con mayor FU, solo el 4 % no está documentado en los 21 países tomados en consideración. Esta última es la cifra que resulta representativa del léxico habitual del español tal como se refleja en los textos de prensa.

Corpus y otros recursos electrónicos mencionados

CORPES: Real Academia Española. *C*orpus del Español del Siglo XXI, http://rae.es/recursos/banco-de-datos/corpes-xxi. Versiones 1.0 y 1.1.

Diccionario de Frecuencias Léxicas (basado en los textos de prensa de la versión 1.0 del CORPES), https://www.rae.es/corpes/assets/rae/files/corpes/guiaDiccionariosFrecuenciasLex.pdf

ESLORA: Corpus para el Estudio del Español Oral. Coord. Victoria Vázquez Rozas, http://eslora.usc.es/. Versión 2.2.

VARILEX: Variación Léxica en Español del Mundo. Coord. Hiroto Ueda, https://h-ueda.sakura.ne.jp/varilex-r/.

Referencias bibliográficas

Andión Herrero, María Antonieta. 2003. La lengua en la prensa española e hispanoamericana en Internet: el fantasma de la diferenciación, *Español Actual*, 76: 71-92.

Atkins, B. T. Sue y Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography,* Nueva York, Oxford University Press.

Ávila, Raúl. 2000. Lenguaje y medios: noticias internacionales, *Anuario de Letras*, 38: 37-65.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics. A practical guide,* Cambridge, Cambridge University Press.

Bybee, Joan. 2007. *Frequency of use and the organization of language,* Oxford, Oxford University Press.

Caravedo, Rocío. 2000. *Léxico del habla culta de Lima,* Lima, Pontificia Universidad Católica del Perú.

Castillo Fadić, María Natalia. 2021. *Léxico básico del español de Chile,* Santiago de Chile, Liberalia Ediciones.

Davies, Mark y Kathy H. Davies. 2018. *A Frequency Dictionary of Spanish. Core Vocabulary for Learners,* Nueva York / Londres, Routledge.

Egbert, Jesse, Brent Burch y Douglas Biber. 2020. Lexical dispersion and corpus design, *International Journal of Corpus Linguistics,* 25, 1: 89-115.

FDSW = Juilland, Alphonse y Eugenio Chang-Rodríguez. 1964. *Frequency Dictionary of Spanish Words,* La Haya, Mouton.

Gries, Stephan. Th. 2008. Dispersions and adjusted frequencies in corpora, *International Journal of Corpus Linguistics*, 13, 4: 403-437.

Haensch, Günther. 2003. Español de América y español de Europa (2.ª parte), *Panace@,* 3, 7: 37-64.

Lipsky, John M. 1994. *Latin American Spanish*. Londres: Longman. Hay trad. esp. de Silvia Iglesias Recuero: *El español de América*. Madrid, Cátedra, 1996.

López Morales, Humberto. 2006. *La globalización del léxico hispánico,* Madrid, Cátedra.

Morales, Amparo, 1986. *Léxico básico del español de Puerto Rico,* San Juan, Academia Puertorriqueña de la Lengua Española.

Moreno de Alba, José G. 1992. *Diferencias léxicas entre España y América,* Madrid, Mapfre.

Moreno de Alba, José G. 1993. *El español en América,* Ciudad de México, Fondo de Cultura Económica (1988).

Moreno Fernández, Francisco. 2016. La búsqueda de un "español global". Ponencia presentada en el CILE celebrado en Puerto Rico 2016. https://congresosdelalengua.es/puerto-rico/paneles-ponencias/espanol-mundo/moreno-fancisco.htm

Nation, I.S.P. 2006. How Large a Vocabulary Is Needed For Reading and Listening?, *The Canadian Modern Language Review/La Revue canadienne des langues vivantes,* 63, 1: 59-82.

Nation, I. S. P. 2016. Word lists. En Nation, I.S.P. *Making and Using Word Lists for Language Learning and Testing*, Amsterdam / Philadelphia, John Benjamins, 3-13.

Rojo, Guillermo. 2008. Lingüística de corpus y lingüística del español, *Actas del XV Congreso de la Asociación de Lingüística y Filología de América Latina*. Montevideo. Edición en CD. http://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf

Rojo, Guillermo. 2011. Frecuencia de inventario y frecuencia de uso, *Revista española de lingüística*, 41, 1: 5-43.

Rojo, Guillermo. 2017. Sobre la configuración estadística de los corpus textuales, *Lingüística* 33, 1: 121-134.
Rojo, Guillermo. 2023. Hacia un nuevo concepto de diccionario de frecuencias. En Dolores Corbella, Josefa Dorta y Rafael Padrón (eds.): *Perspectives en linguistique et philologie romanes (I et II),* París, Éditions de Linguistique et de Philologie, Bibliotèque de Linguistique Romane (BiLiRo), vol. I, 18, 1: 45-63.
Rojo, Guillermo. En prensa. "Un breve apunte sobre frecuencias léxicas".
Schmitt, Norbert, Xiangying Jiang y William Grahe. 2011. The percentage of words known in a text and reading comprehension, *Modern Language Journal*, 95, 1: 26-43.
Tejera, María Josefina. 2003. La tercera norma del español de América, en Francisco Moreno Fernández, Francisco Gimeno Menéndez, José Antonio Samper, María Luz Gutiérrez Araus, María Vaquero y César Hernández (coords.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, Madrid, Arco/Libros: 455-467.

NOTA

El autor de este artículo es el único responsable por su contenido y redacción.

Disponibilidad de datos

El conjunto de datos utilizados para la elaboración de este trabajo es el contenido en el Diccionario de Frecuencias Léxicas asociado a la versión 1.1 del CORPES, disponible en: https://www.rae.es/corpes/assets/rae/files/corpes/corpes_elementos.zip.
No están disponibles los referentes a los lemas documentados exclusivamente en un país.

Nota de aceptación

Este texto ha sido aceptado para publicación por el único Director-Editor de la revista, Adolfo Elizaincín, quien ha actuado de acuerdo a lo establecido en la "Declaración de comportamiento ético" de la revista *Lingüística* https://www.mundoalfal.org/sites/default/files/revista/Declaracion_comp_etico.pdf), primer párrafo del capítulo "Obligaciones del Director-Editor". A esta declaración deben adherir, explícitamente, el Director-Editor, los árbitros y los autores.