El Diccionario de frecuencias léxicas derivado del CORPES
The frequency dictionary derived from CORPES

GUILLERMO ROJO
Universidade de Santiago de Compostela-Real Academia Española
guillermo.rojo@usc.es

Resumen: El artículo pretende resumir las características más importantes del diccionario de frecuencias basado en los datos del CORPES que publicó la Real Academia Española en 2024, con la versión 1.1 del corpus. Además de facilitar la estructura de los dos componentes de este recurso, se analizan someramente algunas de las consecuencias que se pueden deducir de los datos obtenidos y se destacan las ventajas que supone un diccionario de frecuencias léxicas concebido y desarrollado en formato electrónico.

Palabras clave: diccionarios de frecuencia; estadística léxica; distribución diatópica del léxico.

Abstract: The article aims to summarize the most important features of the frequency dictionary based on CORPES (v. 10) data published by the Real Academia Española in 2024, with version 1.1 of the corpus. In addition to providing the structure of the two components of this resource, we briefly analyze some of the consequences that can be deduced from the data obtained and highlight the advantages of a frequency dictionary conceived and developed in electronic format.

Key words: frequency dictionaries; lexical statistics; diatopic distribution of the lexicon.

Fecha de presentación: 12/11/2024 *Fecha de aceptación*: 20/12/2024

1. INTRODUCCIÓN

Aunque sin negar la importancia que tuvieron en su tiempo las listas de frecuencias léxicas elaboradas por Keniston (1920), Cartwright (1925), Buchanan (1927), Keniston (1933 y 1941), Rodríguez Bou (1952) o García Hoz (1953), es de justicia reconocer la influencia ejercida por la parte dedicada al español del proyecto liderado por Alphonse Juilland, que estableció un hito muy señalado en este terreno. Concebido y desarrollado en la década de los años 60 del siglo pasado, el *Frequency Dictionary of Spanish Words* (FDSW = Juilland y Chang-Rodríguez 1964) supuso un más

GUILLERMO ROJO,
«El diccionario de frecuencias léxicas derivado del CORPES»,
Revista de Lexicografía, XXX (2024), pp. 133-159
ISSN: 1134-4539, *e-ISSN*: 2603-667. *doi*: https://doi.org/10.17979/rlex.2024.11847

que significativo avance en los análisis estadísticos y todavía hoy es necesario prestarle atención en cuanto a la concepción y tratamiento de los datos, aunque, como es lógico, los resultados obtenidos no resulten aplicables al español actual[1].

El FDSW fue construido sobre un corpus de 500 000 palabras obtenido mediante sorteo de obras y oraciones correspondientes a textos de autores españoles publicados en el período de entreguerras. La recogida y tratamiento manual de los datos fue complementada en la fase final del proyecto por un procesamiento computacional que permitió obtener con mayor comodidad las frecuencias, índices de dispersión e índices de uso de los 5024 lemas que superaron el índice de uso 3,08 y también de las formas asociadas a cada uno de ellos[2]. La limitación se justifica, por tanto, con un resultado estadístico, pero no se puede olvidar que el FSDW (y, en realidad, todo el proyecto de Juilland) fue concebido para ser publicado en formato impreso. Un libro que contuviera también los lemas con frecuencias e índices de uso inferiores a los fijados habría dado lugar a una publicación de varios miles de páginas, lo cual resultaba impensable en aquel momento.

En términos generales, el FDSW recibió críticas (fundadas) por el escaso tamaño del corpus empleado (incluso para la época), la lematización practicada y la consistencia teórica de los índices utilizados (*cf.*, entre otras, las reseñas de Muller 1965, Bustos Tovar 1966 o el trabajo de Biber *et al.* 2016), pero eso no impide que haya sido durante decenas de años un modelo para obras de carácter similar sobre el español y también sobre otras lenguas. El modelo del FSDW fue posteriormente aplicado al español de Puerto Rico (Morales 1986) y, más recientemente, al de Chile (Castillo Fadić 2001).

Los diccionarios y listas de frecuencias (DF) publicados con posterioridad al FDSW intentaron superar algunos de los problemas señalados o bien fueron construidos sobre materiales de distinto carácter. Por ejemplo, el elaborado por Justicia (1995) utiliza redacciones y test de asociación libre de palabras realizados por es-

[1] Para una interesante perspectiva de las listas de frecuencias léxicas del inglés publicadas antes de la lista de Thorndike-Lorge (1944), *cf*. Bontrager 1991. Para el análisis de las características de las obras mencionadas sobre el léxico español resultan de gran utilidad los trabajos de Ezquerra (1974 y 1977).

[2] En realidad, el apuntado es solo el factor final de la selección. El primer resultado produjo unas 50 000 formas distintas correspondientes a algo menos de 20 000 lemas una vez eliminados extranjerismos, nombres propios, etc. En la fase siguiente, desaparecieron los lemas con frecuencia inferior a 4 (aunque las listas finales dan 5 como frecuencia mínima), con lo que se llegó a 14 000 lemas. Se redujeron a 9000 al eliminar los lemas documentados en solo 3 de los 5 «mundos» y, finalmente, se llegó a 5024 al exigir un índice de uso superior a 3,08 (Juilland y Chang-Rodríguez 1964: lxxiv-lxxvi.

tudiantes de EGB de algunas provincias andaluzas para analizar el desarrollo del léxico y tratar de establecer el vocabulario común. El DF preparado por Alameda y Cuetos (1995) resulta del procesamiento de 2 000 000 de palabras, pero se limita a los listados de frecuencias de formas ortográficas, sílabas, bigramas y letras. Aunque no es en realidad un DF, LEXESP (Sebastián Gallés *et al*. 2000), basado en un corpus etiquetado de 500 000 palabras, permite obtener datos de frecuencias de palabras, lemas y categorías gramaticales, coapariciones, etc. El construido sobre el corpus CUMBRE (Almela *et al*. 2005) proporciona los datos correspondientes a las 10 000 formas ortográficas más frecuentes de un corpus de 20 millones y los 5000 lemas más frecuentes de un subcorpus etiquetado de 2 millones de palabras. Los 5000 lemas más frecuentes del subcorpus lematizado correspondiente al siglo xx del Corpus del Español (género/histórico, CdEhist) son también los incluidos en Davies (2006) y lo mismo se encuentra en Davies y Davies (2018)[3], construido sobre los 2000 millones de formas del Corpus del Español (web/dialectos, CdEweb) más los 20 millones usados en Davies (2006). La Lista de Frecuencias de Palabras del Castellano de Chile (LIFCACH), preparada por Zadowski y Martínez Gamboa, contiene cerca de 850 000 lemas (incluye nombres propios, cifras, etc.) extraídos de un corpus de unos 450 millones de palabras mayoritariamente procedentes de textos de prensa chilenos. Contiene la frecuencia general (FG) y la frecuencia normalizada (FN) de cada lema, además de la FG correspondiente a cada uno de los 102 subcorpus que integran el Corpus Dinámico del Castellano de Chile (CODICACH). En estos casos, las obras se limitan a dar los datos de frecuencia de las formas ortográficas (Alameda y Cuetos), los lemas (LIFCACH y, en otro bloque, también de las formas en Almela *et al*.)[4] y, en el caso de Davies y Davies (2018), a proporcionar además unos índices de uso de lemas elaborados sobre agrupaciones de carácter altamente discutible[5].

A pesar de la posibilidad de usar recursos computacionales desde la misma concepción del trabajo, la única mejora que estas publicaciones suponen sobre el

[3] Presentado como segunda edición de Davies (2006).

[4] Es decir, que no se adscriben las formas a los lemas a los cuales pertenecen.

[5] Según indican los autores (p. 9), el cálculo del índice se hace a base de considerar la frecuencia normalizada y el índice *D* de Juilland para cuatro agrupaciones de textos: ficción, no-ficción, páginas web y textos orales. El resultado ha sido utilizado para seleccionar los 5000 lemas que figuran en el diccionario. En las entradas proporcionan la frecuencia general en los 20 millones del siglo xx del CdEhist, la del CdEweb y, cuando es pertinente, incluyen una indicación general acerca de la preferencia del lema por una de estas agrupaciones.

FDSW radica en el tamaño del corpus utilizado. Resultan claramente inferiores en la estructuración de los subcorpus sobre los que se calculan los índices de dispersión y uso y, sobre todo, en la falta de información sobre la frecuencia de las formas adscritas a cada lema, totalmente inexistente[6].

Aunque las deficiencias señaladas tienen causas diversas, es claro que muchas de ellas proceden del hecho de que estos DF han sido concebidos para su difusión en formato impreso, lo cual limita enormemente el tamaño de las listas publicadas y la riqueza de la información proporcionada. Trabajar con corpus suficientemente amplios, constituidos por textos que han recibido una codificación suficientemente detallada y destinados a ser distribuidos en formato electrónico permite concebir y desarrollar un DF mucho más rico y útil no solo para trabajar con las frecuencias léxicas, sino también con las frecuencias de elementos gramaticales. Eso es lo que se ha pretendido con el DF basado en el CORPES que se describe en los apartados siguientes.

2. El DF basado en el CORPES (versión 1.0)

El DF del CORPES se ha construido sobre un subconjunto de la versión 1.0 de este corpus. Está formado por los textos de prensa correspondientes a 21 de los países hispánicos. De los que tienen textos en el CORPES no se han incluido Guinea Ecuatorial ni Filipinas porque el volumen de textos que se ha podido incorporar es demasiado reducido. El volumen total del subcorpus, una vez eliminados signos de puntuación, nombres propios, cifras, fechas y algunos otros elementos que consideramos irrelevantes para esta finalidad es de 184 millones de elementos gramaticales[7].

La decisión de tomar en consideración únicamente los textos de prensa se basa fundamentalmente en los inconvenientes observados en los DF publicados hasta el momento. Una buena parte de sus insuficiencias a la hora de aplicar los resultados a terrenos como la cobertura léxica, la enseñanza de lenguas, etc. procede básicamente de dos factores. En primer lugar, el hecho de que la inmensa mayoría de los corpus que manejamos habitualmente está constituida por textos escritos y de ca-

[6] Según se indica en https://www.wordfrequency.info/spanish.asp, del CdEweb es posible obtener también la relación de los 40 000 lemas más frecuentes, las formas adscritas a esos 40 000 lemas y las 200 000 formas más frecuentes (sin lematización ni indicación de la clase de palabras).

[7] *Cf. infra*, tabla 1. Los elementos gramaticales son los que resultan de los procesos de segmentación de los textos. Las formas ortográficas *decirlo*, *diciéndolo* o *del* constan de dos elementos gramaticales cada una de ellas. En sentido contrario, la secuencia *a pesar de*, constituida por tres formas ortográficas, es un único elemento gramatical.

rácter formal, lo cual tiene un coste importante cuando se pretende aplicar sus resultados a la fijación de un léxico básico, fundamental[8]. El segundo procede de la misma estructura estadística de los textos. La mayor parte de ellos presenta formas y lemas que no se documentan en otros del mismo tipo. El análisis que Miller y Biber (2015) hicieron sobre diez introducciones a los estudios universitarios en psicología lo muestra con toda claridad. El estudio reveló que, a pesar de la identidad de tema y nivel de tratamiento técnico, los lemas que aparecen exclusivamente en uno de los textos (excluidos nombres propios) se elevan al 41 %, mientras que solo el 13 % puede documentarse en los diez. Parece claro que incluir, por ejemplo, narraciones ambientadas en diferentes marcos temáticos o temporales multiplicará el efecto diversificador y, en consecuencia, distorsionará un tanto los resultados.

Trabajar únicamente con textos de prensa reduce una buena parte de esos inconvenientes. Aunque, por supuesto, existen muchas diferencias entre las publicaciones, el tono general de los periódicos es semejante en todo el mundo hispánico: es lengua formal, pero con un marcado interés por lograr la comprensión general. De otra parte, los temas tratados se inscriben en las mismas esferas. Hay, por supuesto, una importante variedad temática (economía, política, cultura, ocio, salud, etc.), pero esa variedad se da básicamente del mismo modo en todas las publicaciones. Es importante tener en cuenta que el CORPES establece porcentajes para la prensa y para cada área temática en los diferentes países, de modo que el volumen de noticias periodísticas dedicadas a, por ejemplo, temas políticos tiene el mismo porcentaje en todos los países. Por supuesto, el subcorpus resultante tiene una configuración heterogénea, como se muestra en la tabla 1, pero creo que se puede asegurar que la mayor parte de las divergencias observadas proceden fundamentalmente de la propia estructura del léxico y de las diferencias esperables entre las distintas variedades del español.

Una de las mayores ventajas de trabajar sobre un subconjunto de un corpus de referencia como el CORPES radica en el hecho de que los materiales incluidos han pasado previamente los procesos de selección, codificación, lematización y anotación que se aplican a todos los textos. Esto implica que, con los inevitables errores debidos a los procesos automáticos, el tratamiento es homogéneo, de modo que los resultados obtenidos para un país pueden ser contrastados con seguridad con los correspondien-

[8] Estos inconvenientes son los que llevaron a la utilización de acercamientos distintos a la simple frecuencia, como el léxico fundamental o el léxico disponible, para la organización del vocabulario que debería aparecer en, por ejemplo, cursos de ELE. En Ezquerra (1974: 22) puede verse una interesante relación de términos vinculados a la alimentación que figuran en el léxico fundamental incluido en el curso *Vida y diálogos de España* (Rojo Sastre *et al.* 1969) y no aparecen en el FDSW.

tes a otros. Lo mismo puede decirse de la comparación de la prensa con textos narrativos, ensayísticos, etc. Y no es irrelevante tener en cuenta que en proyectos como este no hay costes añadidos, es decir, que se trabaja con el resultado de tratamientos que hay que aplicar a todos los textos del corpus aunque no se vaya a producir un DF.

Elaborar listas de frecuencias léxicas (o un DF) supone, tradicionalmente, purgar ciertos elementos que surgen en el proceso de análisis. Son excluidos habitualmente de los recuentos los nombres propios (tanto de personas como de entidades), cifras, fechas, numerales escritos con caracteres alfabéticos, abreviaturas, símbolos y algunas otras clases de elementos que no se consideran pertenecientes al léxico en sentido estricto. En el caso del DF derivado del CORPES se han fundido en un lema único las variantes que, por algún proceso propio del sistema de anotación, presentaban diferencias entre grafías con mayúsculas y minúsculas, pero recibían la misma clase gramatical. Después de haber realizado toda esta regularización, el corpus utilizado contiene unos 184 millones de elementos gramaticales, con la distribución que aparece en la tabla 1.

Con el conjunto de procesos descritos hasta aquí se obtiene una lista con la FG de cada lema y también la FN (en casos por millón, como es habitual en volúmenes de este tipo). El enriquecimiento de esta lista puede lograrse por diferentes vías. La primera posibilidad es, sin duda, la diferenciación por países. Facilitar la FG y, sobre todo, la FN de cada lema en los 21 países analizados puede proporcionar un panorama preciso de la diversidad léxica del español. Otra posibilidad, igualmente interesante, consiste en establecer las FG y FN de cada lema para cada área temática considerada en el corpus (política y economía, ocio, salud, etc.). Por supuesto, es posible también conjuntar las dos perspectivas: frecuencia del lema en general, frecuencia por país y, dentro de cada uno, frecuencia por área temática.

Aunque todas estas posibilidades resultan enormemente atractivas, es claro que producen lo que la mayoría de los consultantes considerarían exceso de información. Hay que tener en cuenta que el sistema de consultas incorporado al CORPES permite ya obtener datos como los mencionados para un lema o un grupo de lemas. Al tiempo, resulta evidente que la frecuencia de un lema (o de otro elemento cualquiera) es un factor relevante para valorar su «importancia», pero necesita ser complementada con otros aspectos. Si solo atendemos a su frecuencia general, *lempira,* con FG 1267 y FN 6,88 casos por millón (cpm), figura con el rango 7143, en el mismo bloque en que aparecen, por ejemplo, *espinaca*, *gástrico*, *despensa*, *indeseable*, *arrollar*, *diócesis* y muchos otros términos bastante más familiares. La

mayoría de los hispanohablantes, que desconoce el vocablo, se sorprenderá de la posición de este elemento, pero comenzará a entender lo que sucede si observa a continuación que 1236 casos de ese lema (el 97,6 %) proceden de Honduras y que, en consonancia con ese desequilibrio, su índice de dispersión es de 0,96 (para un máximo de 1,0 en el caso de dispersión máxima). Se trata, por tanto, de un término muy frecuente en un país (su rango de frecuencia en ese subconjunto es el 179) y muy poco o nada documentado en los demás, por lo que su introducción en el léxico de un curso de ELE debe estar en las primeras semanas o bien en un nivel muy avanzado en función de la variedad y el ámbito a los que se pretenda atender.

Así pues, la frecuencia necesita ser complementada con la dispersión, esto es, la medida en la que un determinado elemento aparece en los diferentes subcorpus o segmentos de subcorpus. Dado el interés que tiene la distribución geográfica de los lemas, tomamos la decisión de incorporar un índice de dispersión de cada lema y cada forma asociada a él entre los diferentes países. Sabiendo que el índice utilizado (*vid. infra*) oscila entre 0 para los que tienen una distribución totalmente homogénea, y 1 para los que muestran distribución heterogénea, conocer que el índice correspondiente a *lempira* es de 0,96 aclara lo que sucede en casos como este. Consideramos que la inclusión de un índice de dispersión al lado de la frecuencia general y la normalizada incorpora toda la información necesaria para obtener una visión global de la organización de los elementos léxicos en cuanto a su distribución (diatópica en este caso).

En los últimos años se ha publicado una amplia serie de trabajos acerca de las ventajas e inconvenientes de los distintos índices de dispersión que han sido utilizados a partir del famoso índice D propuesto por Juilland y Chang-Rodríguez (1964). En Egbert, Burch y Biber (2020) puede encontrarse un amplio resumen, que incluye la valoración de diferentes aspectos implicados. Para el cálculo de los índices de uso de este DF, hemos decidido utilizar el basado en la diferencia de proporciones (DP) propuesto por Gries (2008) y descrito también en Brezina (2018: 52-53). Se trata de un estadístico sencillo, fácilmente comprensible incluso por personas sin formación técnica en estadística y que produce una fotografía bastante clara de la distribución de los diferentes elementos léxicos entre los diferentes subcorpus. Como ventaja adicional, importante en corpus construidos con propósitos más generales, hay que destacar también que permite trabajar con cualquier número de subcorpus con diferentes tamaños.

El primer paso consiste, lógicamente, en establecer los subcorpus. En el caso del elaborado sobre los textos de prensa del CORPES (versión 1.0), los subcorpus corresponden a los textos procedentes de los 21 países diferentes que ha sido posible tomar en consideración. Obtenemos así una serie de proporciones (0,0811, 0,0237, etc.) cuya suma debe ser 1 (o 100 si se prefiere trabajar con porcentajes). Los datos figuran en la tabla 1.

Subcorpus	Número de formas	Prop. s/ corpus
Argentina	14 924 587	0,0811
Bolivia	4 362 549	0,0237
Chile	10 842 064	0,0589
Colombia	14 549 834	0,0790
Costa Rica	2 494 035	0,0135
Cuba	6 648 104	0,0361
Ecuador	5 469 862	0,0297
El Salvador	2 542 044	0,0138
España	50 307 664	0,2733
Estados Unidos	3 782 503	0,0205
Guatemala	2 550 678	0,0139
Honduras	2 469 098	0,0134
México	21 829 199	0,1186
Nicaragua	2 798 995	0,0152
Panamá	1 790 775	0,0097
Paraguay	4 180 930	0,0227
Perú	5 620 618	0,0305
Puerto Rico	2 760 431	0,0150
República Dominicana	4 578 726	0,0249
Uruguay	5 058 611	0,0275
Venezuela	14 546 846	0,0790
Totales	184 108 153	1

TABLA 1: Tamaño y proporción de los diferentes subcorpus obtenidos. Fuente: CORPES XXI (https://www.rae.es/corpes/assets/rae/files/corpes/guiaDiccionariosFrecuenciasLex.pdf)

En un lema distribuido de forma homogénea, la proporción correspondiente a cada subcorpus debe ser próxima a la que el subcorpus supone sobre el total del corpus. Si no es así, estamos ante un lema distribuido de forma heterogénea y ese desajuste será mayor o menor en función de la diferencia entre la proporción esperada (la que corresponde al peso del subcorpus) y la observada. La suma de las diferencias entre ambas proporciones, dividida entre 2, produce la diferencia de proporciones (DP) que, como se ha indicado ya, oscila entre 0 (distribución totalmente homo-

génea) y 1 (distribución totalmente heterogénea)[9]. La tabla 2 muestra la distribución de *abanicar* y, al tiempo, los cálculos realizados para la obtención de la DP.

Abanicar	Tamaño	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
Argentina	14 924 587	0,0811	1	0,005	15,645	0,081	0,076
Bolivia	4 362 549	0,0237	0	0,000	4,573	0,024	0,024
Chile	10 842 064	0,0589	2	0,010	11,366	0,059	0,049
Colombia	14 549 834	0,0790	2	0,010	15,253	0,079	0,069
Costa Rica	2 494 035	0,0135	3	0,016	2,614	0,014	0,002
Cuba	6 648 104	0,0361	2	0,010	6,969	0,036	0,026
Ecuador	5 469 862	0,0297	0	0,000	5,734	0,030	0,030
El_Salvador	2 542 044	0,0138	0	0,000	2,665	0,014	0,014
España	50 307 664	0,2733	12	0,062	52,737	0,273	0,211
Estados Unidos	3 782 503	0,0205	9	0,047	3,965	0,021	0,026
Guatemala	2 550 678	0,0139	4	0,021	2,674	0,014	0,007
Honduras	2 469 098	0,0134	0	0,000	2,588	0,013	0,013
México	21 829 199	0,1186	25	0,130	22,883	0,119	0,011
Nicaragua	2 798 995	0,0152	17	0,088	2,934	0,015	0,073
Panamá	1 790 775	0,0097	3	0,016	1,877	0,010	0,006
Paraguay	4 180 930	0,0227	0	0,000	4,383	0,023	0,023
Perú	5 620 618	0,0305	1	0,005	5,892	0,031	0,025
Puerto Rico	2 760 431	0,0150	14	0,073	2,894	0,015	0,058
R. Dominicana	4 578 726	0,0249	32	0,166	4,800	0,025	0,141
Uruguay	5 058 611	0,0275	3	0,016	5,303	0,027	0,012
Venezuela	14 546 846	0,0790	63	0,326	15,249	0,079	0,247
Totales	184 108 153	1,0000	193	1	193	1	1,141
						Div. entre 2	0,571

TABLA 2: Muestra de los datos y cálculos realizados para obtener la DP del lema *abanicar* en el DF del CORPES. Fuente: DF del CORPES.

La facilidad de interpretación de la DP presenta, en mi opinión, un único punto gris, que puede resultar inicialmente chocante. Para un subcorpus determinado, cuando la frecuencia que presenta en él un cierto lema es 0, la DP de ese corpus es la misma con independencia de la FG del lema en cuestión. Es decir, en un subcorpus que representa una proporción de 0,1125 sobre el total del corpus, si la frecuencia observada de un lema es 0, su DP será de 0,1125 con independencia de que

[9] Dado que los cálculos pueden producir algunos casos en los que los índices superan el valor 1, Lijffijt y Gries (2012) han propuesto un refinamiento consistente en dividir la DP obtenida según el método anterior entre el tamaño proporcional del subcorpus más pequeño. No parece que la diferencia produzca efectos relevantes, de modo que se ha preferido no utilizarla en este caso.

la FG de ese lema sea 10, 100, 1000, etc. De modo semejante, en caso de un lema se documente solo en uno de los subcorpus la DP parcial de los subcorpus y, por tanto, la total será la misma con independencia de que la FG del lema sea 100 o 1000. En realidad, estos resultados son los esperables puesto que el estadístico trabaja con las proporciones y no directamente con las frecuencias.

La primera salida del DF del CORPES, la más sencilla, contiene FG, FN y DP de cada lema (con clase). A estos campos se ha añadido el rango de frecuencia correspondiente a cada lema y el número de subcorpus (es decir, de países) en los que se ha documentado (en el subcorpus de prensa, como es lógico).

Rango DP	Rango frec.	Lema	Clase	Frecuencia	Frec. norm.	DP	Núm. países
1	3	en	P	5 886 629	31973,755122	0,0062263	21
2	11	con	P	1 921 890	10438,918476	0,00685678	21
3	13	por	P	1 905 485	10349,813243	0,00846239	21
4	1	el	T	24 253 277	131733,856458	0,00963192	21
5	9	se	L	2 901 164	15757,933327	0,00967559	21
6	5	a	P	4 626 346	25128,414601	0,00982385	21
7	2	de	P	15 343 801	83341,235844	0,01226035	21
8	4	y	C	5 461 242	29663,227353	0,01490734	21
9	126	problema	N	112 717	612,232528	0,01588217	21
10	40	uno	M	293 382	1593,530733	0,01642018	21
[…]							
5001	3702	indicio	N	3 845	20,884463	0,11886069	21
5002	5577	umbral	N	1 954	10,613327	0,11886561	21
5003	1237	par	N	16 347	88,790201	0,11887584	21
5004	11496	reciprocidad	N	516	2,8027	0,11887718	21
5005	12454	osado	A	435	2,362742	0,11888626	21
5006	4786	naturalmente	R	2 532	13,752786	0,11889355	21
5007	335	terminar	V	53 258	289,275619	0,11890629	21
5008	9547	rastreo	N	751	4,079124	0,11894274	21
5009	3232	exploración	N	4 752	25,810916	0,11895143	21
5010	7339	devenir	N	1207	6,555929	0,11895584	21
[…]							
15001	21485	hastiado	A	127	0,689812	0,22344837	19
15002	20774	pica	N	139	0,754991	0,22345069	17
15003	5157	icono	N	2225	12,085288	0,22347556	21
15004	29996	metódicamente	R	52	0,282443	0,22349269	15
15005	5930	cometa	N	1748	9,494419	0,22351411	21
15006	8378	guardería	N	959	5,208895	0,22353134	21
15007	21052	megabit	N	134	0,727833	0,22353784	18

15008	25882	pendenciero	A	78	0,423664	0,22353835	14
15009	30484	emancipado	A	50	0,271579	0,223555	14
15010	16544	operado	A	238	1,292718	0,22355728	19

TABLA 3: Muestra del DF basado en el CORPES. Lemas ordenados por DP.
Fuente: DF del CORPES.

La segunda salida, un tanto más compleja, incluye estos mismos datos para cada lema y para cada una de las formas adscritas a él.

Orden	Rango frec.	Lema	Forma	Etiqueta	Frec.	Frec. norm.	DP	Núm. países
291	91813	abocinar		V	1	0,005432	0,977295	1
			Abocinar	V----v0n	1	0,005432	0,9773	1
292	28954	abofetear		V	58	0,315032	0,25868673	15
			abofetea	Vis-3p0n	8	0,043453	0,5152	5
			abofeteaba	Vis-3i0n	3	0,016295	0,7118	2
			abofeteaban	Vip-3i0n	1	0,005432	0,9751	1
			abofeteada	Vfs--d0n	1	0,005432	0,9411	1
			abofeteado	Vms--d0n	3	0,016295	0,6907	2
			abofetean	Vip-3p0n	1	0,005432	0,7268	1
			abofeteando	V----g0n	2	0,010863	0,7268	1
			abofetear	V----v0n	18	0,097769	0,4003	7
			abofetear	V----v1n	4	0,021726	0,5524	3
			abofetee	Vss-3p0n	1	0,005432	0,985	1
			abofeteó	Vis-3t0n	15	0,081474	0,3817	9
			Abofeteó	Vis-3t0n	1	0,005432	0,921	1
293	11054	abogacía		N	565	3,068848	0,46604615	19
			abogacía	Nfsc---n	562	3,052554	0,4684	18
			Abogacía	Nfsc---n	2	0,010863	0,9862	1
			abogacías	Nfpc---n	1	0,005432	0,7268	1

TABLA 4: Muestra del DF basado en CORPES ordenado alfabéticamente.
Fuente: DF del CORPES.

Ambas versiones, descargables de la página web de la RAE (https://www.rae.es/corpes/), consisten en ficheros de texto con campos separados por tabuladores (formato TSV), lo cual permite, sin necesidad de conocimientos profundos en informática, la extracción directa de datos, así como su integración y manejo en hojas de cálculo o bases de datos.

3. Algunos resultados generales destacados

Aunque es posible imprimirlo, el DF del CORPES ha sido concebido como un recurso en formato electrónico. La liberación de las servidumbres del formato impreso ha permitido plantear, por primera vez en la historia de la lexicografía del español, un DF completo, sin la limitación tradicional a los 5000 lemas con mayor índice de uso, difícilmente justificable a la luz de lo que vamos sabiendo acerca de la estructura estadística del léxico. Las consecuencias y derivaciones son muchas y variadas, así que me limitaré aquí a algunos aspectos que me parecen de interés especial.

3.1. *Cuestiones previas*

Como se indica en el apartado 2, la toma en consideración de los datos de frecuencia no es suficiente para entender adecuadamente la forma en que se organizan los elementos léxicos y hay que tener en cuenta también algún estadístico relacionado con las tasas de dispersión. En un nivel probablemente más elemental, pero muy importante, es necesario diferenciar entre las que he propuesto llamar «frecuencia de uso» y «frecuencia de inventario» (*cf.* Rojo 2011)[10]. Corresponden a frecuencias de uso las respuestas a preguntas sobre, por ejemplo, el número de apariciones en un corpus del lema *ejercicio*, el número de apariciones (*tokens*) de los lemas pertenecientes a la clase de los sustantivos, el número de cláusulas con sujeto, complemento directo y complemento indirecto, etc. Corresponden, en cambio, a frecuencias de inventario las respuestas a preguntas sobre el número de lemas o de sustantivos **distintos** (*types*) que aparecen en un texto, el número adjetivos que figuran en el lemario de un diccionario, el número de esquemas sintácticos con estructura triactancial, etc. Podemos decir, así, que en un corpus determinado aparecen 10 000 sustantivos distintos (frecuencia de inventario) que acumulan, en su conjunto, un total de 2 500 000 casos (frecuencia de uso).

3.2. *Frecuencias acumuladas y tasas de cobertura*

La diferencia establecida en el apartado anterior permitirá entender mejor la estructura del léxico, las caracterizaciones generales que se han formulado y sus consecuencias. De acuerdo con la ley de Zipf (1935, 1949), las frecuencias de los elementos léxicos responden a una constante: si la frecuencia de uso (FU) del elemento más abundante de un lemario es *n*, la del segundo será *n*/2, la del tercero *n*/3 y así sucesivamente. Por tanto, la frecuencia de un lema que ocupe el rango 400 en los listados

[10] Conceptos próximos, pero no idénticos, a los de *type frequency* y *token frequency* propuestos por Bybee (2007). *Vid*. Rojo (2011) para más detalles.

de frecuencia será aproximadamente la que corresponde a la frecuencia del que tiene el rango 1 dividido entre 400. El resultado general (*cf*. Nation 2016: 4-5; Miller 2020: 79; Rojo 2021: 131) es que en cualquier texto o corpus existen unos pocos elementos con frecuencias muy altas, muchos elementos con frecuencias bajas o muy bajas y una cantidad importante de elementos que aparecen solo una vez (los hápax)[11]. La distribución general está, por tanto, muy alejada de lo que consideramos la distribución normal de un fenómeno, caracterizada por presentar la mayor parte de los casos concentrados en la zona central. El gráfico 1 muestra el aspecto de la distribución de frecuencias de los lemas según los datos del DF del CORPES[12].

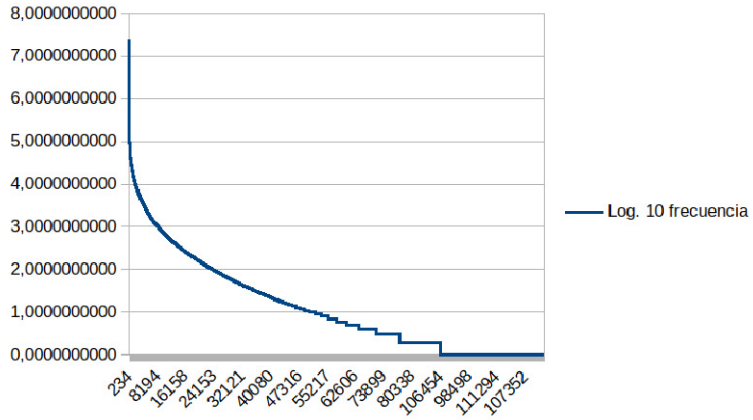

GRÁFICO 1: Representación de las frecuencias de lemas. Eje Y: logaritmo de la frecuencia.

El perfil que se muestra en el gráfico 1 es compatible también con lo que establece el principio de Pareto, según el cual el 80 % del volumen de una cierta variable (la riqueza de un país, por ejemplo) corresponde al 20 % de los efectivos (sus ciudadanos). El principio del 80/20 se queda muy corto cuando se trata de las frecuencias léxicas, como muestra la tabla 5, en la que se ve que el 80 % de los usos

[11] Creo que es importante señalar que, aunque aquí se trabaje con lemas, elementos que muestran mayor capacidad de agrupación que los elementos gramaticales o las formas ortográficas, el porcentaje de hápax con respecto al total de lemas distintos es considerablemente alto. Como muestra la tabla 5, se sitúa en torno al 25 % del total de lemas distintos documentados. Para interpretar correctamente este dato es necesario no olvidar lo ya señalado acerca de la eliminación en los recuentos de nombres propios, cifras, fechas, etc. Los nombres propios muestran una tendencia clara a la aparición muy infrecuente y, por tanto, son integrantes habituales de las listas de hápax.

[12] Utilizo el logaritmo (en base 10) de las frecuencias para que se pueda observar una línea de descenso más «dulce».

del corpus examinado procede de los aproximadamente 1100 lemas más frecuentes (que suponen solo el 9,5 % del inventario de lemas).

Los primeros x lemas	Frecuencias acumuladas por rango	
	Frecuencia general	Porcentaje acum.
10	71 476 162	38,82
25	88 018 180	47,81
50	96 805 394	52,58
100	105 377 049	57,24
500	131 379 918	71,36
1000	145 163 982	78,85
2000	158 784 779	86,25
3000	165 718 819	90,01
4000	169 964 179	92,32
5000	172 798 972	93,86
6000	174 807 685	**94,95**
7000	176 296 768	95,76
8000	177 463 580	96,39
9000	178 397 603	96,90
10000	179 157 502	97,31
12500	180 529 374	**98,06**
15000	181 424 624	98,54
17500	182 048 102	98,88
20000	182 497 159	99,12
25000	183 071 582	99,44
35000	183 616 065	99,73
50000	183 910 931	99,89
100000	184 091 446	99,99
116707	184 108 153	100,00
Total hápax	29172 (= 25 %)	

TABLA 5: Frecuencias acumuladas por rango en el DF basado en el CORPES.
Fuente: DF del CORPES. Elaboración propia

Como es bien sabido y confirma la tabla 5, los lemas más frecuentes tienen frecuencias extraordinariamente altas, por lo que con un porcentaje muy bajo de lemas (por ejemplo, los 50 primeros, que suponen el 0,0004 % del inventario) se reconoce ya algo más del 50 % del uso. Es fácil ver en la tabla que el incremento en los porcentajes de reconocimiento se va reduciendo a medida que bajamos en el rango de frecuencias. Así, pasar de los 100 a los 500 lemas más frecuentes produce

un incremento de 14,12 puntos porcentuales en el reconocimiento, mientras que el esfuerzo que supone la subida de 6000 a 7000 lemas se ve compensado con únicamente un aumento de 0,81 puntos porcentuales. Esa diferencia se va haciendo más importante a medida que nos acercamos a la zona más baja de las frecuencias. El gráfico 2 muestra este proceso con toda claridad.

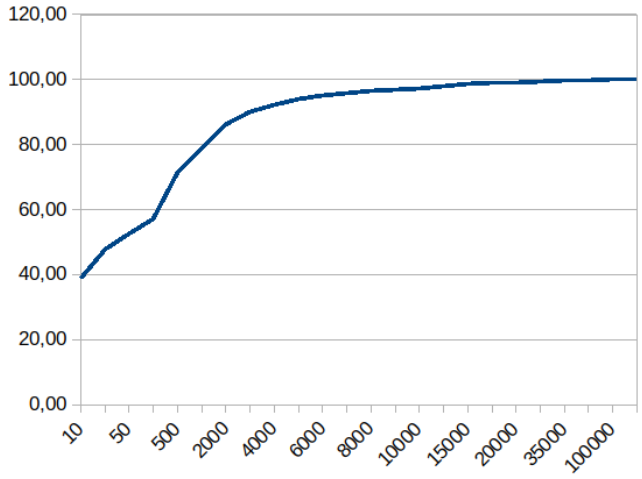

GRÁFICO 2: Evolución de la cobertura de uso
en función de las frecuencias acumuladas

Como ya he señalado, las listas y los diccionarios de frecuencia han sido profusamente utilizados para la organización de la enseñanza del léxico en los cursos de lengua extranjera. Es natural, por tanto, intentar obtener de datos como los que figuran en la tabla 5 una idea acerca del tamaño del léxico que se necesita para entender un cierto porcentaje de un texto. Es evidente que la comprensión de un texto constituye un proceso muy complejo que no se puede evaluar con procedimientos puramente estadísticos como los que estamos utilizando (*cf.* Schmitt *et al.* 2011, Robles-García 2020). Es posible, sin embargo, aproximarse a ese objetivo mediante el análisis de los porcentajes de reconocimiento en la asignación de los lemas según los rangos de frecuencia. En otras palabras, se trata simplemente de comprobar si el lema al que pertenecen las formas incluidas en un texto está incluido en un determinado segmento de frecuencias. Se prescinde, por tanto, de todos los complejos procesos que actúan sobre el aprendizaje e incluso se acepta que la simple identificación de un lema resulta equivalente al dominio de todos sus significados.

Asumiendo esa simplificación, con los datos de la tabla 5 se concluye que el reconocimiento del 95 % de las formas incluidas en un texto exige el dominio de unos 6000 lemas (con clase de palabras). Llegar al 98 % requiere el manejo de unos 12 500 lemas, que es una cifra realmente notable, sobre todo si se piensa en el aprendizaje de lenguas extranjeras. Son, sin embargo, similares a las que se han dado para otras lenguas, como se indica en Nation (2006: 79**)** o se puede ver en las cifras de, por ejemplo, el CORGA (*cf.* Rojo en prensa a). Para valorarlas adecuadamente, debe tenerse en cuenta que el DF del CORPES tiene únicamente textos de prensa y que de su lemario han sido excluidos nombres propios, cifras y algunos otros elementos de nula relevancia en el análisis de las frecuencias léxicas. Trabajamos, por tanto, con lengua escrita, en una versión formal, pero habitualmente producida con el propósito deliberado de llegar a todos los segmentos culturales de la sociedad. Los pocos datos disponibles muestran que el lemario preciso para la comprensión de textos orales es considerablemente menor. Con los datos de ESLORA (*cf*. Rojo 2022, en prensa a) se aprecia que se alcanza el 95 % con unos 2000 lemas y para el 98 % del uso se requieren únicamente unos 4000[13].

El DF basado en el CORPES permite explorar y valorar adecuadamente un aspecto que no ha sido tenido en cuenta hasta el momento. Como es bien sabido, los lemas más frecuentes suponen unos porcentajes muy altos sobre el uso, pero son fundamentalmente artículos, preposiciones o conjunciones. Los 10 lemas más frecuentes según el DF, que suponen en conjunto casi el 40 % de las frecuencias de uso, son los 2 artículos, 3 preposiciones (*de*, *en*, *a*), 2 conjunciones (*y*, *que*), 1 relativo (*que)*, 1 pronombre (*se*) y 1 verbo (*ser*)[14]. Es interesante, por tanto, analizar lo que sucede si reducimos los lemas que vamos a tomar en consideración a las clases con mayor contenido léxico, esto es, sustantivos, adjetivos, verbos y adverbios, como se propone en Rojo (en prensa a y en prensa b). En la tabla 6 se pueden contrastar los porcentajes de uso acumulado correspondientes a todos los elementos en todo el lemario y a las cuatro clases mencionadas sobre el total de elementos de esas clases[15].

[13] Estas cifras corresponden a los lemas sin consideración de nombres propios ni cifras y sin diferenciación de la clase de palabras.

[14] Comprende también los casos en los que funciona como auxiliar de perífrasis.

[15] Téngase en cuenta que, como se deduce de la configuración general de la tabla, los porcentajes se establecen sobre los totales correspondientes en cada columna: la totalidad del corpus en el primer caso y la suma de los adjetivos, adverbios, sustantivos y verbos en el segundo. Por tanto, con los diez lemas más frecuentes del inventario general se reconoce casi el 40 % de los elementos de todo el corpus; con los diez lemas más frecuentes de las clases léxicas, únicamente el 10,56 % de los elementos de estas clases.

	Porcentajes de uso acumulados	
Los primeros x lemas	Todo el lemario	Clases léxicas
10	38,82	10,56
25	47,81	14,60
50	52,58	18,81
100	57,24	24,90
500	71,36	47,57
1000	78,85	60,81
2000	86,25	74,19
3000	90,01	81,15
4000	92,32	85,43
5000	93,86	88,31
6000	94,95	90,36
7000	95,76	91,90
8000	96,39	93,10
9000	96,9	94,07
10000	97,31	94,85
12500	98,06	96,27
15000	98,54	97,21
17500	98,88	97,86
20000	99,12	98,33
25000	99,44	98,92
35000	99,73	99,49
50000	99,89	99,80
100000	99,99	99,99
116707	100,00	100,00
Total lemas	116 707	112 971
Frecuencia total lemas	1841 08 153	92 013 231
Porcentaje hápax	25,00%	24,06%

TABLA 6: Porcentajes acumulados de la totalidad del lemario y las clases léxicas.
Fuente: DF del CORPES. Elaboración propia

Antes de entrar en los detalles de la comparación, es interesante valorar las cifras generales. Las cuatro clases léxicas suponen el 96,79 % de los 116 707 lemas distintos, pero solo el 49,98 % del volumen total del corpus (184 108 153) elementos. Significa esto que al 3,21 % restante (artículos, conjunciones, preposiciones, etc.) le corresponde casi el 50 % de los elementos que forman el corpus; en realidad, de cualquier texto escrito en español. Esas cifras, gruesas, dan una idea clara, me parece, de la importancia de separar estas cuatro clases para obtener una idea más precisa de cómo se organizan las frecuencias de los elementos en los textos.

Las diferencias son realmente notables y proporcionan una visión bastante diferente acerca del volumen del léxico necesario para captar el contenido de un texto. Si nos reducimos a las cuatro clases, alcanzar el 95 % de sus apariciones exige manejar unos 11 000 lemas (frente a unos 6000 en el recuento general) y llegar al 98 % exige casi 20 000 (frente a menos de 12 500 en el global). En el gráfico 3 puede apreciarse el diferente perfil que muestran las curvas de crecimiento correspondientes a los dos enfoques.

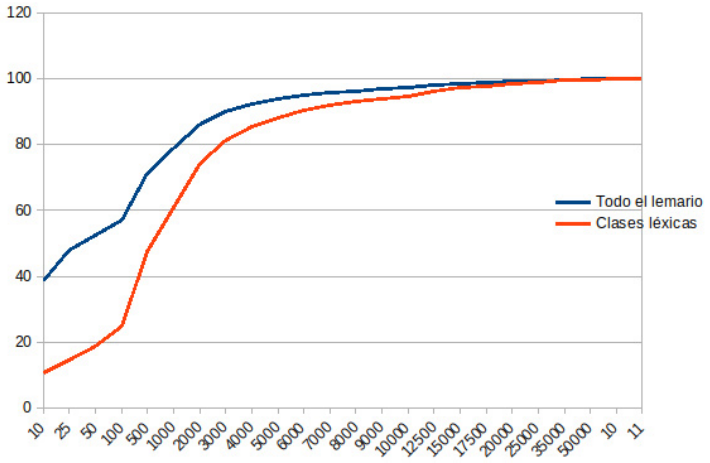

GRÁFICO 3: Porcentajes acumulados de todo el lemario y las clases léxicas.
Fuente: DF del CORPES. Elaboración propia

Creo que las cifras y el gráfico son suficientemente indicativas de que es necesario modificar lo que se dice habitualmente sobre esta cuestión y confirmar que se trata de una consecuencia directa e la posibilidad de construir diccionarios de frecuencia que exploten realmente las posibilidades de los corpus textuales y las herramientas informáticas que tenemos a nuestra disposición.

3.3. *Los índices de dispersión*

Como he indicado en el apartado 2, la frecuencia de un elemento es un factor importante, pero no el único que hay que tener en cuenta a la hora de valorar su peso en el conjunto al que pertenece. Desde cierto punto de vista, la mayor o menor dispersión con que podemos documentarlo resulta incluso más decisiva a la hora de valorar, por ejemplo, el momento en que debería ser introducido un lema concreto en el programa de un curso de ELE.

El DF derivado del CORPES incorpora un índice de dispersión basado en la diferencia de proporciones (DP), como se indica en el apartado 2. La DP oscila entre 0, que corresponde a aquellos lemas que presentan una distribución totalmente homogénea, y 1, que resulta en los que tienen una distribución totalmente heterogénea. Dado que los subcorpus utilizados en este DF han sido constituidos sobre la pertenencia a los 21 países distintos, la DP será, además de una valoración general acerca de la dispersión, un indicador importante del carácter más o menos policéntrico del léxico a lo largo del mundo hispánico.

Lo que ya hemos visto sobre la frecuencia permite hacer una hipótesis acerca de lo que se puede esperar de la dispersión. Es obvio que los elementos que tienen frecuencia más reducida presentarán mayoritariamente un índice de dispersión más alto. En concreto, si sabemos que los hápax constituyen el 25 % de los lemas distintos de este corpus, tenemos ya una considerable cantidad de elementos que forzosamente mostrarán una DP muy elevada. Los datos proporcionados por el DF permiten hacer una consideración mucho más detallada y precisa. En la tabla 7 aparece la distribución de los lemas según tramos de su DP.

DP	% s/inventario
DP<0,05	0,29
DP>=0,05 y DP<0,1	2,57
DP>=0,1 y DP<0,2	8,13
DP>=0,2 y DP<0,3	7,27
DP>=0,3 y DP<0,4	6,88
DP>=0,4 y DP<0,5	6,95
DP>=0,5 y DP<0,6	8,08
DP>=0,6 y DP<0,7	10,55
DP>=0,7 y DP<0,8	16,31
DP>=0,8 y DP<0,9	12,46
DP>=0,9	20,49

TABLA 7. Porcentajes de inventario correspondientes a distintos tramos de las DP en el DF del CORPES. Fuente: DF del CORPES. Elaboración propia

La configuración general que presenta la tabla 7 no es inesperada, pero creo que se puede afirmar que las cifras resultan sorprendentes. Solo el 2,86 % de los lemas tienen una DP inferior a 0,1, mientras que los que presentan DP iguales o superiores a 0,9 constituyen el 20,49 %, siete veces más. Los que tienen DP inferiores a 0,5 suman el 32,09 %, que es, más o menos el que corresponde a los que tienen DP iguales o superiores a 0,8.

La primera conclusión que se obtiene de estos datos es que el léxico del español presenta un grado de dispersión muy alto, con casi el 50 % de los lemas situados en DP iguales o superiores a 0,7. Esta innegable realidad se refiere a los porcentajes sobre las frecuencias de inventario y, como hemos visto en los apartados anteriores, debe ser complementada con las correspondientes a las frecuencias de uso, que son las que figuran en la tabla 8.

DP	% s/uso
DP<0,05	58,92
DP>=0,05 y DP<0,1	22,24
DP>=0,1 y DP<0,2	13,68
DP>=0,2 y DP<0,3	2,86
DP>=0,3 y DP<0,4	1,00
DP>=0,4 y DP<0,5	0,48
DP>=0,5 y DP<0,6	0,38
DP>=0,6 y DP<0,7	0,23
DP>=0,7 y DP<0,8	0,11
DP>=0,8 y DP<0,9	0,07
DP>=0,9	0,04

TABLA 8. Porcentajes de uso correspondientes a distintos tramos de las DP en el DF del CORPES. Fuente: DF del CORPES. Elaboración propia

Esta tabla presenta un panorama general totalmente distinto del anterior. Los lemas que tienen una DP inferior a 0,05 suponen casi el 60 % de los usos registrados en el corpus. Y si elevamos el tope hasta las DP inferiores a 0,2, el porcentaje acumulado asciende a casi el 95 %. Resulta, pues, evidente que la mayoría de los elementos que se pueden documentar en un corpus del estilo del utilizado para este DF presentan unos índices de dispersión muy bajos, lo cual implica que el léxico del español tiene una distribución fundamentalmente homogénea. La diferencia entre ambas consideraciones puede verse con mucha claridad en el gráfico 4.

La consideración de la diferencia entre frecuencia de inventario y frecuencia de uso proporciona la clave para entender correctamente lo que sucede con el inventario léxico del español actual. Es cierto que la mayoría de los lemas muestra unos índices de dispersión altos, pero los elementos que entran en ese grupo tienen habitualmente una frecuencia baja o muy baja y, por tanto, pesan poco sobre la generalidad de los textos.

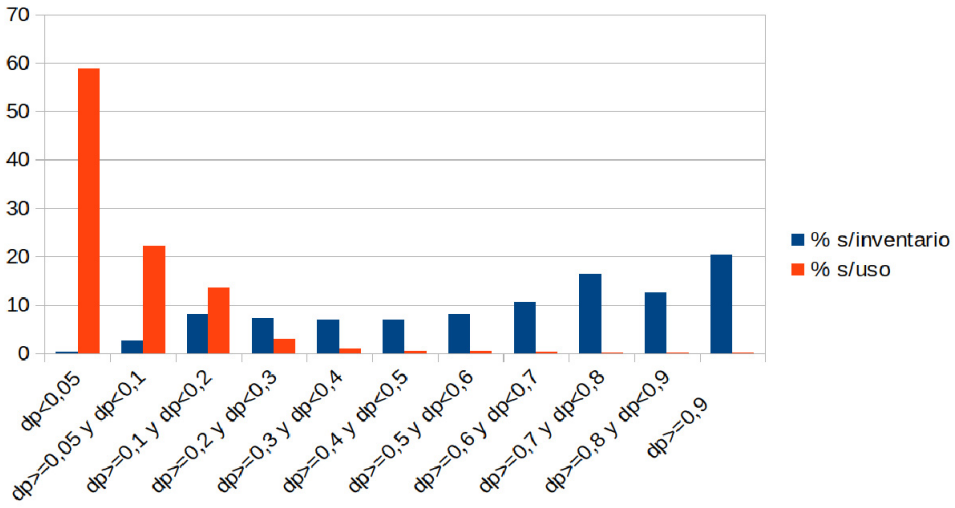

GRÁFICO 4. Porcentajes de uso e inventario de los diferentes tramos de DP del DF del CORPES. Fuente: DF del CORPES. Elaboración propia

Una visión congruente con la anterior es la que se obtiene si, desde una perspectiva parcialmente distinta, comparamos inventario y uso en función del número de países en los que se documenta cada uno de los lemas. Los datos aparecen en la tabla 9.

Número de países	% inventario	% uso
1	34,80	0,052
2	11,95	0,039
3	6,79	0,038
4	4,93	0,040
5	3,60	0,039
6	2,89	0,039
7	2,41	0,040
8	2,22	0,045
9	1,93	0,049
10	1,81	0,056
11	1,64	0,059
12	1,55	0,066
13	1,43	0,072
14	1,37	0,081
15	1,30	0,095
16	1,33	0,122
17	1,29	0,131
18	1,40	0,178

19	1,58	0,252
20	2,22	0,530
21	11,56	97,980

TABLA 9. Porcentajes de inventario y uso por número de países en los que se documentan los lemas. Fuente: DF del CORPES. Elaboración propia

Tal como se podía esperar, los lemas que se documentan en un solo país constituyen el 34,80 % de los registrados en el corpus, pero su frecuencia conjunta no alcanza ni siquiera el 0,1 % de los usos. En el extremo contrario, los lemas que se documentan en todos los países (solo el 11,56 % del total) supone el 98 % de los usos.

En la misma línea de utilizar la distribución por número de países como un factor asociado a la DP podemos cruzar este rasgo con la configuración general de los lemas según su rango de frecuencia. Se trata, por tanto, de averiguar cuántos lemas se documentan en los 21 países en cada uno de los tramos de frecuencia consideramos habitualmente. Los datos, que aparecen en la tabla 10, son, me parece, sobradamente ilustrativos de lo que sucede.

Los x primeros lemas	Total de lemas en los 21 países	Porcentaje sobre los lemas del tramo
1000	1000	100,00
2000	2000	100,00
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
10000	9637	96,37
15000	12 588	83,92
20000	13 393	66,97
25000	13 486	53,94
50000	13 487	26,97
100000	13 487	13,49
116707	13 487	11,56

TABLA 10. Lemas documentados en los 21 países según tramos de frecuencia. Fuente: DF del CORPES. Elaboración propia

Solo 14 de los 5000 lemas (con clase) más frecuentes no se documentan en los 21 países considerados en el DF basado en el CORPES. Es decir, únicamente el

0,28 % del lemario básico del español carece de documentación en alguno de los subcorpus establecidos. Algo similar se puede decir de los 363 lemas que faltan en los 10 000 más frecuentes (el 3,63 %). Las consideraciones cruzadas revelan el juego existente en el léxico español entre lo correspondiente a la frecuencia de uso y a la frecuencia de inventario.

3.4. *Estadísticas gramaticales*

El hecho de utilizar datos procedentes de un corpus codificado, anotado morfosintácticamente y lematizado permite que lo inicialmente concebido como un simple diccionario de frecuencias léxicas facilite también explotaciones de carácter gramatical, aunque, como es lógico, solo hasta el punto en que la anotación añadida lo hace posible. En la facilidad con que se puede hacer esa explotación tiene un papel fundamental el carácter digital con que ha sido concebido este DF. Como se indica en el apartado 2, los lemas llevan también la indicación de la clase a la que pertenecen, de modo que, con una simple hoja de cálculo, una base de datos o directamente con algunos comandos del sistema operativo (*cf.* Rojo 2023) es muy sencillo obtener estadísticas tanto de inventario como de uso de los diferentes tipos de lemas.

Por ejemplo, la distribución por frecuencias de inventario de las clases léxicas es el que muestra la tabla 11:

Adjetivos	32 811
Adverbios	4686
Sustantivos comunes	65 077
Verbos	10 399

TABLA 11. Frecuencia de inventario de las *clases léxicas.*
Fuente: DF del CORPES. Elaboración propia

Dado que el DF incluye las formas adscritas a cada lema y la etiqueta que les corresponde, es igualmente sencillo obtener la frecuencia de cada tipo de forma, siempre, como es lógico, con el límite de la etiquetación incorporada y la utilización de la abreviatura correspondiente a la etiqueta, como muestra la tabla 12:

Sust. en fem.	52385	Sust. fem. sing.	34854
		Sust. fem. plural	17531
Sust. en masc.	66918	Sust. masc. sing.	40159
		Sust. masc. plural	26759

TABLA 12. Frecuencias de inventario de algunas *subcategorías gramaticales.*
Fuente: DF del CORPES. Elaboración propia

Del mismo modo, con recuentos sencillos de formular, es posible obtener, por ejemplo, la frecuencia de las diferentes formas verbales, etc.

4. CONCLUSIÓN

Los rasgos y resultados analizados en los apartados anteriores dan una idea suficiente de las características y posibilidades de explotación del DF basado en los textos de prensa incluidos en la versión 1.0 del CORPES. Sus evidentes ventajas sobre los DF anteriores proceden, sin duda, de que ha sido concebido como un diccionario en formato electrónico. La liberación de las servidumbres de los formatos impresos permite facilitar datos de todos los lemas y formas existentes en el corpus y también información acerca de su distribución según categorías y subcategorías gramaticales. Por otro lado, el formato TSV, abierto, de los datos permite la obtención de una amplia gama de resultados distintos mediante la utilización de aplicaciones informáticas de uso generalizado, como hojas de cálculo o bases de datos, o de comandos sencillos incluidos en los sistemas operativos más utilizados.

CORPUS MENCIONADOS EN EL TEXTO

[CdEhist]: Corpus del Español (Género / Histórico). Dir. Mark Davies, <https://www.corpusdelespanol.org/hist-gen/>.

[CdEweb]: Corpus del Español (Web / Dialectos). Dir. Mark Davies, <https://www.corpusdelespanol.org/web-dial/>.

[CODICACH]: Corpus Dinámico del Castellano de Chile. Dir. Scott Sadowsky, <https://sadowsky.cl/codicach-es.html>.

[CORGA]: Centro Ramón Piñeiro para a Investigación en Humanidades. Corpus do Galego Actual. Dirs. Guillermo Rojo, María Sol López Martínez y Vítor Míguez Rego, < https://corpus.cirp.es/corga/>.

[CORPES]: Real Academia Española. Corpus del Español del Siglo XXI, <http://rae.es/recursos/banco-de-datos/corpes-xxi>. Versiones 1.0, 1.1 y 1.2.

[DF del CORPES]: 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 2.3.

[LIFCACH]: Lista de Frecuencia de Palabras del Castellano de Chile. Dirs. Scott Sadowsky y Ricardo Martínez Gamboa, <https://sadowsky.cl/lifcach-es.html>.

REFERENCIAS BIBLIOGRÁFICAS

ALAMEDA, José Ramón y Fernando CUETOS (1995): *Diccionario de frecuencias de las unidades lingüísticas del castellano*, Oviedo, Universidad de Oviedo, 2 vols.

ALMELA PÉREZ, Ramón, Pascual CANTOS, Aquilino SÁNCHEZ, Ramón SARMIENTO y Moisés ALMELA (2005): *Frecuencias del español: Diccionario y estudios léxicos y morfológicos*, Madrid, Universitas.

BIBER, Douglas, Randi REPPEN, Erin SCHNUR y Romy GHANEM (2016): «On the (non)utility of Juilland's D to measure lexical dispersion in large corpora», *International Journal of Corpus Linguistics*, 21, pp. 439-464.

BONTRAGER, Terry (1991): «The development of Word Frequency Lists Prior to the 1944 Thorndike-Lorge List», *Reading Psychology*, 12, 2, pp. 91-116.

BREZINA, Vaclav (2018): *Statistics in Corpus Linguistics. A practical guide,* Cambridge, Cambridge University Press.

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

BUSTOS TOVAR, Eugenio de (1966): «Un nuevo recuento del vocabulario español», *Filología moderna*, 25-26, pp. 171-192.

BYBEE, Joan (2007): *Frequency of Use and the Organization of Language*, Oxford, Oxford University Press.

CARTWRIGHT, C. W. (1925): «A Study of the Vocabularies of Eleven Spanish Grammars and Fifteen Spanish Reading Texts», *The Modern Language Journal*, 10, 1, pp. 1-14.

CASTILLO FADIĆ, María Natalia (2021): *Léxico básico del español de Chile*, Santiago de Chile, Liberalia Ediciones.

DAVIES, Mark (2006): *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*, Nueva York-Londres, Routledge.

DAVIES, Mark y Kathy H. DAVIES (2018): *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*, Nueva York-Londres, Routledge. Presentado como segunda edición de Davies (2006).

EGBERT, Jesse, Brent BURCH y Douglas BIBER (2020): «Lexical dispersion and corpus design», *International Journal of Corpus Linguistics*, 25, 1, pp. 89-115.

EZQUERRA, Raimundo (1974): «Los diccionarios de frecuencia en español», *Boletín de la Asociación Europea de Profesores de Español*, 10, pp. 3-27.

EZQUERRA, Raimundo (1977): «Los diccionarios de frecuencia en español II». *Boletín de la Asociación Europea de Profesores de Español*, 16, pp. 43-54.

GARCÍA HOZ, Víctor (1953): *Vocabulario usual, vocabulario común y vocabulario fundamental*, Madrid, CSIC, Instituto «San José de Calasanz».

GRIES, Stefan Th. (2008): «Dispersions and adjusted frequencies in corpora», *International Journal of Corpus Linguistics*, 13, 4, pp. 403-437.

JUILLAND, Alphonse y Eugenio CHANG-RODRIGUez (1964): *Frequency Dictionary of Spanish Words*, La Haya, Mouton.

JUSTICIA, Fernando (1995): *El desarrollo del vocabulario. Diccionario de frecuencias*, Granada, Universidad de Granada.

KENISTON, Hayward (1920): «Common words in Spanish», *Hispania*, 3, 2, pp. 85-96.

KENISTON, Hayward (1933): *List of Spanish Words and Idioms*, Chicago, University of Chicago Press.

KENISTON, Hayward (1941): *A Standard List of Spanish Words and Idioms*, Boston, D. C. Heath and Co.

LIJFFIJT, Jefrey y Stefan Th. GRIES (2012): «Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", *International Journal of Corpus Linguistics*, 13:4 (2008), 403-437», *International Journal of Corpus Linguistics*, 17, 1, pp. 147-149

MILLER, Don (2020): «Analysing Frequency Lists», en M. Paquot y S. Th. Gries, eds., *A Practical Handbook of Corpus Linguistics*, Springer, pp. 77-97.

MILLER, Don y Douglas BIBER (2015): «Evaluating reliability in quantitative vocabulary studies. The influence of corpus design and composition», *International Journal of Corpus Linguistics*, 20, 1, pp. 30-53.

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

MULLER, Charles (1965): «Un dictionnaire de fréquence de l'espagnol moderne», *Zeitschrift für romanische Philologie*, 81, 5-6, pp. 476-483.

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, pp. 59-82.

NATION, I. S. P. (2016): *Making and Using Word Lists for Language Learning and Testing*, Amsterdam-Filadelfia, John Benjamins.

ROBLES-GARCÍA, Pablo (2020): «3K-LEx. Desarrollo y validación de una prueba de amplitud léxica en español», *Journal of Spanish Language Teaching*, 7, 1, pp. 64-76.

RODRÍGUEZ BOU, L. (1952): *Recuento de vocabulario español*, Río Piedras, Universidad de Puerto Rico.

ROJO, Guillermo (2011): «Frecuencia de inventario y frecuencia de uso», *Revista Española de lingüística*, 41, 1, pp. 5-43.

ROJO, Guillermo (2022): «Sobre algunos rasgos estadísticos del léxico de la lengua oral», en Carmen Díaz Alayón, coord., *Studia philologica in honorem José Antonio Samper*, Madrid, Academia Canaria de la Lengua-Arco/Libros, pp. 877-894.

ROJO, Guillermo (2021): *Introducción a la lingüística de corpus en español*, Londres-Nueva York, Routledge.

ROJO, Guillermo (2023): *Análisis informatizado de textos*, Santiago de Compostela, Universidade de Santiago de Compostela (Servizo de Publicacións).

ROJO, Guillermo (en prensa a). «Un breve apunte sobre las frecuencias léxicas».

ROJO, Guillermo (en prensa b). «¿Cuántos lemas hay que dominar para entender la prensa uruguaya?».

ROJO SASTRE, Antonio José, Paul RIVENC y Adán FERRER (1969): *Vida y diálogos de España*, París, Didier, 4 vols.

SCHMITT, Norbert, Xiangying JIANG y William GRAHE (2011): «The percentage of words known in a text and reading comprehension», *Modern Language Journal*, 95, 1, pp. 26-43.

SEBASTIÁN GALLÉS, Nuria, coord., M.ª Antònia MARTÍ ANTONÍN, Manuel Francisco CARREIRAS VALIÑA, Fernando CUETOS VEGA (2000): *Léxico informatizado del español*, Barcelona, Universitat de Barcelona.

THORNDIKE, Edward L. e Irving Lorge (1944): *The teacher's word book of 30,000 words*, Nueva York, Teachers College Press.

ZIPF, George Kingsley (1935): *The Psycho-biology of Language. An Introduction to the Dynamic Philology*, Cambridge (MA), MIT Press.

ZIPF, George Kingsley (1949): *Human Behaviour and the Principle of Least-Effort,* Cambridge (MA), Addison-Wesley.