Un breve apunte sobre las frecuencias léxicas*

Guillermo Rojo

Universidade de Santiago de Compostela / Real Academia Española

La lingüística de corpus (LC) ha supuesto una importante revolución metodológica en los estudios lingüísticos, que han modificado fuertemente sus modos de trabajo incluso en las visiones más formalistas, tradicionalmente alejadas del análisis directo de los actos lingüísticos realizados. Una de las características más destacadas de la LC es, sin duda, su carácter marcadamente cuantitativo, centrado en los últimos años en estudiar las diferentes frecuencias con las que un elemento o fenómeno se presenta en distintos corpus o subcorpus. Resulta por ello realmente curioso que los diccionarios de frecuencias (DF) no hayan experimentado un cambio que refleje y explote las nuevas opciones que surgen de la posibilidad de obtener datos procedentes de conjuntos formados por cientos o miles de millones de palabras y que habitualmente están aceptablemente codificados y anotados. En Rojo (2023) se muestra que los DF posteriores a la publicación del *Frequency Dictionary of Spanish Words* (Juilland y Chang-Rodríguez 1964) superan las deficiencias que presenta esta obra en el volumen de corpus manejado (solo 500 000 palabras), pero no lo mejoran o incluso resultan inferiores a él en algunos otros aspectos. A mi modo de ver, este defecto procede de la insuficiente explotación de las posibilidades que brindan los corpus y, asociado a ella, el peso que, con independencia de los formatos de salida, todavía tiene la versión impresa de los DF. En los párrafos siguientes trataré de mostrar los cambios que produce la explotación de los corpus textuales.

Si nos restringimos al módulo léxico, los DF y las listas de frecuencias han sido utilizados fundamentalmente para establecer el volumen y la secuencia de introducción del léxico en los cursos de lengua extranjera. Desde esta perspectiva, la “utilidad” de un lema se mide básicamente en la cobertura que proporciona en los textos, esto es, en el porcentaje de un texto que se puede entender si se conoce el significado de un cierto conjunto de lemas. En este sentido, se alude habitualmente a las leyes de Zipf y Pareto, con las que se explica que en cualquier texto o conjunto de textos hay siempre unas pocas palabras que tienen frecuencia muy alta, muchas palabras con frecuencias bajas o muy bajas y también un alto porcentaje que, independientemente del tamaño del corpus manejado, aparecen solo una vez (son hápax). Tomando los datos del DF del español incluido en la versión 1.1 del CORPES, se comprueba que con solo 50 lemas (caracterizados

* **Borrador final.** Enviado para su publicación en el homenaje a Johannes Kabatek (*Discursos tradicionales y particulares. Festschrift für Johannes Kabatek zum 60. Geburtstag*. Madrid/Frankfurt: Iberoamericana/Vervuert, 2025).

también por la clase de palabras) se da cuenta del 50 % de todo el corpus manejado (unos 180 millones de palabras procedentes de textos de prensa). Naturalmente, esa observación va seguida inmediatamente de la restricción esperable: los lemas más frecuentes son artículos, preposiciones, conjunciones, etc., de modo que la comprensión de los textos que se puede lograr con el conocimiento de su significado es más bien escasa. En la bibliografía referida al inglés, hay acuerdo general en considerar que para alcanzar el reconocimiento del 98 % de los lemas presentes en un texto (se entiende que sin tener en cuenta nombres propios, cifras y algunos otros tipos de elementos) se requiere el dominio de unas 8000 o 9000 familias de lemas.¹ Los datos de los que dispongo para español y gallego, referidos a lemas y con inclusión de la clase de palabras son los que figuran en la tabla 1.

Tabla 1: Porcentajes de lemas en diferentes corpus de español y gallego

Los primeros x lemas	Todo el lemarío			Adjetivos, adverbios, sustantivos y verbos		
	CORPES 1.0 (prensa)	CORGA 4.1 (2001_2022)	ESLORA 2.2	CORPES 1.0 (prensa)	CORGA 4.1 (2001-2022)	ESLORA 2.2
10	38,82	38,65	30,78	10,56	12,13	30,06
25	47,81	48,27	46,93	14,6	17,68	42,64
50	52,58	54,04	60,85	18,81	22,86	54,09
100	57,24	59,45	72,9	24,9	29,12	63,79
500	71,36	72,75	88,68	47,57	49,84	82,11
1000	78,85	79,34	92,83	60,81	61,75	88,16
2000	86,24	86,19	96	74,19	74,29	93,24
3000	90,01	89,71	97,43	81,15	80,83	95,63
4000	92,37	91,92	98,26	85,43	84,96	97,02
5000	93,86	93,45	98,78	88,31	87,83	97,94
6000	94,95	94,58	99,14	90,36	89,96	98,55
7000	95,76	95,45	99,41	91,9	91,58	99,05
8000	96,39	96,13	99,58	93,1	92,86	99,33
9000	96,9	96,68	99,72	94,07	93,89	99,61
10000	97,31	97,13	99,86	94,85	94,74	99,89
15000	98,54	98,54		97,21	97,39	
20000	99,12	99,23		98,33	98,67	
25000	99,44	99,6		98,92	99,34	
50000	99,89	100		99,8	100	
Total lemas	116 707	49 603	10 990	112 971	46 148	10 386
Hápax	29 172 (25 %)	7700 (15,52 %)	3719 (33,83 %)		6510 (14,10 %)	3606 (14,60 %)
Total subcorpus	184 108 153	25 831 840	728 038	92 013 231	12 643 355	358 032

Fuente: CORPES, CORGA y ESLORA. Elaboración propia

1 Las familias de lemas agrupan los relacionados mediante procedimientos gramaticales, como, por ejemplo, la derivación. Así *limpio*, *limpiar*, *limpieza*, *limpiador*, *limpiamente* y algunos otros son lemas distintos, pertenecientes a distintas clases de palabras, pero forman parte de la misma familia. Pues, bien, según indica Nation (2006, 79):

If we take 98% as the ideal coverage, a 8,000–9,000 word-family vocabulary is needed for dealing with written text, and 6,000–7,000 families for dealing with spoken text.

La parte izquierda de la tabla 1 muestra la coincidencia en este punto entre español y gallego en el número de lemas necesarios para alcanzar una cobertura del 95 % y el 98 %. En cambio, en ESLORA, formado por textos orales, las cifras respectivas bajan a 2000 y 4000 lemas, respectivamente. Las cifras son válidas en sí mismas, pero es necesario tener mucha precaución en la comparación, puesto que son corpus que tienen características de composición y tamaño bastante diferentes.

En la tabla se muestra también un aspecto al que, precisamente por las insuficiencias del modelo clásico de DF, no se ha prestado atención suficiente. Si nos limitamos a excluir de nuestra consideración elementos como nombres propios, cifras, fechas y similares, los porcentajes de cobertura arrojarán irremediabilmente resultados en los que los elementos de significado gramatical seguirán teniendo un gran peso como consecuencia de la alta frecuencia que tienen muchos de ellos. No es seguro, pues, que la imagen que proyectan estos resultados sobre la comprensión del contenido de un texto (en realidad, del reconocimiento de los lemas) refleje realmente el funcionamiento de ese módulo. Podemos intentar evitar esos efectos distorsionadores reduciendo nuestra consideración a los lemas que tienen significado léxico en el sentido más estricto, es decir, sustantivos, verbos, adjetivos y adverbios.

Eso es lo que muestran las columnas de la parte derecha de la tabla 1. En los tres corpus considerados la frecuencia de uso de los lemas pertenecientes a estas cuatro clases se sitúa en torno al 50 % del total, pero su frecuencia de inventario (*cf.* Rojo 2011) ronda el 95 %. Esas cifras son sumamente significativas y adelantan lo que sucede con la cobertura: en los textos de prensa del CORPES, el reconocimiento del 98 % de los lemas de estas cuatro clases requiere unos 18 500, que es un 150 % de los necesarios para todo el lemario (12 500). Algo parecido sucede en el CORGA y la cifra asciende a unos 5000 en el caso de ESLORA.

Esta diferencia, que considero relevante, se basa en el carácter diferente de las unidades léxicas consideradas, pero también a un hecho relacionado con la estructura estadística de los textos al que no hemos prestado la atención necesaria hasta ahora. Cualquier comparación de los lemas contenidos en una serie de textos muestra que los porcentajes de los que aparecen en todos ellos es bastante baja. En un experimento que me parece realmente demostrativo, Miller y Biber compararon los lemas contenidos en diez textos universitarios de introducción a la psicología utilizados en universidades estadounidenses. El corpus produce un total de 32 598 lemas distintos (sin contar nombres propios, como es lógico). A pesar de la identidad de tema, objetivo y nivel, los presentes en todos los textos son el 13 % y el 41 % se encuentra en solo uno de los textos. Es lógico pensar que esa distribución tan escasamente homogénea procede, sobre todo, de elementos

pertenecientes a estas cuatro clases.

El que hemos tocado en esta breve nota de homenaje a nuestro querido colega es uno de los varios aspectos que debemos reformar para que las frecuencias léxicas den resultados realmente útiles en la enseñanza de lenguas extranjeras y en nuestra comprensión general del funcionamiento de las lenguas.

Corpus mencionados en el texto

CORGA: Centro Ramón Piñeiro para a Investigación en Humanidades. Corpus do Galego Actual.

Dir. Guillermo Rojo y María Sol López Martínez, <<http://www.cirp.gal/corga/>>. Versión 4.1.

CORPES XXI: Real Academia Española y Asociación de Academias de la lengua española. Corpus del Español del Siglo XXI, <<http://rae.es/recursos/banco-de-datos/corpes-xxi>>. Versión 1.1.

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

Referencias bibliográficas

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

Miller, Don y Douglas Biber, 2015. "Evaluating reliability in quantitative vocabulary studies. The influence of corpus design and composition". *IJCL*, 20/1, 30-53.

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

Rojo, Guillermo, 2011. "Frecuencia de inventario y frecuencia de uso", *Revista Española de lingüística*, 41/1, 5-43.

Rojo, Guillermo, 2023. "Hacia un nuevo concepto de diccionario de frecuencias". En Dolores Corbella, Josefa Dorta, Rafael Padrón (eds.): *Perspectives de recherche en linguistique et philologie romanes (I et II)*. París: Éditions de Linguistique et de Philologie, Bibliothèque de Linguistique Romane (BiLiRo), n° 18/1, 45-63.