Diccionario de frecuencias léxicas de ESLORA

Índices de uso

Sin que ello implique infravalorar su importancia, es claro que la frecuencia global y la frecuencia normalizada de un lema no permiten captar totalmente el papel de los elementos léxicos en la comprensión de los textos. El grado en que los lemas se distribuyen en los textos y sus diferentes tipos, esto es, su dispersión, es una medida mucho más reveladora de su relevancia. El importante desarrollo que han experimentado las técnicas estadísticas utilizadas en lingüística de corpus (LC) ha tenido efecto también, como era de esperar, en los índices de uso utilizados en el análisis de los inventarios léxicos. Puede verse un amplio y detallado resumen, que incluye valoraciones de los más diferentes aspectos, en Egbert, Burch y Biber (2020). Para el cálculo de los índices de uso en ESLORA que se incorpora a la versión 2.2, hemos decidido utilizar el basado en la diferencia de proporciones (DP) propuesto por Gries (2008). Se trata de un estadístico muy intuitivo, fácil de calcular y capaz de trabajar con cualquier número de subcorpus de diferentes tamaños.

En muy pocas palabras, el método de la diferencia de proporciones (DP) se basa en la comparación de la frecuencia que presenta un elemento en cada uno de los subcorpus establecidos con la frecuencia esperada en función del porcentaje que ese subcorpus representa sobre el subcorpus total. La suma de esas diferencias arroja una cifra comprendida habitualmente entre 0 y 1. Los elementos distribuidos de modo más irregular presentan DP más cercanas a 1, mientras que los que se distribuyen de forma más homogénea presentan índices más próximos a 0.

Las operaciones necesarias para obtener estos índices son las siguientes:

- En primer lugar es necesario decidir el tamaño y la composición de los subcorpus, teniendo en cuenta que este procedimiento puede trabajar con cualquier número de subcorpus y que los tamaños no tienen que ser ni siquiera similares.
- En el segundo paso, hay que calcular la proporción que supone cada subcorpus sobre el total del corpus.[1]
- Obtener las frecuencias que cada elemento presenta en cada uno de los subcorpus (frecuencias observadas) y la proporción que cada una supone sobre la frecuencia total del elemento.
- Obtener las frecuencias esperadas de cada elemento en cada uno de los subcorpus y la proporción que cada una supone sobre la frecuencia total del elemento.
- Obtener la diferencia entre la proporción observada y la proporción esperada de cada elemento en cada subcorpus.
- Sumar las diferencias y dividir entre 2.

Para esta primera aproximación a los índices de uso en el corpus ESLORA, hemos organizado los 83 ficheros que contienen los textos anotados morfosintácticamente en diez subcorpus con la distribución que figura en la Tabla 1. Todas las conversaciones están integradas en el Subcorpus 1. Las entrevistas se distribuyen en los nueve subcorpus restantes sin tener en cuenta los parámetros que intervienen en la configuración del corpus (edad, sexo y nivel sociocultural). De los listados de formas se han eliminado los signos ortográficos, las etiquetas, los nombres propios y los elementos que no han podido ser lematizados en el proceso de anotación automática.

1 Si resulta preferible, se puede trabajar con porcentajes. Naturalmente, en ese caso los índices oscilarán entre 0 y 100.

Tabla 1*:* Composición y tamaño de los diez subcorpus de ESLORA[2]

		Tamaño	Proporción subcorpus_x/ ESLORA
SCOM_C_[0-3][0-9]	subcorpus1	108 108	0,148
SCOM*00[1-6]	subcorpus2	72 851	0,100
SCOM*00[7-9] + SCOM*01[0-2]	subcorpus3	72 623	0,099
SCOM*01[3-9]	subcorpus4	71 898	0,098
SCOM*02[0-5]	subcorpus5	68 926	0,094
SCOM*02[6-9] + SCOM*03[0-1]	subcorpus6	67 074	0,092
SCOM*03[2-7]	subcorpus7	62 637	0,086
SCOM*03[89] + SCOM*04[0-2]	subcorpus8	73 738	0,101
SCOM*04[3-7]	subcorpus9	68 586	0,094
SCOM*04[89] + SCOM*05[0-9]	subcorpus10	61 708	0,084
Totales		728 149	1

En el paso siguiente, extrajimos los lemas contenidos en cada uno de esos subcorpus y calculamos la frecuencia de cada uno de ellos en los diferentes subcorpus para, de esta forma, poder aplicar el método de la diferencia de proporciones observada y esperada.

En las tablas 2, 3 y 4 se pueden observar, como ilustración del procedimiento aplicado, los datos correspondientes al cálculo del DP de tres lemas distintos: *paciente*, *tren* y *tintorería*. Los dos primeros tienen una frecuencia similar, pero presentan una distribución diferente y, por tanto, arrojan DP distintas. El último es un caso especial: muestra una frecuencia relativamente alta (19 casos), pero con la peculiaridad de que están todos en el mismo texto y, por tanto, en un único corpus. En las columnas 4 y 5 se dan los datos observados. En las dos columnas siguientes, la frecuencia y la proporción esperada según el peso de cada uno de ellos. En la última, el valor absoluto de la diferencia entre la proporción observada y la esperada.

Tabla 2: Cálculo de la diferencia de proporciones para el sustantivo *paciente* (61 casos) en ESLORA

		Tamaño	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
SCOM_C_[0-3][0-9]	subcorpus1	108 108	0,148	0	0,000	9,057	0,148	0,148
SCOM*00[1-6]	subcorpus2	72 851	0,100	2	0,033	6,103	0,100	0,067
SCOM*00[7-9] + SCOM*01[0-2]	subcorpus3	72 623	0,099	38	0,623	6,084	0,100	0,523
SCOM*01[3-9]	subcorpus4	71 898	0,098	14	0,230	6,023	0,099	0,131
SCOM*02[0-5]	subcorpus5	68 926	0,094	0	0,000	5,774	0,095	0,095
SCOM*02[6-9] + SCOM*03[0-1]	subcorpus6	67 074	0,092	0	0,000	5,619	0,092	0,092
SCOM*03[2-7]	subcorpus7	62 637	0,086	0	0,000	5,247	0,086	0,086
SCOM*03[89] + SCOM*04[0-2]	subcorpus8	73 738	0,101	0	0,000	6,177	0,101	0,101
SCOM*04[3-7]	subcorpus9	68 586	0,094	6	0,098	5,746	0,094	0,004
SCOM*04[89] + SCOM*05[0-9]	subcorpus10	61 708	0,084	1	0,016	5,170	0,085	0,068
	Total	728 149	1	61	1	61	1	1,316
							Div. por 2	0,658

2 Los totales que figuran en las dos últimas columnas eliminan signos de puntuación, marcas de pausa, nombres propios y elementos que no han podido ser lematizados.

En el caso de *paciente* (sustantivo), se observa enseguida que hay cinco subcorpus en los que este lema no se documenta y también que la mayor parte de sus apariciones (el 83 %) se concentra en los subcorpus 3 y 4 (que solo suponen un 20 % del corpus completo). Por ejemplo, en el primero de ellos se encuentran 36 casos, lo cual significa una proporción de 0,59 sobre el total, mientras que la proporción esperada es de 0,1 y, por tanto, si la distribución fuera homogénea aquí debería haber únicamente 6 o 7 casos. Parece claro que la dispersión de este lema es bastante alta y eso es lo que indica el valor que le atribuye el método de la diferencia de proporciones: 0,658.

Tabla 3: Cálculo de la diferencia de proporciones para el sustantivo *tren* (67 casos) en ESLORA

		Tamaño	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
SCOM_C_[0-3][0-9]	subcorpus1	108 108	0,148	13	0,191	10,096	0,148	0,043
SCOM*00[1-6]	subcorpus2	72 851	0,100	4	0,059	6,803	0,100	0,041
SCOM*00[7-9] + SCOM*01[0-2]	subcorpus3	72 623	0,099	4	0,059	6,782	0,100	0,041
SCOM*01[3-9]	subcorpus4	71 898	0,098	6	0,088	6,714	0,099	0,011
SCOM*02[0-5]	subcorpus5	68 926	0,094	9	0,132	6,437	0,095	0,038
SCOM*02[6-9] + SCOM*03[0-1]	subcorpus6	67 074	0,092	6	0,088	6,264	0,092	0,004
SCOM*03[2-7]	subcorpus7	62 637	0,086	6	0,088	5,850	0,086	0,002
SCOM*03[89] + SCOM*04[0-2]	subcorpus8	73 738	0,101	11	0,162	6,886	0,101	0,060
SCOM*04[3-7]	subcorpus9	68 586	0,094	9	0,132	6,405	0,094	0,038
SCOM*04[89] + SCOM*05[0-9]	subcorpus10	61 708	0,084	0	0,000	5,763	0,085	0,085
	Totales	728 149	1	68	1	68	1	0,363
							Div. por 2	0,181

En el caso de *tren*, en cambio, el lema se documenta en 9 de los 10 subcorpus y las diferencias entre la proporción observada y la esperada es relativamente baja en todos ellos. Por tanto, el índice de uso atribuido por el método de la diferencia de proporciones es de únicamente 0,181, bastante cerca del 0 que indicaría una distribución totalmente homogénea de este lema.

Como demostración del modo en que este procedimiento trata los casos extremos de concentración de uso podemos ver las cifras correspondientes a *tintorería*, que muestra un comportamiento peculiar. Sus 21 casos se concentran en un documento único (y, por tanto, solo en uno de los subcorpus). El resultado obtenido se aproxima al valor 1.

Tabla 4: Cálculo de la diferencia de proporciones para el sustantivo *tintorería* (21 casos) en ESLORA

		Tamaño	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
SCOM_C_[0-3][0-9]	subcorpus1	108 108	0,148	0	0,000	2,821	0,148	0,148
SCOM*00[1-6]	subcorpus2	72 851	0,100	0	0,000	1,901	0,100	0,100
SCOM*00[7-9] + SCOM*01[0-2]	subcorpus3	72 623	0,099	0	0,000	1,895	0,100	0,100
SCOM*01[3-9]	subcorpus4	71 898	0,098	0	0,000	1,876	0,099	0,099
SCOM*02[0-5]	subcorpus5	68 926	0,094	0	0,000	1,799	0,095	0,095
SCOM*02[6-9] + SCOM*03[0-1]	subcorpus6	67 074	0,092	0	0,000	1,750	0,092	0,092
SCOM*03[2-7]	subcorpus7	62 637	0,086	19	1,000	1,634	0,086	0,914
SCOM*03[89] + SCOM*04[0-2]	subcorpus8	73 738	0,101	0	0,000	1,924	0,101	0,101
SCOM*04[3-7]	subcorpus9	68 586	0,094	0	0,000	1,790	0,094	0,094
SCOM*04[89] + SCOM*05[0-9]	subcorpus10	61 708	0,084	0	0,000	1,610	0,085	0,085
	Total	728 149	1	19	1	19	1	1,828
							Div. por 2	0,914

Para interpretar adecuadamente la información proporcionada por la DP es preciso no olvidar que se basa en la diferencia entre las proporciones esperada y observada en los subcorpus establecidos y no directamente en la frecuencia. Esto significa, con un ejemplo muy claro, que en los casos en los que los lemas se registran únicamente en uno de los subcorpus, el DP correspondiente será el mismo con independencia de que la frecuencia sea igual a 1, 100 o 1000. En todos los casos, la proporción correspondiente a ese corpus será 1 y la de los otros nueve será 0. Por tanto, en casos de este tipo, el DP será el mismo en cada subcorpus. Y lo mismo ocurre en todos los casos en los que las proporciones se mantengan aunque las frecuencias sean diferentes (dos subcorpus con proporciones 0,3 y 0,7, etc.). Por la misma razón, la DP de un subcorpus con 0 casos es también el mismo con independencia de cuál sea la frecuencia total del lema en cuestión.

Los listados

La versión actual de ESLORA facilita dos listados relacionados con los diccionarios de frecuencia. El primero de ellos (lista_dp_eslora.csv) contiene los lemas documentados en ESLORA ordenados por frecuencia creciente de la DP (es decir, comenzando por los que tienen una dispersión menor). Presenta el aspecto siguiente:

Orden	Rango por frecuencia	Lema	Clase	DP	Frec.	Frec. norm.	Sub-corpus
1.	[1]	el	D	0.0242	49026	67329.63	10
2.	[28]	hacer	V	0.0243	5951	8172.78	10
3.	[7]	a	X	0.0285	18294	25123.98	10
4.	[6]	que	C	0.0314	19622	26947.78	10
5.	[8]	un	D	0.0332	14701	20189.55	10
6.	[95]	ni	C	0.0360	1148	1576.60	10
7.	[19]	te	P	0.0360	6840	9393.68	10
8.	[15]	ir	V	0.0375	8925	12257.11	10
9.	[3]	de	X	0.0398	24398	33506.88	10
10.	[2]	y	C	0.0419	29519	40539.78	10

Al lado del número de orden según la DP se añade, entre corchetes, el rango del lema en el listado de frecuencia decreciente, para que se pueda hacer la comparación entre las dos caracterizaciones. Figuran luego el lema, la clase gramatical a la que pertenece,[3] la DP, la frecuencia total, la frecuencia normalizada y, en la última columna, el número de subcorpus en los que se documenta.

El segundo listado (diccionario_frecuencias_eslora.csv) está organizado al modo tradicional de los diccionarios de frecuencias léxicas, con los datos correspondientes al lema y luego los de cada una de las formas asociadas a él.

Lema	Elem. gram.	Clase/ Etiqueta	Frecuencia	Frec. norm.	DP	Subcorpus	[Lema]
abalanzar		V	2	2.75	0.8154	2	
	abalanzar	VNP	1	1.37	0.9004	1	[abalanzar]
	abalanzó	VIS3S	1	1.37	0.9141	1	[abalanzar]
abandonado		A	7	9.61	0.7281	3	
	abandonada	AFS	3	4.12	0.8120	2	[abandonado]
	abandonado	AAS	1	1.37	0.9138	1	[abandonado]
	abandonado	ANS	2	2.75	0.9138	1	[abandonado]
	abandonados	AMP	1	1.37	0.9138	1	[abandonado]
abandonar		V	12	16.48	0.3330	7	
	abandono	VIP1S	1	1.37	0.8547	1	[abandonar]
	abandonó	VIS3S	3	4.12	0.6528	3	[abandonar]
	abandonar	VNP	3	4.12	0.8070	2	[abandonar]
	abandoné	VIS1S	1	1.37	0.9012	1	[abandonar]
	abandonaba	VII3S	1	1.37	0.9141	1	[abandonar]
	abandona	VIP3S	1	1.37	0.9141	1	[abandonar]
	abandonado	VPAS	1	1.37	0.9138	1	[abandonar]
	abandonado	VPMS	1	1.37	0.9138	1	[abandonar]
abandono		N	1	1.37	0.8524	1	
	abandono	NCMS	1	1.37	0.8547	1	[abandono]

3 Para las claves utilizadas en las etiquetas, *cf.* **https://eslora.nlpgo.com/guide_tags**.

Ambos ficheros están en formato `csv`, con campos separados por tabuladores, de modo que se pueden manejar con cualquier herramienta informática de extracción de datos (como `grep` y similares), editor de texto, procesador de texto u hoja de cálculo.[4]

En este segundo listado, las formas están sangradas al estilo tradicional, de modo que la primera columna contiene lemas (o blancos), la segunda formas (o blancos), etc. Para permitir la recuperación, las filas que contienen formas incluyen también el lema entre corchetes, para diferenciarlos así de los lemas que están en sus líneas propias. Teniendo en cuenta la organización y el contenido de las columnas es posible recuperar datos referentes a elementos gramaticales (frecuencia de todos los verbos, de todos los adjetivos en femenino singular, de todos los copretéritos de indicativo de la segunda conjugación, etc.).

Referencias bibliográficas

Egbert, Jesse, Brent Burch y Douglas Biber (2020): “Lexical dispersion and corpus design”, en *International Journal of Corpus Linguistics,* 25/1 (2020), pp. 89–115.

Gries, S. Th. (2008): “Dispersions and adjusted frequencies in corpora”, en *International Journal of Corpus Linguistics*, 13/4, (2008), pp. 403–437.

4 En este caso, debe tenerse en cuenta que la separación entre la parte entera y la decimal se marca con un punto (.) para facilitar su tratamiento con herramientas informáticas generales. Es necesario, por tanto, asegurarse de que la hoja de cálculo manejada (Excel, Calc, etc.) está configurada de esta forma.