Dicionario de frecuencias léxicas baseado no CORGA

1 Índices de uso

Sen que iso implique infravalorar a súa importancia, é claro que a frecuencia global e a frecuencia normalizada dun lema non permiten valorar integramente o papel dos elementos léxicos na comprensión dos contidos. O grao en que os lemas se distribúen nos diferentes tipos de textos, isto é, a súa dispersión, é unha medida moito máis reveladora da súa relevancia. O importante desenvolvemento que experimentaron as técnicas estatísticas utilizadas en lingüística de corpus (LC) tivo efecto tamén, como era de esperar, nos índices de uso utilizados na análise dos inventarios léxicos. Pode verse un amplo e detallado resumo, que inclúe valoracións dos máis diferentes aspectos, en Egbert, Burch e Biber (2020). Para o cálculo dos índices de uso que se incorpora á versión 4.1 do CORGA, decidimos utilizar o baseado na diferenza de proporcións (DP) proposto por Gries (2008) e descrito tamén en Brezina (2018: 52–53). Trátase dun estatístico moi intuitivo, fácil de calcular e capaz de traballar con calquera número de subcorpus de diferentes tamaños.

En moi poucas palabras, o método da diferenza de proporcións (DP) baséase na comparación da frecuencia que presenta un elemento en cada un dos subcorpus establecidos coa frecuencia esperada en función da porcentaxe que ese subcorpus representa sobre o corpus total. A suma desas diferenzas presenta como resultado unha cifra comprendida habitualmente entre 0 e 1. Os elementos distribuídos de modo máis irregular revelan DP máis próximas a 1, mentres que os que se documentan de forma máis homoxénea mostran índices máis próximos a 0.

As operacións necesarias para obter estes índices son as seguintes:

- En primeiro lugar é necesario decidir o tamaño e a composición dos subcorpus, tendo en conta que este procedemento pode traballar con calquera número de subcorpus e que os tamaños non teñen que ser nin sequera similares.
- No segundo paso, hai que calcular a proporción que supón cada subcorpus sobre o total do corpus.[1]
- Obter as frecuencias que cada elemento presenta en cada un dos subcorpus (frecuencias observadas) e a proporción que cada unha supón sobre a frecuencia total do elemento.
- Obter as frecuencias esperadas de cada elemento en cada un dos subcorpus e a proporción que cada unha supón sobre a frecuencia total do elemento.
- Obter a diferenza absoluta entre a proporción observada e a proporción esperada de cada elemento en cada subcorpus.
- Sumar as diferenzas e dividir entre 2.[2]

2 Dicionario de frecuencias léxicas baseado no CORGA

Para esta primeira aproximación aos índices de uso no CORGA, seleccionamos unicamente os textos publicados ou producidos entre 2001 e 2021.

O corpus utilizado suma algo máis de 25 millóns de palabras unha vez eliminados os casos correspondentes a nomes propios, cifras, formas non lematizadas e outros elementos que non se inclúen habitualmente nos dicionarios de frecuencias léxicas. Todo ese material estruturouse en 10

1 Se resulta preferible, pódese traballar con porcentaxes. Naturalmente, nese caso os índices oscilarán entre 0 e 100.
2 *Vid.* os cálculos amosados nas táboas do apartado 2.

subcorpus diferentes, construídos mediante un proceso aleatorio de asignación. A distribución dos subcorpus é a que figura na táboa 1.

Táboa 1: Tamaño e proporción dos diferentes subcorpus obtidos do CORGA

Subcorpus	Número de formas	Prop. s/ corpus
	Tamaño	Prop. s/ corpus
subcorpus1	2587622	0,10017
subcorpus2	2583441	0,10001
subcorpus3	2581250	0,09993
subcorpus4	2578657	0,09983
subcorpus5	2569104	0,09946
subcorpus6	2589005	0,10023
subcorpus7	2582660	0,09998
subcorpus8	2602902	0,10077
subcorpus9	2586054	0,10011
subcorpus10	2570438	0,09951
Total	25831133	1

Extraemos os lemas e mais as formas documentadas neses subcorpus e calculamos a frecuencia dos diferentes elementos en cada un deles e logo, tendo en conta o volume de cada subcorpus e, polo tanto, as frecuencias observada e esperada, fixemos o cálculo da DP para cada lema e para cada forma gramatical.

Nas táboas seguintes pódense observar, como ilustración do procedemento aplicado, os datos correspondentes ao cálculo do DP de varios lemas diferentes, seleccionados para ilustrar, ademais do procedemento de cálculo, os DP resultantes do procesamento de lemas con diferentes frecuencias e distribucións. Nas tres primeiras columnas aparecen o subcorpus, o seu tamaño e a proporción sobre a totalidade. Nas columnas 4 e 5 danse os datos observados. Nas dúas columnas seguintes, a frecuencia e a proporción esperada segundo o peso de cada un deles. Na última, o valor absoluto da diferenza entre a proporción observada e a esperada.

Táboa 2: Frecuencia e distribución do hiperlema *ata* nos diferentes subcorpus do CORGA e cálculo do índice DP

Hiperlema *ata*	Tamaño subcorpus	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
subcorpus1	2587622	0,10017	3211	0,102	3147,284	0,100	0,002
subcorpus2	2583441	0,10001	2924	0,093	3142,199	0,100	0,007
subcorpus3	2581250	0,09993	3097	0,099	3139,534	0,100	0,001
subcorpus4	2578657	0,09983	3295	0,105	3136,380	0,100	0,005
subcorpus5	2569104	0,09946	2866	0,091	3124,761	0,099	0,008
subcorpus6	2589005	0,10023	3121	0,099	3148,966	0,100	0,001
subcorpus7	2582660	0,09998	3517	0,112	3141,249	0,100	0,012
subcorpus8	2602902	0,10077	3129	0,100	3165,869	0,101	0,001
subcorpus9	2586054	0,10011	3171	0,101	3145,377	0,100	0,001
subcorpus10	2570438	0,09951	3087	0,098	3126,383	0,100	0,001
Total	25831133	1	31418	1	31418	1	0,040
						DP	0,020

No caso do hiperlema *ata* atopamos a distribución esperable nun elemento que se documenta en todos os subcorpus e se distribúe de forma homoxénea, o cal explica que o índice DP obtido sexa

próximo a 0. O lema *parroquia* ten unha frecuencia bastante máis reducida e a súa distribución é moito menos homoxénea. En consecuencia, o índice DP que revela sitúase nunha zona media, como mostra a táboa 3:

Táboa 3: Frecuencia e distribución do lema *parroquia* nos diferentes subcorpus do CORGA e cálculo do índice DP

Lema *parroquia*	Tamaño	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
subcorpus1	2587622	0,10017	237	0,06735	352,514	0,100	0,033
subcorpus2	2583441	0,10001	235	0,06678	351,945	0,100	0,033
subcorpus3	2581250	0,09993	132	0,03751	351,646	0,100	0,062
subcorpus4	2578657	0,09983	1905	0,54135	351,293	0,100	0,442
subcorpus5	2569104	0,09946	201	0,05712	349,991	0,099	0,042
subcorpus6	2589005	0,10023	277	0,07872	352,703	0,100	0,022
subcorpus7	2582660	0,09998	188	0,05342	351,838	0,100	0,047
subcorpus8	2602902	0,10077	125	0,03552	354,596	0,101	0,065
subcorpus9	2586054	0,10011	115	0,03268	352,301	0,100	0,067
subcorpus10	2570438	0,09951	104	0,02955	350,173	0,100	0,070
Total	25831133	1	3519	1	3519	1	0,883
						DP	0,442

O lema *ganga* documéntase en 8 dos 10 subcorpus e amosa unha distribución non excesivamente discordante en 7 deles, pero a acumulación de casos no restante fai que o DP suba ata achegarse a 1, como amosa a táboa 4.

Táboa 4: Frecuencia e distribución do lema *ganga* nos diferentes subcorpus do CORGA e cálculo do índice DP

Lema *ganga*	Tamaño	Prop. s/ corpus	Frec. obs.	Prop. obs.	Frec. esp.	Prop. esp.	Dif.
subcorpus1	2587622	0,10017	4	0,01544	25,945	0,100	0,085
subcorpus2	2583441	0,10001	4	0,01544	25,903	0,100	0,085
subcorpus3	2581250	0,09993	1	0,00386	25,881	0,100	0,096
subcorpus4	2578657	0,09983	2	0,00772	25,855	0,100	0,092
subcorpus5	2569104	0,09946	8	0,03089	25,760	0,099	0,069
subcorpus6	2589005	0,10023	0	0,00000	25,959	0,100	0,100
subcorpus7	2582660	0,09998	1	0,00386	25,895	0,100	0,096
subcorpus8	2602902	0,10077	3	0,01158	26,098	0,101	0,089
subcorpus9	2586054	0,10011	0	0,00000	25,929	0,100	0,100
subcorpus10	2570438	0,09951	236	0,91120	25,773	0,100	0,812
Total	25831133	1	259	1	259	1	1,623
						DP	0,812

Para interpretar adecuadamente a información que proporciona o índice DP cómpre non esquecer que se basea na diferenza entre as proporcións esperada e observada nos subcorpus establecidos e non directamente na frecuencia. Isto significa, cun exemplo moi claro, que nos casos nos que os lemas se rexistran unicamente nun dos subcorpus, o DP correspondente será o mesmo para ese subcorpus con independencia de que a frecuencia sexa igual a 1, 100 ou 1000. En todos os casos, a proporción correspondente a ese corpus será 1 e a dos outros nove será 0. Polo tanto, en casos deste tipo, o DP será o mesmo en cada subcorpus. E o mesmo ocorre en todos os casos nos que as proporcións se manteñan aínda que as frecuencias sexan diferentes (dous subcorpus con proporcións 0,3 e 0,7 etc.). Pola mesma razón, a DP dun subcorpus con 0 casos é tamén o mesmo con independencia de cal sexa a frecuencia total do lema en cuestión.

3 As listaxes

A versión actual do CORGA facilita dúas listaxes relacionadas cos dicionarios de frecuencia. A primeira delas (**lista_dp_corga.csv**) contén os lemas documentados no CORGA ordenados por frecuencia crecente da DP (é dicir, comezando polos que teñen unha dispersión menor). Presenta o aspecto seguinte:

Núm. orde	Rango frec.	Lema/Hiperlema	Var. lema	Clase	Frecuencia	Frec. norm.	DP	Tot. subc.
1.	6	a		P	534501	20692124	0.00524	10
2.	13	por		P	257842	9981831	0.00675	10
3.	1	o		D	3607172	139644359	0.00705	10
4.	15	PARA		P	213582	8268395	0.00713	10
			pa	P	18	697	0.28822	8
			para	P	211880	8202505	0.00956	10
			pra	P	1684	65193	0.40116	10
5.	11	con		P	280348	10853105	0.00904	10
6.	4	E		C	757413	29321710	0.00949	10
			e	C	755935	29264493	0.00941	10
			y	C	1478	57218	0.18225	10
7.	14	seu		M	225061	8712781	0.01010	10
8.	3	en		P	781564	30256667	0.01016	10
9.	52	dous		N	38715	1498773	0.01100	10
10.	10	se		R	347186	13440603	0.01130	10
11.	2	de		P	2004656	77606197	0.01163	10
12.	22	todo		I	107444	4159477	0.01182	10
13.	5	un		D	584858	22641593	0.01182	10
14.	44	ENTRE		P	46392	1795972	0.01242	10
			antre	P	28	1084	0.35626	8
			entre	P	46364	1794888	0.01235	10
15.	7	que		T	499927	19353661	0.01263	10
16.	117	tres		N	17290	669347	0.01305	10
17.	25	poder		V	92985	3599726	0.01307	10
18.	20	máis		W	128262	4965404	0.01311	10
19.	54	MESMO		I	36887	1428006	0.01341	10
			mesmo	I	36692	1420456	0.01342	10
			mismo	I	195	7549	0.22404	10
[…]								
16557.	25754	aprazado		A	12	465	0.26790	8
16558.	6667	pousada		S	213	8246	0.26791	10
16559.	19551	embuste		S	27	1045	0.26794	10
16560.	20469	feudalismo		S	24	929	0.26794	9
16561.	21492	multipremiar		V	21	813	0.26794	9
16562.	21362	calcinado		A	21	813	0.26794	8
16563.	8488	fechado		A	142	5497	0.26799	10
16564.	16682	localista		A	39	1510	0.26799	10
16565.	13121	nudez		S	64	2478	0.26799	10
16566.	7664	lodo		S	169	6542	0.26800	10
16567.	16863	noventa e cinco		N	38	1471	0.26800	9
16568.	15730	xestante		A	44	1703	0.26804	10
16569.	10655	infindo		A	95	3678	0.26808	10
16570.	14066	culleriña		S	56	2168	0.26809	10
16571.	19752	pestanexo		S	26	1007	0.26814	9
16572.	16882	inversor		S	38	1471	0.26817	9
16573.	14082	aparvado		A	56	2168	0.26822	9

16574.	8376	real	S	145	5613	0.26826	10
16575.	15735	torrencial	A	44	1703	0.26826	10
16576.	16627	virtuosismo	S	39	1510	0.26834	10
16577.	7222	gramática	S	188	7278	0.26835	10
[…]							
30346.	29357	indixesto	A	8	310	0.60004	4
30347.	29247	nepalí	A	8	310	0.60004	4
30348.	29085	sensualmente	W	8	310	0.60004	4
30349.	31146	térreo	A	6	232	0.60004	4
30350.	28135	urda	S	9	348	0.60004	4
30351.	32488	vicioso	S	5	194	0.60004	4
30352.	30865	almofadón	S	7	271	0.60005	4
30353.	28748	antroponímico	A	9	348	0.60005	4
30354.	28645	colexiación	S	9	348	0.60005	4
30355.	31716	fanal	S	6	232	0.60005	4
30356.	33193	fanerógamo	A	5	194	0.60005	4
30357.	29425	foliar	A	8	310	0.60005	4
30358.	33091	ibis	S	5	194	0.60005	4
30359.	27543	micropartícula	S	10	387	0.60005	4
30360.	32537	trampullada	S	5	194	0.60005	4
30361.	27286	xenealoxista	S	10	387	0.60005	4

Ao lado do número de orde segundo a DP engádese, entre corchetes, o rango do lema na listaxe de frecuencia decrecente, para que se poida facer a comparación entre as dúas caracterizacións. Figuran logo o lema ou o hiperlema, as variantes do hiperlema, a clase gramatical á que pertence, a frecuencia total, a frecuencia normalizada, o índice DP e, na última columna, o número de subcorpus nos que se documenta cada elemento.

A segunda listaxe (**dicionario_frecuencias_corga.csv**) está organizada ao modo tradicional dos dicionarios de frecuencias léxicas, cos datos correspondentes ao lema e deseguido os de cada unha das formas pertencentes a el.

Núm. orde	Rango frec.	HIPERLEMA	Lema	Forma	Etiqueta	Frecuencia	Frec. norm.	DP	Núm. subc.
1	6		a		P	534501	20692,124	0,00524	10
			a	a	P	534501	20692,12	0,00524	10
2	49128		a		S	1	0,039	0,89982	1
			a	as	Scmp	1	0,04	0,89982	1
3	4890	Á			S	347	13,433	0,25199	10
			ala	ala	Scfs	65	2,52	0,1931	10
			ala	alas	Scfp	44	1,7	0,4368	10
			á	á	Scfs	2	0,08	0,80038	2
			áa	áas	Scfp	4	0,15	0,90049	1
			á	ás	Scfp	232	8,98	0,2985	10
4	49127		a a bartola		W	1	0,039	0,89977	1
			a a bartola	a a bartola	Wn	1	0,04	0,89977	1
5	18008		a a beira		W	33	1,278	0,206	10
			a a beira	a a beira	Wn	33	1,28	0,206	10
6	26391		a a beira de		P	12	0,465	0,34991	7
			a a beira de	a a beira de	P	12	0,46	0,34991	7
7	16799	A A CUSTA DE			P	39	1,51	0,32015	9
			a a custa de	a a custa de	P	34	1,32	0,39326	8
			a costa de	a costa de	P	5	0,19	0,60007	4

8	49125	a a envorca		W	1	0,039	0,89999	1
		a a envorca	a a envorca	Wn	1	0,04	0,89999	1
9	7480	a a fin		W	177	6,852	0,22673	10
		a a fin	a a fin	Wn	177	6,85	0,22673	10
10	6935	A A FIN E A O CABO		W	201	7,781	0,16674	10
		a a fin e a o cabo	a a fin e a o cabo	Wn	111	4,3	0,23022	10
		a o fin e a o cabo	a o fin e a o cabo	Wn	90	3,48	0,16598	10
[…]								
36173	2219	permiso		S	1034	40,029	0,08589	10
		permiso	permiso	Scms	835	32,33	0,08316	10
		permiso	permisos	Scmp	199	7,7	0,15198	10
36174	171	permitir		V	12056	466,724	0,03004	10
		permitir	permita	Vps30s	829	32,09	0,05667	10
		permitir	permitades	Vps20p	2	0,08	0,80056	2
		permitir	permitamos	Vps10p	2	0,08	0,79941	2
		permitir	permitan	Vps30p	543	21,02	0,13183	10
		permitir	permitas	Vps20s	6	0,23	0,60069	4
		permitir	permite	Vpi30s	3632	140,61	0,05361	10
		permitir	permite	V0m20s	2	0,08	0,80024	2
		permitir	permiten	Vpi30p	1272	49,24	0,04864	10
		permitir	permites	Vpi20s	24	0,93	0,2324	9
		permitir	permitiamos	Vii10p	3	0,12	0,80066	2
		permitir	permitiches	Vei20s	3	0,12	0,6995	3
		permitir	permitida	V0p0fs	51	1,97	0,16817	10
		permitir	permitidas	V0p0fp	31	1,2	0,14861	10
		permitir	permitide	V0m20p	29	1,12	0,327	9
		permitir	permitides	Vpi20p	21	0,81	0,26622	9
		permitir	permitido	V0p0ms	187	7,24	0,11098	10
		permitir	permitidos	V0p0mp	42	1,63	0,14788	10
		permitir	permitimos	Vpi10p	28	1,08	0,32014	8
		permitir	permitimos	Vei10p	9	0,35	0,49986	5
		permitir	permitindo	V0x000	283	10,96	0,11808	10
		permitir	permitir	V0f000	1214	47	0,04796	10
		permitir	permitira	Ves30s	33	1,28	0,10546	10
		permitir	permitira	Vli30s	32	1,24	0,16795	10
		permitir	permitira	Vli10s	1	0,04	0,90007	1
		permitir	permitiran	Ves30p	9	0,35	0,39966	6
		permitir	permitiran	Vli30p	22	0,85	0,23634	10
		permitir	permitirdes	V0f20p	2	0,08	0,80024	2
		permitir	permitiredes	Vfi20p	1	0,04	0,89989	1
		permitir	permitirei	Vfi10s	8	0,31	0,2998	7
		permitir	permitiremos	Vfi10p	13	0,5	0,36961	7
		permitir	permitiren	V0f30p	5	0,19	0,50016	5
		permitir	permitiriades	Vci20p	1	0,04	0,90007	1
		permitir	permitiriamos	Vci10p	1	0,04	0,89924	1
		permitir	permitirmos	V0f10p	5	0,19	0,59929	4
		permitir	permitiron	Vei30p	297	11,5	0,05459	10
		permitir	permitirá	Vfi30s	786	30,43	0,0437	10
		permitir	permitirán	Vfi30p	189	7,32	0,12403	10
		permitir	permitirás	Vfi20s	1	0,04	0,89989	1
		permitir	permitiría	Vci30s	318	12,31	0,06401	10
		permitir	permitiría	Vcia0s	2	0,08	0,79994	2

		permitir	permitirían	Vci30p	74	2,86	0,16211	10
		permitir	permitirías	Vci20s	1	0,04	0,90049	1
		permitir	permitise	Ves30s	183	7,08	0,11525	10
		permitir	permitise	Vesa0s	2	0,08	0,80066	2
		permitir	permitisen	Ves30p	60	2,32	0,1505	10
		permitir	permitises	Ves20s	2	0,08	0,80103	2
		permitir	permitistes	Vei20p	1	0,04	0,89999	1
		permitir	permitiu	Vei30s	1069	41,38	0,05871	10
		permitir	permito	Vpi10s	38	1,47	0,21003	10
		permitir	permitía	Vii30s	476	18,43	0,06947	10
		permitir	permitía	Vii10s	3	0,12	0,80051	2
		permitir	permitía	Viia0s	2	0,08	0,80038	2
		permitir	permitían	Vii30p	177	6,85	0,0575	10
		permitir	permitín	Vei10s	16	0,62	0,31243	8
		permitir	permitísemos	Ves10p	1	0,04	0,89924	1
		permitir	permitíu	Vei30s	12	0,46	0,45044	6
36175	24716	permuta		S	14	0,542	0,48602	6
		permuta	permuta	Scfs	9	0,35	0,59954	4
		permuta	permutas	Scfp	5	0,19	0,60027	4

Ambos os ficheiros están en formato `tsv`, con campos separados por tabuladores, de xeito que se poden manexar con calquera ferramenta informática de extracción de datos (como `grep` e similares), editores de texto, procesadores de texto, follas de cálculo ou bases de datos.[3] Por esa razón, aínda que o aspecto dos ficheiros poida resultar un tanto incómodo de entrada, a información está distribuída en columnas homoxéneas, co que é sinxelo recuperar a información referente a hiperlemas, lemas, formas e etiquetas.

4 Referencias bibliográficas

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

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

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

3 Neste caso, debe terse en conta que a separación entre a parte enteira e a decimal márcase con punto (.) para facilitar o seu tratamento con ferramentas informáticas xerais. É necesario, por tanto, asegurarse de que a folla de cálculo manexada (Excel, Calc etc.) está configurada con esta opción.