SOBRE ALGUNOS RASGOS ESTADÍSTICOS DEL LÉXICO DE LA LENGUA ORAL[*] [**]

Guillermo Rojo
Universidade de Santiago de Compostela

1 INTRODUCCIÓN

Los estudios acerca de la configuración estadística del léxico del español no son demasiado abundantes. Como he tratado de mostrar en otro lugar (Rojo: en prensa), los diccionarios de frecuencias publicados en los últimos años presentan deficiencias importantes. Una parte de ellas puede ser atribuida a que la finalidad perseguida se reduce casi siempre a su utilización para la enseñanza del español como lengua extranjera y otra se debe a los condicionamientos que impone el formato impreso, pero el gran problema es, sin duda, de tipo estructural. El léxico utilizado en los textos es el resultado de la conjunción de factores de muy diverso tipo, con lo que el único camino que nos puede llevar a la conversión de las frecuencias léxicas en un instrumento realmente útil para la investigación pasa por lograr un recurso dinámico, de modo que la recuperación de las listas de frecuencias sea sensible a los diversos parámetros que han sido tomados en cuenta en la configuración del corpus que contiene los materiales de trabajo (*cf*. Rojo: en prensa).

Además de este factor general, en el que no es posible entrar aquí, los intentos de aplicación de los estudios de frecuencias léxicas tropiezan con el hecho, tan claro como difícilmente salvable, de que los corpus utilizados como material de base están formados de modo exclusivo o mayoritario por textos escritos. De diferentes tipos y procedencias, por supuesto, pero básicamente escritos, con lo que no es seguro que las conclusiones alcanzadas en el análisis de materiales de ese tipo puedan ser aplicados también a los textos orales.

En este trabajo, dedicado a la memoria de José Antonio Samper y Clara Hernández, me propongo contribuir al mejor conocimiento de algunas de las características que presenta el léxico de la lengua oral mediante la comparación de las frecuencias obtenidas en un corpus de referencia (el Corpus del Español del Siglo XXI, CORPES) con las que proporciona el Corpus para el Estudio del Español Oral (ESLORA), formado exclusivamente por entrevistas semidirigidas y

* El presente trabajo ha sido realizado en el marco del proyecto de investigación “El corpus ESLORA de español oral: enriquecimiento, análisis lingüístico y aplicaciones” (ref. PID2020-118133GB-I00), financiado por el Ministerio de Ciencia e Innovación (2021-2024).

** Borrador final. Presentado para su publicación en el *Homenaje a José Antonio Samper* que editará la Academia Canaria de la Lengua.

conversaciones procedentes de hablantes de español de la zona de Santiago de Compostela (*cf.* Vázquez Rozas *et al.*: 2020).

2 CARACTERÍSTICAS GENERALES

En sentido estricto, la distinción entre lengua oral y lengua escrita se basa en el medio utilizado para establecer la comunicación y, por tanto, presenta características muy cambiantes. Un ensayo académico, una noticia de prensa y una carta particular son ejemplos de lengua escrita, pero es evidente que muestran rasgos muy diferentes. Por otro lado, una clase, una intervención en una tertulia radiofónica y una charla informal con amigos son ejemplos de lengua oral, pero sus características son también distintas. Sin embargo, la oposición habitual entre "lengua escrita" y "lengua oral" se asocia a la que se establece entre textos escritos formales y textos orales coloquiales, informales. Esta consideración se ha visto afectada en los últimos años por la aparición de nuevos géneros textuales vinculados a las nuevas formas de transmisión de información y, muy especialmente, Internet. Se trata de comunicaciones escritas (blogs, los mensajes electrónicos, los tuits, whatsapps, etc.) que presentan los rasgos de informalidad y coloquialismo propios de la comunicación oral. Aquí utilizaré "lengua oral" y expresiones semejantes en el sentido más habitual, atendiendo, sobre todo, a los rasgos presentes en textos como entrevistas semidirigidas o conversaciones informales que constituyen los materiales recogidos en el corpus ESLORA.

La primera cuestión que cabe plantear se refiere a la "riqueza" del léxico utilizado en los textos orales. Son habituales las alusiones a que el número de lemas distintos que figuran en los textos orales es bastante inferior al que se puede encontrar en los escritos, pero no se proporcionan habitualmente datos cuantitativos que permitan contrastar y valorar adecuadamente esas afirmaciones. Son bien conocidos los problemas que surgen en cualquier proceso de lematización, sea manual o automática. Es forzoso tomar decisiones con respecto a cómo tratar diferentes factores, algunos de los cuales proceden de cuestiones teóricas y otros, en cambio, son más bien de tipo operativo. ¿Se considera que los determinantes son una clase que luego se subdivide en demostrativos, posesivos, etc.? ¿Son los artículos una clase aparte o entran en los determinantes? ¿Qué clase corresponde a las denominaciones de los años? ¿Qué hacer con los numerales? ¿Cómo tratar fechas, cifras, etc.? Estas y muchas otras cuestiones relevantes son, como se puede apreciar, de muy distinta naturaleza, pero todas ellas inciden al final sobre los resultados obtenidos y, por tanto, en las dificultades inevitables cuando se intenta comparar lo que se puede extraer de corpus diferentes.

Comencemos por una aproximación muy general. Disponemos de los datos extraídos de un corpus de referencia, el CORPES XXI (versión 0.91), y de un corpus constituido exclusivamente por textos orales, ESLORA (versión 2.1). El CORPES constaba, en esa versión, de unos 300 millones de elementos gramaticales procedentes de textos de los más diferentes temas y tipos (incluido un porcentaje reducido de textos orales), generados en todos los países del ámbito hispánico. El corpus ESLORA consta de unos 900 000 elementos gramaticales procedentes de un total de 83 conversaciones y entrevistas semidirigidas que dan lugar a unas 80 horas de grabaciones de hablantes de español que residen en Santiago de Compostela o sus proximidades. Como es evidente que las diferencias de tamaño entre ambos corpus producen por sí mismas efectos que podrían dar lugar a malas interpretaciones de las divergencias encontradas, he construido una muestra reducida del CORPES constituida por un subconjunto de textos narrativos, ensayísticos y periodísticos con un volumen semejante al de ESLORA y formado únicamente por textos generados en España, para evitar la posible influencia de la diversidad dialectal en los inventarios léxicos.

La caracterización más general se refiere a aspectos relacionados con el número de lemas diferentes que se obtiene de cada uno de estos corpus. Para que las cifras resulten significativas en lo que a configuración del léxico se requiere, he eliminado de los recuentos los nombres propios.[1] Los resultados son los que se pueden observar en la tabla 1.

Tabla 1: Tamaño y número de lemas distintos en el CORPES, una muestra reducida del CORPES y ESLORA (sin nombres propios)

	CORPES 0.91	Muestra CORPES	ESLORA 2.1
Lemas distintos	125 285	19 863	10 613
Total elementos gramaticales	268 071 799	936 211	729 660

Fuentes: CORPES (0.91) y ESLORA (2.1). Elaboración propia.

Dejando a un lado de momento los resultados procedentes de la totalidad del CORPES, que tiene unas características muy diferentes en cuanto a tamaño y composición, la comparación entre la muestra del CORPES (textos escritos procedentes de España) y ESLORA da ya un indicio claro de las divergencias que podemos esperar. El número de lemas distintos del corpus escrito es casi el doble del que aparece en el corpus oral, aunque la diferencia en tamaño es solo de 200 000 elementos. Por buscar un índice que ilustre las diferencias, la proporción entre el total de elementos y el número de lemas distintos es de 43,13 en el caso de la muestra del CORPES y de 68,75 en ESLORA.[2]

1 Y también, por supuesto, los signos ortográficos, que es forzoso reconocer y etiquetar en el análisis automático de los textos.

2 Aunque esto no tenga mucho sentido, esta cifra es aproximable a la frecuencia media de cada uno de los lemas.

Como he indicado anteriormente, de estos recuentos están excluidos los nombres propios, pero se toma en cuenta la clase de palabras a la que es adscrito cada lema. Eso puede dar lugar a ligeras discrepancias en los recuentos y, sobre todo, trabaja con diferencias que es forzoso tomar en cuenta en la lematización general, pero que no tienen demasiada utilidad cuando lo que se busca está reducido a la cara puramente léxica. Diferenciar entre los casos en los que *pontevedrés* se comporta como un adjetivo o un sustantivo o entre los usos adjetivos y adverbiales de *fácil* no resulta de utilidad a la hora de considerar el conocimiento del léxico, su distribución, etc. La realización de los recuentos sin considerar la clase de palabras atribuida a los lemas arroja los resultados que figuran en la tabla 2:

Tabla 2: Tamaño y número de lemas distintos en el CORPES, una muestra reducida del CORPES y ESLORA (sin diferenciación de clase)

	CORPES 0.91	Muestra CORPES	ESLORA 2.1
Lemas distintos	112 711	18 619	9981
Total elementos gramaticales	268 071 799	936 211	729 660

Fuentes: CORPES y ESLORA. Elaboración propia.

La proporción es ahora 50,28 en el caso del CORPES y 73,10 en ESLORA. Las diferencias, pues, se mantienen aproximadamente como en la consideración previa.

Está claro que la variedad de elementos léxicos es más reducida en el caso del corpus oral que en el corpus escrito, lo cual apunta a una mayor concentración de elementos en el primer caso. Un modo razonable de averiguar en qué medida se produce este fenómeno e incorporarle consideraciones adicionales consiste en poner en relación las cifras relacionadas con el inventario y con el uso, en el sentido señalado en Rojo (2011). El número de lemas distintos documentados en un corpus es lo que proporciona la frecuencia de inventario y la suma total de sus apariciones produce la frecuencia de uso.[3] Para cada lema individual, la frecuencia de inventario es, evidentemente, 1 y la frecuencia de uso corresponde al número de veces que ese lema aparece en el conjunto de los textos. Naturalmente, los lemas tienen frecuencias distintas y puede resultar interesante investigar si en este punto se producen diferencias significativas entre los textos escritos y los textos orales.

Como es bien sabido, la distribución estadística de los elementos de un texto o un corpus responden a la llamada ley de Zipf, según la cual la frecuencia de los distintos elementos (lemas,

3 La frecuencia de inventario de los sustantivos de un texto es, por tanto, el número de sustantivos diferentes que contiene. La frecuencia de uso es la cantidad total de sustantivos. En casos de este tipo, esta distinción es formulable también como la existente entre *types* y *tokens*. Para detalles, *cf.* Rojo (2011).

palabras ortográficas, fonemas, esquemas sintácticos, etc.) resulta de una constante de acuerdo con la cual la frecuencia del elemento que ocupa el rango 2 es la mitad de la que tiene el que ocupa el rango 1, la del elemento que aparece en el rango 3 es la tercera parte del primero, etc. Es decir, la frecuencia esperable en un elemento que ocupe el rango *n* es la que presenta el primer elemento partido por *n*. De ahí se deducen tres consecuencias claras (*cf.* Nation 2016: 4; Rojo 2021: 131). La primera es que existe un número reducido de elementos que presentan frecuencias muy altas. En segundo lugar, hay un número muy elevado de elementos que presentan frecuencias bajas o muy bajas. Por último, como un caso especial del punto anterior, existe una cantidad importante de elementos que tienen frecuencia igual a 1 (los hápax).

Si tenemos en cuenta ambos factores, podemos analizar el peso relativo de los lemas mediante la comprobación de qué porcentaje del total de las formas contenidas en un texto o en un corpus son identificables con los x lemas más frecuentes. Aceptando la simplificación que supone igualar la captación del significado con el reconocimiento de la pertenencia a un cierto lema es posible comparar los porcentajes acumulados de comprensión del significado en diferentes textos o corpus. La tabla 3 refleja estas características en el CORPES, la muestra que estamos utilizando y ESLORA.[4]

Tabla 3: Porcentajes acumulados de reconocimiento de formas en CORPES, muestra de CORPES y ESLORA

Los x lemas		Muestra	
más frecuentes	CORPES 0.91	CORPES	ESLORA 2.1
10	34,8	32,61	32,25
25	45,37	44,03	48,14
50	51,46	52,65	62,60
100	56,79	59,26	74,34
500	71,23	75,35	89,60
1000	78,29	82,52	93,51
2000	85,64	89,08	96,48
3000	89,42	92,42	97,82
4000	91,75	94,41	98,58
5000	93,32	95,72	99,05
6000	94,44	96,66	99,36
7000	95,28	97,35	99,60
8000	95,95	97,89	99,74
9000	96,49	98,32	99,87
10 000	96,93	98,64	100
15 000	98,35	99,57	
20 000	99,05	99,95	
25 000	99,42	99,95	
50 000	99,92	100	
[Total lemas]	[112 711]	[18 619]	[9981]

4 Las cifras corresponden a los lemas sin consideración de nombres propios ni cifras y sin diferenciación de la clase de palabras. Para una visión general de los factores implicados en el reconocimiento (y comprensión) de las formas, *cf.* Robles-García (2020).

Fuentes: CORPES (0.91) y ESLORA (2.1). Elaboración propia.

La tabla 3 muestra un panorama bien conocido: los lemas más frecuentes tienen una altísima frecuencia de uso acumulada, de modo que con solo 10 lemas se puede dar cuenta de aproximadamente un tercio de los los elementos contenidos en los textos y con 25 se llega cerca de la mitad. Esos lemas tan frecuentes son fundamentalmente artículos, preposiciones, conjunciones y verbos como *ser* o *haber*. A partir de ahí el incremento en el reconocimiento se va reduciendo y, especialmente importante en este caso, la diferencia entre los textos escritos y los orales se va haciendo mayor. Con los 100 lemas más frecuentes se llega a casi el 75 % de los usos en ESLORA, mientras que los corpus escritos están por debajo del 60 %. Por buscar un referente generalizado en estudios de este tipo, para llegar al 95 % de los textos, en el CORPES hay que trabajar con casi 7000 lemas, en la muestra se necesitan unos 5000 y, en los textos orales es suficiente con algo menos de 2000 lemas. Algo muy parecido se obtiene si se compara el número de lemas necesario para alcanzar el reconocimiento del 98 % del texto: en los textos orales es suficiente con 4000 lemas, mientras que en la muestra del CORPES hacen falta casi 9000 lemas.[5]

Es una diferencia importante, que muestra una configuración del léxico bastante distinta en los dos tipos de textos. Parece evidente que el léxico utilizado en los textos orales es mucho más reducido que el que encontramos en la muestra de textos escritos, lo cual resulta conforme con la visión generalizada. Sin embargo, esa constatación no debería llevarnos a pensar que los dos inventarios pueden compararse de modo tal que establezca una relación entre los 9981 lemas de ESLORA y los 10 000 lemas más frecuentes en el CORPES. Para intentar ver con mayor profundidad lo que sucede, he cruzado el lemario obtenido en ESLORA con el lemario del CORPES, ordenado por frecuencias y diferenciando en él distintos tramos. El objetivo es tener una idea más adecuada de a qué rangos de frecuencia (en un lemario general, como el que deriva del CORPES) corresponden los lemas localizados en ESLORA. Los datos básicos están en la tabla 4.

Tabla 4: Comparación de los tramos del lemario del CORPES con el lemario de ESLORA

Los x lemas más frecuentes en CORPES 0.91	Presentes en ESLORA 2.1			
	Inventario	Uso	% inventario	% uso
1000	956	601 162	9,58	82,39

5 Los porcentajes correspondientes a ESLORA que figuran en la tabla 3 coinciden con los obtenidos por Ávila Muñoz (1999: 92) para la lengua hablada en Málaga. Los 1000 lemas más frecuentes dan cuenta del 93 % de los elementos contenidos en los textos, con los primeros 2000 se llega al 96,6 % y con 5000 se alcanza el 99,7 %. La coincidencia con lo que resulta del análisis de ESLORA es, como se ve, casi total.

2000	1835	628 054	18,38	86,07
3000	2624	645 482	26,29	88,46
4000	3332	652 570	33,38	89,43
5000	3936	657 513	39,43	90,11
6000	4484	661 218	44,93	90,62
7000	4934	663 396	49,43	90,92
8000	5350	665 315	53,60	91,18
9000	5699	667 008	57,10	91,41
10 000	5989	668 135	60,00	91,57
11 000	6251	669 068	62,63	91,70
12 000	6504	669 958	65,16	91,82
13 000	6754	670 687	67,67	91,92
14 000	6964	671 441	69,77	92,02
15 000	7157	672 362	71,71	92,15
16 000	7321	672 912	73,35	92,22
17 000	7471	673 381	74,85	92,29
18 000	7610	673 819	76,24	92,35
19 000	7754	674 255	77,69	92,41
20 000	7895	675 354	79,10	92,56
25 000	8336	680 857	83,52	93,31
50 000	9092	682 583	91,09	93,55
100 000	9188	683 507	92,05	93,67
150 000	9193	683 513	92,10	93,68

Fuentes: CORPES y ESLORA.

La tabla 4 proporciona algunos datos interesantes. El más llamativo es, sin duda, el hecho de que solo 9193 lemas documentados en ESLORA (el 92,10 % de los 9981 que contiene) se encuentren en el lemario resultante de esta versión del CORPES. El análisis detenido de los casos que difieren en las dos listas muestra que la causa está casi siempre en las diferencias en el sistema de anotación y lematización seguido en cada caso, lo cual produce fracasos en la identificación que no proceden realmente de la ausencia de esos elementos en el CORPES. Por ejemplo, la relación de expresiones multipalabra de ESLORA es mucho más amplia que la utilizada en CORPES. Lo mismo sucede con el inventario de elementos fáticos, muy abundantes en la lengua hablada. Hay también errores en la asignación automática del lema en ambos corpus y algunos casos de palabras propias del español de Galicia que no se documentan en el CORPES. Volveremos sobre la cuestión de las ausencias más abajo.

Aun con las precauciones necesarias en función de lo indicado en el párrafo anterior, la tabla 4 muestra los efectos de las discrepancias existentes entre ambos conjuntos léxicos. Con los 1000 lemas más frecuentes del CORPES se reconoce el 82,39 % de las formas presentes en ESLORA, que parece una cifra acorde con lo que ya sabemos, pero los 5000 más frecuentes del CORPES sirven para asignar solo el 90,62 % de las formas, lo cual indica que hay que pensar en otros factores que expliquen las divergencias.

Para obtener una visión más detallada de lo que sucede, he cruzado el lemario del CORPES con el lemario de la muestra de este corpus descrita en el apdo. 1. Dado que se trata de un subconjunto del corpus general, la totalidad de los lemas extraídos de la muestra figuran en el lemario completo, pero lo que interesa aquí sobre todo es la posibilidad de comprobar si los porcentajes acumulados de reconocimiento de lemas presentan diferencias importantes. En la tabla 5 se comparan los porcentajes de uso explicables con los distintos tramos del CORPES en ESLORA y la muestra de CORPES.[6]

Tabla 5: Porcentajes acumulados de identificación de formas en ESLORA y la muestra del CORPES

Los x más frecuentes en CORPES 0.91	% sobre el uso	
	ESLORA	Muestra CORPES
1000	82,39	80,22
2000	86,07	86,87
3000	88,46	90,54
4000	89,43	92,63
5000	90,11	94,05
6000	90,62	95,83
7000	90,92	96,44
8000	91,18	96,44
9000	91,41	96,88
10 000	91,57	97,27
11 000	90,87	97,60
12 000	91,82	97,91
13 000	91,92	98,14
14 000	92,02	98,38
15 000	92,15	98,55
16 000	92,22	98,72
17 000	92,29	98,85
18 000	92,35	98,97
19 000	92,41	99,07
20 000	92,56	99,17
25 000	93,31	99,54
50 000	93,55	99,95
100 000	93,67	100
150 000	93,68	----

Fuentes: ESLORA y CORPES.

Salvo en el primer tramo, los porcentajes de la muestra superan siempre a los encontrados en ESLORA y la divergencia supera los 3 puntos porcentuales a partir del cuarto tramo. Hay que insistir en las diferencias del sistema de lematización, que es sin duda responsable de una parte de

6 No tiene sentido comparar los porcentajes referidos al inventario, puesto que el de la muestra es casi el doble del que se obtiene de ESLORA (*cf. supra*, tabla 2).

los fallos de reconocimiento, pero hay otra parte que se debe a la diferente configuración de los lemarios en los textos escritos y los orales.

Por las razones apuntadas, no se puede dar demasiada importancia a las cifras de lemas de ESLORA ausentes de CORPES, que requieren un análisis cualitativo que veremos posteriormente. Sin abandonar la perspectiva cuantitativa existe, sin embargo, la posibilidad de hacer una comparación ilustrativa entre ambos lemarios. Consiste en analizar la distribución de los lemas de cada tramo de ESLORA entre los diferentes tramos del CORPES. En esta perspectiva se trabaja solo con coincidencias, de modo que los fracasos pesan mucho menos en el panorama general que se obtiene. Los datos correspondientes aparecen en la tabla 6.[7]

Tabla 6: Distribución de los lemas de ESLORA en diferentes tramos de frecuencia del CORPES

	Rango en CORPES														
Rango en ESLORA	1-1000	1001-2000	2001-3000	3001-4000	4001-5000	5001-6000	6001-7000	7001-8000	8001-9000	9001-10000	10001-15000	15001-20000	20001-25000	>=25001	Total
1-1000	552	170	94	43	28	17	5	4	6	3	9	6	4	2	943
1001-2000	199	242	149	91	56	38	27	21	17	13	47	16	13	7	936
2001-3000	103	165	144	110	62	69	35	45	38	19	70	38	23	25	946
3001-4000	46	91	104	104	78	72	71	49	41	34	89	70	30	56	935
4001-5000	18	74	87	103	91	74	55	47	35	27	129	63	46	84	933
5001-6000	10	40	64	86	63	61	60	44	45	29	153	96	44	89	884
6001-7000	12	35	43	65	73	52	48	56	45	29	176	102	58	103	897
7001-8000	7	20	37	27	41	61	49	50	36	44	183	101	76	179	911
8001-9000	5	16	38	42	49	49	62	55	38	46	164	127	79	171	941
9001-10000	4	26	29	37	63	55	38	45	48	46	148	118	70	140	867
Total	956	879	789	708	604	548	450	416	349	290	1168	737	443	856	9193

Fuentes: CORPES y ESLORA.

Como se puede apreciar, solo el 55,2 % de los 1000 lemas más frecuentes de ESLORA están entre los 1000 más frecuentes del CORPES. El 45 % restante se distribuye en otros tramos y se puede destacar que el 5,5 % de los 1000 más frecuentes en ESLORA se encuentra más allá del rango 5001 del CORPES. En la tabla 7 se encuentran estos mismos porcentajes para los primeros 5 tramos de ESLORA. Algo más de un tercio de los que figuran entre los rangos 2001 y 3000 en ESLORA se sitúan más allá de la posición 5001 del CORPES.

7 Como los tramos tienen el mismo tamaño se puede trabajar directamente con las frecuencias.

Tabla 7: Porcentajes de lemas de los primeros tramos de ESLORA situados en rangos superiores a 5001 en el CORPES

Rango en ESLORA	% en rango >=5001 en CORPES
1-1000	5,60
1001-2000	19,90
2001-3000	36,20
3001-4000	51,20
4001-5000	56,00

Fuentes: CORPES y ESLORA

Parece ahora claro que la diferencia más notable entre ambos corpus reside en la distinta distribución que muestran ambos lemarios en cuanto a su situación en rangos de frecuencia. Si invertimos la perspectiva, obtendremos la distribución en ESLORA de los lemas más frecuentes en el CORPES, como muestra la tabla 8:

Tabla 8: Porcentajes de lemas de los primeros tramos del CORPES situados en rangos superiores a 5001 en ESLORA

Rango en CORPES	% en rango =>5001 en ESLORA
1-1000	3,97
1001-2000	15,59
2001-3000	26,64
3001-4000	36,30
4001-5000	47,85

Fuentes: CORPES y ESLORA

Podemos ahora tratar de complementar esta visión puramente cuantitativa de las diferencias entre los dos lemarios identificando los lemas que aparecen en los tres primeros tramos de ESLORA y se encuentran, sin embargo, más allá del rango 15 000 en el CORPES. Es una distancia considerable y, por tanto, puede ilustrarnos sobre sus causas posibles. Dejando a un lado los elementos fáticos o los que pueden resultar de las divergencias entre los sistemas de anotación, entre los 1000 más frecuentes de ESLORA aparecen elementos como *guay, jo, jolín, ajá, flipar, mogollón* o *coña*. Todos ellos son característicos de la lengua coloquial y, por tanto, lo esperable es que en el CORPES aparezcan solo textos orales (que suponen todavía un porcentaje muy bajo), en textos narrativos que contienen fragmentos coloquiales o bien en comunicaciones informales, lo cual explica su posición marginal en las listas de frecuencias habituales.

En el segundo tramo de frecuencias de ESLORA (de los rangos 1001 a 2000) hay 36 lemas que ocupan puestos posteriores al 15 001 en el CORPES. Los más destacados son

> chao, espabilar, botellón, selectividad, currar, notaría, chungo, pela, esquiar, pinzar, reválida, tintorería, veranear, fisioterapeuta, chollo, folk, follón, camping

En la lista anterior figuran, como en la primera, lemas de carácter netamente coloquial, como *chao*, *currar*, *pela*, *chollo* o *follón,* pero aparecen también elementos que no tienen ese carácter informal y cuya presencia en este tramo de frecuencias resulta sorprendente. Me refiero a casos como *notaría* o *tintorería*. Interesa tenerlos en cuenta porque resultan reveladores de otros factores que pueden actuar sobre la distribución de los elementos léxicos. *Tintorería*, por ejemplo, aparece 18 veces en total, todas ellas en la misma entrevista (16 de la informante y 2 de la entrevistadora). Algo parecido sucede en los 21 casos de *notaría*. Esta concentración, derivada de factores relacionados con el tema de la conversación o la entrevista, hace que un cierto lema adquiera gran relevancia y, apoyado por el reducido tamaño del corpus, aparezca, desde una perspectiva general, en una posición mucho más elevada de la que en realidad le corresponde en la configuración global del léxico. Por supuesto, estos problemas se dan en todos los corpus[8] y su efecto perturbador desaparece o, cuando menos, disminuye si, además de la frecuencia total, es posible contar con datos de dispersión y uso.

Observaciones del mismo tipo son aplicables a los 86 del tramo siguiente de ESLORA (2001 a 3000), entre los que menciono únicamente los más frecuentes:

> animalada, capón, chapar, discal, estupendamente, mentalizar, pachanga, positivar, soportal, tirachinas, cojonudo, estresar, opositar, nulo, sedar, vieira, antitabaco, autentificar, catamarán, hablador, mazar, membrillo, mercería, putear, rectorado, calcetar, chipirón, chupito, ciática, cursillo, gaita, goloso, jorobar, mirto, preuniversitario, putada, visera, becar, burrada, cutre, fontanero, paraninfo, pasantía, pazo, pescadería, párkinson, relax

Para completar el panorama de las divergencias entre los dos lemarios, podemos adoptar la perspectiva complementaria, endureciendo las exigencias. Por ejemplo, hay 44 lemas situados en el primer tramo del CORPES (rangos 1 a 1000) que no se documentan en ESLORA. No se trata ya de diferencias debidas al lugar que ocupan en las escalas respectivas, sino de la aparición en una posición marcada en un corpus de unos cientos de millones de elementos y la ausencia total en otro corpus. Los casos más significativos son

> agregar, constituir, nación, víctima, rostro, concluir, candidato, entidad, representante, aumento, ejecutivo, electoral, costo, agente, inmediato

En el segundo tramo (1001 a 2000) son ya 121 lemas los que no aparecen en ESLORA. Entre ellos:

> evento, inicio, poseer, contribuir, ambiental, promover, revelar, indígena, productor, hallar, federal, triunfo, dirigente, judicial, vencer, voluntad, tecnológico, permanente, incremento, ausencia, sugerir, estimar, reconocimiento, comité, efectivo, cultivo, vincular, negociación

Aunque es evidente que el tema requiere una investigación detenida, da la sensación de que en muchos casos se trata de palabras habituales en las noticias de prensa, sector que tiene un gran peso en la configuración del CORPES. *Electoral*, por ejemplo tiene una frecuencia normalizada de 573,48 casos por millón en el área temática correspondiente a política, economía y justicia (textos

8 Se cita con mucha frecuencia el caso de *mucosa* en el BNC (*cf.* Atkins y Rundell 2008, 69).

de prensa, ensayo o divulgación) y de solo 8,84 en el área de ciencia y tecnología o 5,92 en novela. En otros casos se trata, en cambio, de elementos más vinculados al lenguaje literario. *Rostro* tiene una frecuencia normalizada de 280 casos por millón en novela, 331 en relativo y 45,15 en actualidad o 16,63 en textos de política.

3 DISTRIBUCIÓN DE LEMAS EN ESLORA

Como hemos visto en los apartados anteriores, los elementos léxicos presentes en ESLORA son menos numerosos que los que se encuentran habitualmente en corpus escritos de tamaño semejante y presentan una tasa más alta de concentración en los usos, gracias a lo cual con solo los 2000 lemas (sin clase ni nombres propios) más frecuentes se reconoce el 96,5 % de las formas presentes en el corpus, mientras que en la muestra de CORPES que hemos estado manejando hacen falta unos 6000 lemas para alcanzar una tasa semejante.

Esas cifras, indiscutibles, no reflejan, sin embargo, todo lo que se puede apreciar sobre la distribución de las formas. En el apartado 2 me he referido a que una de las consecuencias más claras de la distribución estadística de los elementos léxicos en un corpus es el alto porcentaje de hápax que figuran en él. Si bien los rasgos anteriores podrían hacer pensar en una mayor concentración de lemas en los textos orales y, por tanto, un porcentaje menor de hápax, la comparación de CORPES, la muestra de CORPES y ESLORA indica más bien lo contrario: el porcentaje de lemas (como siempre, sin clases ni nombres propios) con frecuencia igual a 1 sobre el total de los lemas es mayor en ESLORA que el que se da en la muestra y bastante más alto que el que se encuentra en CORPES, como refleja la tabla 9:

Tabla 9: Hápax documentados en el CORPES, la muestra del CORPES y ESLORA

	CORPES 0.91	Muestra CORPES	ESLORA 2.1
Lemas sin clase ni nombres propios	112 711	18 619	9981
Hápax	27 453	5472	3182
Porcentaje de hápax	24,36	29,39	31,88

Fuentes: CORPES y ESLORA.

Las causas de esta cifra más elevada de hápax en ESLORA están probablemente en el volumen reducido del corpus, el tamaño, también reducido, de los textos que lo componen (unos 9000 elementos gramaticales por término medio) y la pluralidad temática intertextual (no intratextual).

Esta diferencia entre la pluralidad que muestran los textos entre sí y la concentración temática que se puede ver en cada uno de ellos apunta hacia otro factor que puede resultar de interés: la distribución de los lemas entre los textos que componen ESLORA. También aquí hay una cierta sorpresa con respecto a lo que harían suponer los datos generales. Como hemos visto, ESLORA documenta 9981 lemas diferentes (sin clase y sin nombres propios). Pues bien, solo 130 de esos lemas (es decir, el 1,31 %) figuran en 70 o más de los 83 textos que componen el corpus. Se trata, claro, de lemas de gran frecuencia, que reúnen en conjunto el 77,40 % de las formas presentes en ESLORA. En el extremo opuesto podemos obtener los lemas que solo aparecen en uno de los documentos: 4180 (el 42,13 %) del total de los lemas, que explican en conjunto únicamente 6125 formas (el 0,81 % de ESLORA).[9]

La diferencia entre 4180 y 6125 muestra que no se trata solo de lemas que se concentran en un documento porque solo aparecen una vez. Tras lo que hemos visto sobre *notaría* o *tintorería* no sorprende descubrir que hay 18 lemas presentes únicamente en un texto, pero tienen una frecuencia general igual o superior a 10 casos. Son los siguientes:

> notaría, contador, tintorería, incapacidad, coral, folk, gua, cirugía, scout, insignia, calambre, gama, hípster, capón, rea(nimación), sedar, autentificar, membrillo.

4 Distribución por clases de palabras

Los cruces de tramos de lemarios como los que hemos visto en el apartado anterior resultan ilustrativos de las características de los textos orales. De forma indesligable a la configuración estadística de los corpus, el análisis de los resultados debe tener en cuenta algunos factores que aconsejan proceder con cierta prudencia en su interpretación. En primer lugar, sin duda, los problemas derivados de la comparación entre conjuntos diferentes de distintos sistemas de anotación y codificación. En segundo lugar, la insuficiencia de contar únicamente con la frecuencia general (o la normalizada), sin manejar también los índices de dispersión o de uso, que pueden proporcionar una perspectiva más adecuada en tanto que reducen drásticamente la importancia de elementos muy frecuentes, pero reducidos a un texto o un tipo de textos. Este rasgo se vincula al de los diferentes tipos de textos incluidos habitualmente en los corpus generales: narración, ensayo, prensa, etc. por una parte, diferentes áreas temáticas por otra y distintos países en el caso del CORPES incrementan el valor de los datos obtenidos, sin duda, pero también hacen que las comparaciones se vean a veces perjudicadas por el peso que uno de estos componentes puede dar a ciertos lemas. En el caso de los corpus orales hay que tener en cuenta también lo reducido de su

9 Para una visión más general de este fenómeno, *cf.*, por ejemplo, Miller y Biber (2015) y Rojo (en prensa).

volumen, que puede conceder mucha importancia a elementos que están presentes en un único texto.

Las dificultades derivadas de todos estos factores disminuyen en la medida en que sea posible incrementar la granularidad del análisis, pero también lo hacen en tanto que las características manejadas se hacen más generales, menos dependientes de lo que sucede con un lema específico. En este sentido, es probable que sea posible obtener una visión interesante de las características de la lengua oral si comparamos la frecuencia de las clases de palabras, tanto en el inventario como en el uso. Por supuesto, también en la consideración del número y frecuencia de sustantivos comunes o adjetivos aparecen problemas derivados de los diferentes sistemas de anotación. Podemos intentar reducir su importancia limitando los recuentos a las clases “llenas”, para evitar la notable cantidad de diferencias que pueden surgir en la consideración de determinantes, demostrativos, numerales, exclamaciones, fechas, etc. Con esto se reducen los problemas, pero no desaparecen por completo. El factor que puede introducir más discrepancias sistemáticas es, probablemente, la consideración de los casos en los que formas como *cansado* son caracterizadas como adjetivos o participio.

Los datos correspondientes al inventario de elementos de estas cuatro clases aparece en la tabla 10.[10]

Tabla 10: Frecuencias de inventario de clases léxicas "llenas" en CORPES y ESLORA

	CORPES 0.91	ESLORA 2.1	% CORPES 0.91	% ESLORA 2.1
Adjetivos	30 737	1724	25,43	17,01
Adverbios	5316	662	4,4	6,53
Sustantivos comunes	75 747	5801	62,67	57,24
Verbos	9069	1948	7,5	19,22
Total de lemas	120 869	10 135	100,00	100,00

Fuentes: CORPES y ESLORA

La configuración observable en los dos corpus es bastante distinta, pero resalta especialmente la diferencia existente en los porcentajes de los adjetivos y los verbos: en el corpus oral el porcentaje de adjetivos sobre el inventario de lemas es bastante menor que en el corpus general y, a cambio, el correspondiente a los verbos es más del doble.

Veamos ahora los datos correspondientes a los usos, que aparecen en la tabla 11:[11]

10 Téngase en cuenta que el total de lemas es superior al que aparece en las tablas 1 y 2 porque en este caso es forzoso hacer los recuentos teniendo en cuenta la clase. Es posible fundir los lemas que presenten dos clases, pero, en ese caso, la asignación automática de la clase mantenida resultaría totalmente arbitraria. En cualquier caso, el sistema de recuento se aplica del mismo modo a los dos corpus.

11 Una de las diferencias existentes entre los sistemas de anotación de CORPES y ESLORA radica en que en CORPES se reconoce la existencia de formas compuestas como tales y, en consecuencia, una forma como *hemos llegado* cuenta una vez y se atribuye al verbo *llegar*. En la versión de ESLORA utilizada aquí, en estas formas se cuentan dos casos y se atribuye uno al verbo *haber* y otro al verbo *llegar*. Para evitar el efecto de la discrepancia, de la cifra correspondiente a los casos de los verbos en ESLORA he restado las 1483 apariciones de *haber* seguidas de

Tabla 11: Frecuencias de uso de clases léxicas "llenas" en CORPES y ESLORA

	CORPES 0.91	ESLORA 2.1	% CORPES 0.91	% ESLORA 2.1
Adjetivos	18 697 604	14 159	13,58	3,96
Adverbios	13 713 450	110 426	9,96	30,91
Sustantivos comunes	63 201 706	87 664	45,91	24,54
Verbos	42 039 795	145 007	30,54	40,59
Total de lemas	137 652 555	357 256	100,00	100,00

Fuentes: CORPES y ESLORA

Los datos de uso muestran la misma tendencia que hemos visto en el inventario, pero de un modo más radical. La pérdida de peso de los sustantivos y, sobre todo, de los adjetivos contrasta con el aumento de los verbo y, sobre todo, de los adverbios.

Aunque es evidente que la anotación automática practicada en esta versión de ESLORA contiene errores, no parece que eso pueda llevarnos a dudar de la fiabilidad de estos porcentajes. De hecho, las mismas tendencias pueden ser observadas en, por ejemplo, las estadísticas sobre clases de palabras proporcionadas por Terrádez Gurea (2001) sobre los 500 lemas más frecuentes de un corpus de español coloquial de 100 000 formas. Los datos aparecen en la tabla 12:[12]

Tabla 12: Porcentajes de inventario y uso de las cuatro clases léxicas

	% inventario	% uso
Adjetivos	8,23	5,88
Adverbios	13,72	24,68
Sust. comunes	41,16	23,56
Verbos	36,89	45,89
Total	100	100

Fuente: Terrádez Gurea (2001: 73 y 84-85).

5 CONCLUSIONES

En las páginas anteriores he tratado de presentar un panorama –todavía muy provisional– de las características básicas que presenta el léxico de los textos orales tal como se manifiesta en los textos incluidos en la versión 2.1 de ESLORA. A las diferencias esperables como consecuencia del carácter coloquial, informal, de estos textos podemos añadir ahora un par de rasgos sobre los que es necesario investigar con más profundidad y, de ser posible, con un volumen mayor de textos orales. El primero de ellos consiste en que la diferencia más visible entre el lemario de un corpus oral y el

un participio.

12 La distribución de los elementos más frecuentes presenta siempre algunas características especiales. Para más detalles, *cf*. Rojo (2011).

lemario de un corpus general parece radicar en divergencias importantes en la distribución según los rangos de frecuencia. El segundo factor diferencial afecta al inventario y la frecuencia de uso de las clases léxicas "llenas". Hay discrepancias fuertes en el peso de los adjetivos y sustantivos comunes, que se reducen mucho con respecto a lo que se observa en los textos escritos, y de adverbios y verbos, que muestran un incremento igualmente notable.

Recursos computacionales citados

CORPES: Real Academia Española. *C*orpus del Español del Siglo XXI, <**http://rae.es/recursos/banco-de-datos/corpes-xxi**>. Versión 0.91.

ESLORA: Corpus para el Estudio del Español Oral. Coord. Victoria Vázquez Rozas, <**http://eslora.usc.es/**>. Versión 2.1

Referencias bibliográficas

Atkins, Sue y Rundell, Michael (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Ávila Muñoz, Antonio Manuel (1999): *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

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

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

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.

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

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

Rojo, Guillermo (en prensa): "Hacia un nuevo concepto de diccionario de frecuencias". Ponencia presentada en el XXX Congreso Internacional de Lingüística y Filología Románicas (Universidad de La Laguna, 4-9 de julio de 2022).

Terrádez Gurea, Marcial (2001): *Frecuencias léxicas del español coloquial: Análisis cuantitativo y cualitativo*. Valencia: Universitat de València.

Vázquez Rozas, Victoria, Mario Barcala, Eva Domínguez Noya, Alba Fernández Sanmartín, Guillermo Rojo y María Paula Santalla (2020): "Codificación y anotación del habla en un contexto bilingüe: el corpus ESLORA de español de Galicia", en Gallego, Ángel J. y Francesc Roca Urgell (eds.): *Dialectología digital del español*, Anexo 80 de *Verba,* Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, pp. 191-226.