Lingüística
Vol. 33-1, junio 2017: 121-134
ISSN 2079-312X en línea
ISSN 1132-0214 impresa
DOI: 10.5935/2079-312X.20170008

SOBRE LA CONFIGURACIÓN ESTADÍSTICA DE LOS CORPUS TEXTUALES

ON THE STATISTICAL STRUCTURE OF TEXTUAL CORPORA

Guillermo Rojo
Universidade de Santiago de Compostela
guillermo.rojo@usc.es

La estructura estadística de los textos y de los corpus textuales es un tema al que se ha prestado muy escasa atención en la lingüística hispánica. El presente trabajo se propone revisar algunos de sus aspectos más importantes en dos direcciones distintas. Por una parte, mediante la aplicación a los datos procedentes de una versión intermedia del CORPES de los análisis realizados previamente sobre el CREA. Por otra, aprovechando los resultados de la anotación morfosintáctica del CORPES, tomando en consideración no solo las formas ortográficas, sino también los lemas.

Palabras clave: lingüística estadística, frecuencia, lingüística de corpus.

Key words: statistical linguistics, frequency, corpus linguistics.

The statistical structure of texts and textual corpora is a topic to which little attention has been paid in Hispanic linguistics. This paper tries to review some of their more relevant aspects in two different senses. On the one hand, through the application of the techniques previously applied to CREA on an intermediate version of CORPES. On the other hand, working on the morphosyntactically tagged version of CORPES, taking into consideration not only orthographic forms, but also lemmas.

(Recibido: 21/04/2017; Aceptado: 18/05/2017)

1. Introducción

El estudio de los aspectos estadísticos de los textos se ha centrado tradicionalmente en el análisis de la frecuencia de los elementos y fenómenos que se encuentran en ellos, especialmente en lo referido al componente léxico. Las razones de este predominio son bastante claras: de una parte, el conocimiento de las frecuencias léxicas resulta de gran interés para la confección de listas o diccionarios de frecuencias que puedan ser utilizados en la enseñanza de lenguas, especialmente como L2; de otra, la recolección de datos en este terreno resulta relativamente sencilla, aunque no por ello menos tediosa. La difusión del empleo de computadoras permite ahora acometer con rapidez y facilidad los aspectos más mecánicos de este tipo de trabajo. Con las diferencias esperables por la naturaleza de los distintos objetos de estudio, algo parecido puede decirse de los análisis estadísticos de fenómenos gramaticales o fonéticos.

Al lado de los enfoques anteriores, centrados en los recuentos que cabe realizar con los elementos y fenómenos contenidos en los textos, existe otra orientación más centrada en el estudio del texto como tal, cuyos antecedentes pueden ser rastreados en la lingüística anterior a la difusión del empleo de computadores.

Apuntando únicamente a los puntos más evidentes en esta línea, el primero es el constituido por la ley de Zipf, formulada a finales de los años cuarenta del siglo pasado. Según esta ley, la relación entre la frecuencia de un elemento y el rango que le corresponde da lugar a una constante (dentro de ciertos límites). Lo esperable es que la frecuencia del segundo elemento en la ordenación por rangos sea aproximadamente la mitad de la que tiene el primero, la del tercero equivalga a un tercio, etc. Su generalización es el hecho, bien conocido, según el cual las distribuciones suponen siempre que en cualquier texto existen unas pocas formas o palabras que tienen una frecuencia muy elevada y muchas formas o palabras con una frecuencia baja o muy baja[1].

El segundo punto de interés en esta dirección puede ser el constituido por el análisis de lo que se llama habitualmente la riqueza léxica de un texto. Consiste básicamente en poner en relación el número total de formas de un texto (*tokens*) con el de formas distintas (*types*). Su formulación básica, realizada por Templin a finales de los años cincuenta, es el índice conocido como *type-token ratio* (TTR), consiste en dividir el número de formas distintas entre el número total de formas, de modo que el resultado oscila entre 0 y 1 y se considera que el texto es más 'rico' desde el punto de vista léxico cuanto más cerca de 1 esté el TTR. El refinamiento de estos índices puede venir de la parte matemática, que no interesa directamente a nuestros propósitos[2] o bien de la parte lingüística. Por ejemplo, reconvertir la distinción habitual entre *types* y *tokens* y aplicarla no a formas, sino a lemas proporciona una visión bastante más adecuada de la variedad léxica que puede encontrarse en un texto.

La llegada de las computadoras a la lingüística permitió aligerar la pesadez de la realización manual de los recuentos y, como consecuencia de su automatización, aplicar el análisis de los índices obtenidos a textos o conjuntos de textos de volúmenes crecientes. Esta ampliación puso de relieve inmediatamente un factor de gran importancia: con independencia de las características individuales de los textos, la relación entre el aumento del volumen del conjunto considerado (los *tokens*) y el de las formas o lemas distintos contenidos en él (los *types*) se hacía cada vez más distante y la curva correspondiente al aumento de las formas o lemas distintos tendía a aplanarse. En 1967, muy cerca de lo que se considera el nacimiento oficial de la lingüística de corpus (LC), John B. Carroll afirmó, según Kučera (1992: 407), que "the number of new lexical items as the size of the text increases gradually slows to a trickle, to reach, for example, just barely over 200 000 in a sample of 100 million tokens". Afortunadamente, la predicción de Carroll resultó errónea y la realidad es que el número de formas distintas no deja de incrementarse con el aumento del tamaño del corpus, aunque, por supuesto, lo hace a un ritmo decreciente[3].

En el presente trabajo me propongo revisar la configuración estadística de los textos escritos en español, revisando lo ya señalado en Rojo (2008) en dos aspectos diferentes. Por una parte, en la sección 2 retomaré los datos sobre la distribución de formas ortográficas procedentes del *Corpus de referencia del español actual* (CREA) y añadiré los que podemos manejar ahora, procedentes de una versión intermedia del *Corpus del español del siglo XXI* (CORPES), que tiene ya un tamaño bastante superior. Por otra, examinaré en la sección 3 los elementos diferenciales que surgen cuando se trabaja no con formas ortográficas, sino con lemas.

[1] Como es lógico, esto no se aplica únicamente a la distribución de formas en un texto, sino a muchas otras esferas de la realidad. Es lo que postula la ley de Pareto, conocida también como regla del 80/20.
[2] Para un análisis detenido de estos índices, vid. Torruella y Capsada (2013) Capsada y Torruella (en prensa).
[3] Cf. Rojo (2008) para algunos datos complementarios relacionados con este punto. En dirección muy diferente a la insinuada por Carroll, vid. las fórmulas para calcular el número de formas distintas en grandes volúmenes de texto expuestas en Sánchez y Cantos (1997) y Cantos y Sánchez (2011).

2. Análisis de frecuencias de formas ortográficas

El recuento automático de las formas ortográficas de un texto o un conjunto de textos es una operación sencilla desde el punto de vista computacional, puesto que, en definitiva, se limita a identificar y aislar las secuencias alfanuméricas que están situadas entre dos espacios en blanco, un signo ortográfico y un espacio en blanco o dos signos ortográficos y luego hacer los recuentos correspondientes. Hay, sin embargo, algunos aspectos en este proceso que requieren la incorporación de un cierto conocimiento lingüístico y la consiguiente toma de decisiones. En primer lugar, aunque es sencillo conseguir rutinas que eliminen los signos ortográficos que pueden figurar inmediatamente antes o después de las secuencias alfanuméricas que constituyen lo que se considera una 'palabra ortográfica', es necesario también tomar decisiones que dependen del sistema ortográfico de la variedad lingüística con que se trabaje acerca de signos que, como los apóstrofos o los guiones, pueden aparecer en medio de una palabra o bien ser considerados como separadores de palabras. Es necesario también tomar decisiones acerca de si se mantiene o no la diferencia entre caracteres en mayúsculas y minúsculas y, por fin, si se toman en cuenta o no las secuencias de dígitos. Como sucede habitualmente, cualquiera de las decisiones posibles presenta ventajas e inconvenientes que deben ser valorados en cada caso en función de la finalidad con que se realicen los recuentos.

Los resultados de estos análisis son bien conocidos. Aunque no es una sorpresa, resulta siempre llamativo el hecho de que unas pocas formas, muy frecuentes, suponen un porcentaje muy importante del volumen total del texto o del corpus. En segundo término, la relación entre el número total de formas y el número de formas distintas cambia de forma muy marcada a medida que el tamaño total del corpus va aumentando. Este hecho llevó a algunos autores a pensar que, a partir de un determinado punto, la línea de formas distintas se haría plana, es decir, que no aparecerían formas nuevas a partir de, por ejemplo, un volumen total de cien millones de formas (cf. supra, apdo. 1). Sin embargo, como muestran con toda claridad (cf. Rojo 2008) los análisis de las formas del *Corpus de referencia del español actual* (CREA), no solo no sucede eso, sino que se puede demostrar que el porcentaje de formas con frecuencia igual a 1 (*hápax*) con relación al total de formas distintas se mantiene relativamente estable con independencia del tamaño del corpus tomado en consideración.

El experimento llevado a cabo con los textos del CREA consistió en realizar cortes con bloques de diferente tamaño, ir acumulándolos y obtener, para cada tramo, la proporción de formas distintas con relación al volumen y también el porcentaje que sobre el total de formas distintas suponen los hápax. El resultado, que reproduzco aquí como tabla 1, es muy claro: la relación entre el número total de formas (*tokens*) y el de formas distintas (*types*) aumenta de forma espectacular a medida que lo hace el tamaño del corpus y pasa de una forma diferente cada 63,3 formas cuando el tramo consta de unos 14 millones de palabras a 206,8 cuando se considera la totalidad del CREA (algo más de 152 millones de formas)[4]. Sin embargo, en contra de lo que las cifras anteriores podrían hacer pensar, el porcentaje de formas con frecuencia igual a 1 sobre el total de formas distintas se mantiene en torno al 40% con independencia del tamaño del tramo tomado en consideración.

[4] Como es lógico, este hecho pesa sobre los índices de riqueza léxica, que en su formulación más básica (cf. supra, apdo. 1) 'castigan' a los textos más largos. Por ejemplo, con los datos de la tabla 1, el TTR de la primera fila (1,6 millones de formas en total) seria 0,043 y el de la última (152,6 millones de formas) 0,005. De ahí que algunas variantes de los TTR introduzcan la longitud del texto como uno de los factores que intervienen en la fórmula (cf. Torruella y Capsada (2013) y Capsada y Torruella (en prensa). Algunas aplicaciones para análisis de corpus, como WordSmith, optan por buscar la solución a este problema por una línea distinta: admiten la reinicialización de los cálculos del TTR cada cierto número de formas (1000, por defecto) y luego proporcionan la media de los TTR parciales obtenidos.

Datos de la parte escrita del CREA (situación en abril de 2008)							
			Formas diferentes			Hápax	
Núm. ficheros	MBytes	Núm. total de formas	Núm. formas diferentes	% formas diferentes	1 forma diferente cada	Total	% sobre formas diferentes
25	9,7	1 602 351	68 468	4,27	23,4	29 440	42,9
50	19,1	3 172 859	96 623	3,04	32,8	39 809	41,2
150	41,5	6 885 997	149 565	2,17	46,0	60 403	40,4
310	83,0	13 838 517	218 743	1,58	63,3	86 824	39,7
750	166,6	27 798 451	320 549	1,15	86,7	127 649	39,8
1500	318,6	53 319 062	440 682	0,82	121,0	179 607	40,7
3212	700,7	117 070 367	644 841	0,55	181,5	271 615	42,1
4188	905,7	147 180 549	717 149	0,49	205,2	303 924	42,4
5426	937,7	152 558 294	737 799	0,48	206,8	314 065	42,6

Tabla 1: Comparación del número de formas distintas, porcentaje que suponen sobre el total de formas, número de hápax y porcentaje sobre el total de formas distintas en diferentes segmentaciones del CREA. Los recuentos no toman en consideración signos de puntuación ni cifras y anulan la diferencia entre mayúsculas y minúsculas.
Fuente: Rojo (2008: tabla 4).

Los datos contenidos en la tabla 1 anulan los temores de que la incorporación de formas nuevas se detenga a partir de un determinado tamaño del corpus y reafirma la conveniencia de que los corpus tengan el mayor tamaño posible, puesto que esa es la única forma de obtener documentación de aquellos elementos con frecuencias más bajas (una aparición cada cien millones de palabras como media, por ejemplo). Al tiempo, la cara menos positiva del fenómeno radica en la evidencia de que, en muchos casos, disponer de documentación (escasa) de una forma determinada es consecuencia de la inclusión en el corpus del texto que la contiene, lo cual es siempre un factor con un alto grado de casualidad.

La claridad de los datos obtenidos no puede hacernos olvidar que los algo más de 152 millones de formas ortográficas contenidos en el CREA no constituyen una cantidad demasiado elevada según los tamaños que poseen en la actualidad los corpus de referencia. Es por ello de gran interés poder contrastarlos con los que proporciona ahora el *Corpus del español del siglo XXI* (CORPES). Al final de su tercera etapa (diciembre de 2018), el CORPES estará formado por 25 millones de formas correspondientes a cada uno de los años comprendidos entre 2001 y 2016, es decir, un total de 400 millones de formas. Para este trabajo, he procesado los datos de una versión intermedia, posterior a la publicada en abril de 2016 (la 0.83), y que consta de unos 240 millones de formas ortográficas (siempre sin cifras y anulando la diferencia entre mayúsculas y minúsculas, para conservar los parámetros con que se obtuvieron las correspondientes al CREA). Dispondremos así de la posibilidad de observar lo que sucede en un corpus que tiene casi cien millones de formas más que el CREA y, sobre todo, podremos comprobar qué es lo que cambia si pasamos de trabajar con formas a hacerlo con lemas que analizaremos en el apartado 3. En esta versión intermedia, ninguno de los años está completo y apenas se han incorporado textos orales.

La tabla 2 muestra los datos segmentados según el año al que pertenecen los textos. Es fácil observar que la relación entre el total de formas correspondiente a cada año y el número de formas distintas es bastante regular, congruente con lo que se observa en el CREA para estos volúmenes y aceptar que la peculiaridad de los años más recientes se debe al escaso tamaño que tienen todavía. Como era de esperar, mayor semejanza se observa en el porcentaje que suponen los hápax sobre el total de formas distintas, situado siempre en torno al 40%[5].

[5] La desviación que aparece en el año 2016 es perfectamente explicable también por su escaso volumen.

	Formas ortográficas			Hápax	
	Total formas (*tokens*)	Total formas distintas (*types*)	1 forma diferente cada	Total	% sobre formas diferentes
2001	16 111 269	243 154	66,26	97 102	39,93
2002	16 828 420	242 687	69,34	95 157	39,21
2003	15 858 120	244 027	64,99	98 186	40,24
2004	17 230 257	244 562	70,45	95 615	39,10
2005	20 262 226	267 670	75,70	105 481	39,41
2006	21 460 475	270 889	79,22	106 720	39,40
2007	21 810 756	278 242	78,39	110 063	39,56
2008	21 186 650	271 185	78,13	106 944	39,44
2009	20 906 916	268 528	77,86	104 939	39,08
2010	20 469 692	267 599	76,49	105 893	39,57
2011	20 729 984	261 863	79,16	101 540	38,78
2012	18 159 073	246 055	73,80	95 753	38,92
2013	2 692 527	97 487	27,62	40 144	41,18
2014	3 257 009	103 780	31,38	41 586	40,07
2015	1 735 716	72 103	24,07	29 694	41,18
2016	870 377	52 951	16,44	23 709	44,78

Tabla 2. Comparación del número de formas distintas, porcentaje que suponen sobre el total de formas, número de hápax y porcentaje sobre el total de formas distintas en diferentes segmentaciones del CORPES. Los recuentos no toman en consideración signos de puntuación ni cifras y anulan la diferencia entre mayúsculas y minúsculas. Fuente: Real Academia Española (http://http://www.rae.es/recursos/banco-de-datos/corpes-xxi). Elaboración propia.

Mayor interés poseen los datos incluidos en la tabla 3. Han sido obtenidos a base de ir acumulando las cifras correspondientes a un año determinado con todos los anteriores, de modo que podemos observar en qué medida el aumento, gradual, del tamaño total tiene sobre las formas distintas y el porcentaje que sobre ellos suponen los hápax. El resultado encaja perfectamente con lo que se obtuvo en el experimento realizado con el CREA. Por una parte, el volumen de formas necesarias por término medio para obtener una forma nueva sigue aumentando hasta alcanzar las 260 al llegar a 240 millones de tamaño total. La relación entre el crecimiento del total de formas y el que muestran las formas distintas se observa con toda claridad en el gráfico 1, en el que, para mayor congruencia de la representación, he dado el valor 100 a los totales correspondientes al año 2001 y he recalculado con relación a él los que van resultando de la acumulación de los años posteriores. Lo importante es que la curva no deja de subir, aunque es evidente que la pendiente se dulcifica de un modo considerable[6]. Por otra, el porcentaje de hápax sobre el total de formas diferentes sigue situado ligeramente por encima del 40%, con independencia del tamaño del corpus[7].

[6] Nótese, por otra parte, la más que notable coincidencia con los datos del CREA cuando, al añadir las formas de 2008, se alcanza el tamaño de este corpus: una forma nueva cada 206,8 / 206,14 y un porcentaje de hápax situado en 42,6% / 42,74%. Para valorar la semejanza de los resultados debe tenerse en cuenta que la agrupación de ficheros se ha hecho de forma distinta en cada caso: en el CREA la acumulación se realizó según el orden con que la computadora iba accediendo a los ficheros del corpus; en el caso del CORPES, en cambio, la agrupación se hace por el año al que corresponde cada texto.

[7] El aplanamiento del tramo final de la curva se debe sobre todo al escaso peso que tienen los años más recientes en la configuración actual del CORPES. Como puede verse, el mismo aplanamiento se da en la línea correspondiente al total de las formas.

En consecuencia, alcanzamos de nuevo las dos conclusiones obtenidas en el análisis del CREA: es importante construir corpus de gran tamaño porque es el único modo de poder documentar formas, fenómenos o usos de baja frecuencia y, al mismo tiempo, es conveniente no olvidar el grado de aleatoriedad derivado de la elección de un texto u otro para formar parte del corpus.

	Frecuencia total	Total formas distintas	1 forma diferente cada	Hápax	% sobre formas distintas
2001	16 111 269	243 154	66,26	97 102	39,93
+2002	32 939 689	341 001	96,60	137 688	40,38
+2003	48 797 809	418 988	116,47	172 737	41,23
+2004	66 028 066	483 571	136,54	200 758	41,52
+2005	86 290 293	553 808	155,81	232 461	41,98
+2006	107 750 768	616 653	174,73	260 715	42,28
+2007	129 561 527	677 377	191,27	282 451	41,70
+2008	150 680 157	730 953	206,14	312 398	42,74
+2009	172 019 610	780 758	220,32	335 691	43,00
+2010	192 489 302	828 271	232,40	358 402	43,27
+2011	213 219 286	870 729	244,87	378 242	43,44
+2012	231 398 339	907 312	255,04	395 454	43,59
+2013	234 090 866	912 315	256,59	397 713	43,59
+2014	237 347 875	917 988	258,55	400 065	43,58
+2015	239 083 591	920 479	259,74	401 139	43,58
+2016	239 953 968	922 433	260,13	402 070	43,59

Tabla 3. Tamaño total, número de formas distintas y hápax correspondientes a la acumulación de textos correspondientes a diferentes años del CORPES. Los recuentos no toman en consideración signos de puntuación ni cifras y anulan la diferencia entre mayúsculas y minúsculas. Fuente: Real Academia Española (http://www.rae.es/recursos/banco-de-datos/corpes-xxi). Elaboración propia.

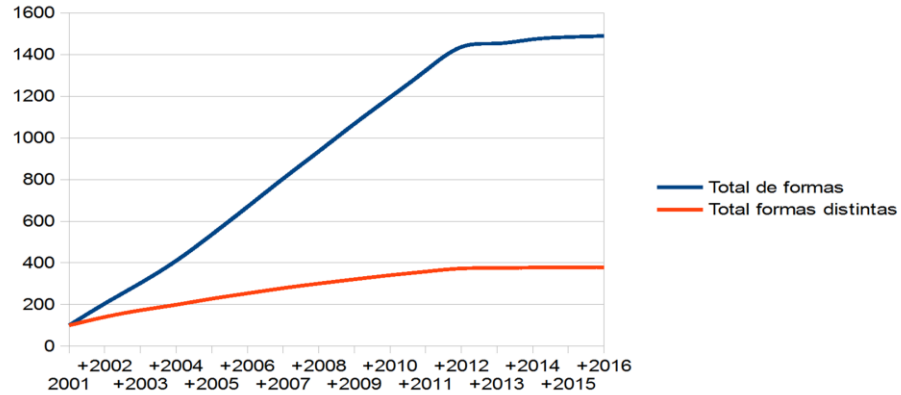

Gráfico 1: Evolución del total general de formas y del total de formas distintas en el CORPES 0.83.

3. Análisis de frecuencias de lemas

Sin ánimo de infravalorar la información que se puede obtener de recuentos como los descritos o mencionados en el apartado anterior, es evidente que solo mediante la adición de información lingüística, de carácter cuando menos morfosintáctico, es posible llegar a análisis mejor fundamentados desde el punto de vista lingüístico y, por tanto, mucho más interesantes para conocer la composición interna de los textos.

Añadir información de carácter morfosintáctico es la única vía para obtener estadísticas aplicadas a, por ejemplo, la distribución cuantitativa de las subcategorías vinculadas a una categoría (las formas temporales de los verbos, los géneros de los sustantivos y adjetivos, etc.) y, por supuesto, todo lo que implique unidades superiores a la palabra, siempre, claro está, que se disponga de los recursos adecuados para la clasificación (estructuras sintácticas clausales, por ejemplo, cf. Rojo 2003). Además, la adición de información morfosintáctica permite arrojar una luz diferente sobre las estadísticas léxicas, como trataré de mostrar al oponerlas a las examinadas en el apartado anterior.

El primer aspecto en el que difieren los dos tipos de recuentos es evidente: los que trabajan con las formas gráficas las usan como elementos básicos (en realidad, los únicos) del recuento y, por tanto, presentan desajustes en todos aquellos aspectos en los que se rompe la relación entre las formas gráficas y los elementos gramaticales, como sucede en las contracciones, las grafías con elementos enclíticos y las unidades multipalabra[8]. Las contracciones son solo dos en español actual, pero no atribuir sus apariciones a las preposiciones que las formas y el artículo con el que se combinan supone distorsionar la frecuencia de estos elementos[9]. Las formas gráficas como *llévalo*, *llevándome*, *llevárselo* implican dos o tres elementos gramaticales diferentes, lo cual hace que los recuentos correspondientes a imperativos, gerundios, infinitivos y pronombres átonos resulten muy distorsionadas si se trabaja únicamente con formas gráficas en textos no anotados. Por fin, las convenciones ortográficas (con variantes como *enseguida* frente a *en seguida*) impiden reconocer la existencia de elementos unitarios como *sin embargo*, *sin la menor duda* y expresiones similares si no hay un tratamiento posterior al simple aislamiento de las formas gráficas.

El segundo aspecto en que difieren estas dos vías es el relacionado con la lematización. En sus comienzos, la obtención de recuentos automáticos sobre textos electrónicos produjo una notable decepción. Las estadísticas tradicionales hacían la lematización ya en la fase de recogida de datos, agrupando directamente, en una entidad única, por ejemplo todas las formas del paradigma de un verbo (con la posibilidad, cómoda, aunque poco rentable, de olvidar las formas concretas, con lo que se simplificaba la estadística de los lemas, pero se bloqueaba la referida a las subcategorías). En cambio, los recuentos realizados directamente sobre la versión ortográfica de textos no pueden tener en cuenta la información gramatical, de modo que *llega*, *llegaré*, *llegaremos*, etc. aparecen como elementos diferentes, no vinculables más que mediante su similaridad gráfica, y, como es lógico, se computan por separado.

Obtener la frecuencia de un verbo con recuentos de este tipo supone una carga de trabajo importante y, además, arriesgado, puesto que no hay forma de decidir qué hay que contabilizar en los casos de homografía (*casa*, *vino*, *canto*, etc.)[10]. Por otra parte, los sistemas ortográficos imponen siempre ciertas condiciones sobre la presentación de los elementos gramaticales, con lo que las estadísticas resultantes son diferentes.

[8] Por supuesto, soy consciente de que es posible lograr recursos importantes usando únicamente las formas ortográficas, como ocurre con los ngramas de Google o incluso de coapariciones (*collocations*), pero eso ya son análisis de otro tipo, mejorables sin duda en el momento en que se hagan no por formas, sino por lemas.

[9] En otras épocas del español (y, claro, en otras lenguas), las contracciones son muchas más. Por otro lado, tendríamos que considerar también lo que sucede en textos con peculiaridades gráficas que pretenden reflejar ciertas variedades distintas de las estándares (del tipo *pa'l*, etc.).

[10] Es el mayor problema que presentan recursos como la *Lista de frecuencias de palabras del castellano de Chile* (LIFCACH), elaborada por Sadowsky y Martínez Gamboa. Su versión 2.0. (2012) consiste en una lista de frecuencias derivada de un conjunto de 102 listas parciales que suman en total unos 800 millones de formas. La anotación se ha hecho directamente sobre las listas, con lo que la desambiguación necesaria en los casos de homografía no puede usar la información existente en el contexto sintáctico inmediato.

Hay otros aspectos en los que las ventajas de los recuentos realizados sobre textos anotados son también muy claras. En el apartado anterior he aludido a la conveniencia de trabajar suprimiendo la diferencia entre mayúsculas y minúsculas por un lado y no tener en cuenta las cifras por otro. Sin duda, esas opciones presentan ventajas sobre las alternativas contrarias, pero implican también ciertos costes. Eliminar la diferencia entre mayúsculas y minúsculas, por ejemplo, permite obtener la frecuencia de una forma con independencia de si aparece con mayúscula inicial por las convenciones gráficas, pero distorsiona los resultados de los recuentos de, por ejemplo, los días de la semana (*domingo* frente a *Domingo*) o los meses del año (*julio* frente a *Julio*).

Los nombres de personas, entidades comerciales, instituciones, países, ciudades, regiones, productos, etc. son otro factor que es necesario tomar en consideración. Parece claro que, con este tipo de elementos, la fusión de mayúsculas y minúsculas es más bien un factor contraproducente, puesto que anula diferencias que pueden ser cruciales en su procesamiento. La única vía razonable es, por supuesto, tratar de reconocerlos como elementos especiales (*named entities*)[11] e identificarlos en toda su extensión.

Esto es, localizar las apariciones de secuencias del tipo *Ministerio de Educación, ciencia y deporte*, *Juan Domínguez Vázquez*, *Miranda de Ebro*, etc. y sus paralelos en nombres de empresas, entidades y productos comerciales, etc.

El tratamiento de los numerales es otro aspecto en el que las diferencias entre los dos enfoques son importantes. No tener en cuenta las cifras es lo lógico para evitar la distorsión que supondría computar la enorme cantidad de secuencias de dígitos que aparecen en los textos, pero resulta insuficiente, porque hay casos en los que la indicación de cantidades se hace a través de un sistema mixto. Cadenas como *3250*, *tres mil doscientos cincuenta*, *3 mil doscientos cincuenta* y otras variantes posibles para la expresión de la misma cifra son tratadas de modo distinto según el enfoque que se adopte en cada caso, pero lo ideal es considerar que se trata en todos los casos de un numeral y, si se estima necesario, mantener la indicación de que constituyen expresiones diferentes de la misma cantidad.

La aplicación de los programas de *tokenización*, anotación y desambiguación a la versión 0.83 del CORPES, publicada en abril de 2016, presenta un total de 245 949 127 elementos lingüísticos. Los tipos generales a los que estos elementos pertenecen se muestran en la tabla 4. Como se puede observar, se diferencia entre aquellos a los que nos referimos habitualmente cuando pensamos en el componente léxico (palabras y locuciones) y otros elementos que el análisis identifica, que son necesarios para entender lo que contienen los textos, pero que tienen un carácter que los distancia del léxico y también los diferencia entre sí.

Los más alejados son, por supuesto, los signos ortográficos, pero cerca se encuentran también las abreviaturas, los nombres de personas, entidades, productos, nombres científicos, expresiones de fecha y hora, así como las cifras y expresiones próximas a ellas. En total, 2 167 027 elementos lingüísticos distintos adscribibles a 1 529 179 lemas o elementos abstractos equivalentes (cuando no se trata de elementos léxicos en sentido estricto).

En el caso de los lemas netamente léxicos y gramaticales se ha tenido en cuenta la clase de palabras atribuida en cada caso, de modo que, por ejemplo el lema *a* figura dos veces, una como preposición y otra como sustantivo.

[11] Se dividen habitualmente en ENAMEX (*entity name expression*) y NUMEX (*numerical expression*).

Clase	Número de elementos distintos	Número de lemas o entidades equivalentes	Total de elementos
Signos de puntuación	165	165	29 995 190
Abreviaturas y acrónimos	4704	3336	480 578
Cifras y expresiones mixtas	140 341	118 172	2 326 417
Fechas, horas, etc.	42 903	39 481	216 945
Entidades nombradas, nombres científicos, etc.	748 657	663 681	5 564 368
Referencias electrónicas	18 040	18 040	24 891
Elementos no identificados	522 826	508 085	3 837 530
Palabras	679 092	173 745	200 063 978
Locuciones	10 299	4474	3 439 230
	2 167 027	1 529 179	245 949 127

Tabla 4: Distribución de tipos de elementos en la versión 0.83 del CORPES. Fuente: Real Academia Española (http://www.rae.es/recursos/banco-de-datos/corpes-xxi). Elaboración propia.

El primer dato que salta a la vista es, por supuesto, el del número de lemas (en sentido estricto) identificados en esta versión del CORPES: 173 45, casi el doble de los que figuran en, por ejemplo, la 23.ª edición del DLE, que contiene 93 11 entradas (DLE 23.ª: xi). Sin duda, en el CORPES hay muchas palabras que no aparecen en el DLE (ni en otros diccionarios) por razones relacionadas con su carácter reciente, escasa frecuencia, pertenencia a ámbitos geográficos o técnicos específicos, etc., y que quizá no vayan a aparecer nunca por su carácter pasajero, pero hay factores más generales que explican esa discrepancia. La consulta a los diccionarios (tanto impresos como electrónicos) exige en muchas ocasiones la aportación de conocimientos gramaticales por parte de quien la hace. Si, por ejemplo, se sabe (y se indica de alguna forma en el diccionario) que los adverbios en *-mente* tienen un significado deducible a partir del que tiene la base adjetiva sobre la que están formados, es posible no incluir una cierta cantidad de adverbios de este tipo, puesto que se trabaja con la garantía —razonable— de que quien se encuentre un elemento de este tipo en un texto llegará a su significado después de consultar el del adjetivo correspondiente en caso necesario.

En cambio, a la hora de etiquetar el contenido de un texto, es necesario atribuir un lema a todo elemento presente en él. En el DLE (23.ª) figuran 161 entradas correspondientes a elementos que comienzan por *a-* y terminan en *-mente*. Con estas mismas características, en el lemario de la versión 0.83 del CORPES aparecen 501 elementos, 499 de los cuales están etiquetados como adverbios. Algo parecido sucede con elementos que implican procesos de prefijación o sufijación. En un diccionario, puede considerarse suficiente con incluir una entrada para explicar los valores de *anti-* y prescindir de todos aquellos derivados que sean explicables como la adición del significado del prefijo al de la base sobre la que se fija. En el DLE (23.ª) hay 23 casos de palabras que comienzan por *antia-*, mientras que en el CORPES 0.83. encontramos 127 con estas mismas características.

Otra fuente de divergencias, más general e importante, procede del modo de reflejar, en la organización de las entradas, las diferencias en las clases gramaticales a las que puede corresponder una palabra.

En la tradición hispánica, por ejemplo, lo habitual es que los significados de una palabra como *militar* aparezcan distribuidos en dos entradas: en una aparecen los significados del verbo y en la otra los que corresponden a *militar* cuando es adjetivo o sustantivo[12]. Este segundo caso es muy común, de modo que el contraste con lo que resulta en el proceso de anotación de un corpus en el que es obligatorio atribuir una de estas clases explica una buena cantidad del exceso de entradas que cualquier corpus de tamaño medio puede presentar con respecto al lemario de un diccionario. Por supuesto, siempre es posible hacer recuentos de lemas contenidos en un diccionario teniendo en cuenta la clase gramatical y diferenciando, por tanto, según ese carácter, pero lo habitual es hacer los recuentos a partir del número de entradas (presentadas como lemas) y de acepciones.

Veamos ahora, teniendo en cuenta todo lo expuesto anteriormente, la distribución y peso en los textos de diferentes clases de palabras. La tabla número 5 da las frecuencias totales y los porcentajes correspondientes a las clases consideradas en el proceso de anotación aplicado en la versión 0.83 del CORPES y que reciben la consideración general de "palabra", lo cual significa que quedan excluidos de estos recuentos todos los demás tipos que figuran en la tabla 5.

Clase de palabras	Frecuencia	Porcentaje
Artículos	26 175 594	13,08
Preposiciones	27 076 471	13,53
Conjunciones	12 690 034	6,34
Contracciones	3 361 474	1,68
Interjecciones	92 365	0,05
Relativos	4 271 974	2,14
Interrogativos	416 907	0,21
Cuantificadores	3 032 564	1,52
Numerales	1 319 163	0,66
Demostrativos	2 105 967	1,05
Posesivos	3 040 298	1,52
Pronombres personales	7 108 847	3,55
Adverbios	9 764 443	4,88
Adjetivos	13 997 787	7,00
Sustantivos	49 418 985	24,70
Verbos	36 191 105	18,09
Totales	200 063 978	100,00

Tabla 5: Frecuencia y porcentajes de diferentes clases de palabras en el CORPES 0.83. Fuente: Real Academia Española (http://www.rae.es/recursos/banco-de-datos/corpes-xxi). Elaboración propia.

Como sucede en todos los recuentos de este tipo, con independencia del tamaño de los corpus, las palabras gramaticales suponen un porcentaje muy alto de los textos, especialmente elevado en el caso de artículos y preposiciones, que superan en ambos casos el 13% del total. En realidad, el porcentaje es más alto, puesto que habría que anular las contracciones del listado de elementos y atribuir las cifras correspondientes a artículos por un lado (*el*) y preposiciones por otro (*a* y *de*)[13].

[12] Esto es lo que sucede en el DLE y en la mayor parte de los diccionarios. La excepción más importante es la primera edición del *Diccionario de uso del español* (DUE) de María Moliner que tiene una única entrada en la que, de acuerdo con el sistema de acepciones y subacepciones se reflejan todos los significados de la palabra en cada una de las clases en las que puede aparecer. Pero esta característica ha desaparecido en las ediciones posteriores (1998, 2007 y 2016), que las distribuyen en dos entradas organizadas del modo más habitual en la lexicografía española.

[13] Esa operación cambiaría también los totales, puesto que *al* y *del*, que han sido considerados en estos recuentos como un elemento en cada una de sus apariciones, pasarían a ser dos elementos. Con ello, el porcentaje de las preposiciones se convertiría en el 14,92% de un total de 204 001 452.

En total, todos estos elementos suponen conjuntamente el 34,68%, de modo que una de cada tres palabras de cualquier texto de español contemporáneo pertenece a alguno de estos tipos netamente gramaticales.

El segundo bloque, constituido por los pronombres y adjetivos (no calificativos) de la gramática tradicional suma en conjunto el 10,65% del total[14]. Por fin, el formado por las palabras de las que se dice habitualmente que están dotadas de contenido léxico suman conjuntamente el 55,48% y, lógicamente, el volumen más importante corresponde a sustantivos (casi el 25%) y verbos (el 18,09%). Es especialmente importante tener en cuenta que, salvo error, todos los nombres propios han sido excluidos de esta tabla, puesto que están integrados en la clase de las entidades nombradas. Se trata, pues, únicamente de los sustantivos comunes, lo cual puede explicar posibles discrepancias con recuentos efectuados sobre una clasificación diferente de los elementos.

Las conocidas diferencias entre la frecuencia de inventario y la frecuencia en el texto pueden verse aquí con toda claridad[15]. Por citar únicamente algunos casos, las preposiciones suponen únicamente el 0,019% del inventario léxico del español, pero les corresponde el 13,5% (o el 14,9% si anulamos las contracciones como entrada independiente) de las frecuencias en los textos. Totalmente distinto es el caso de los sustantivos, que constituyen el 62,74% del inventario de elementos distintos (*types*) y solo el 24,7% de la frecuencia en los textos. En una situación media, los verbos son el 7,63% del inventario y el 18,09% de los textos.

En cuanto a la concentración de las frecuencias, los datos del CORPES confirman, como era de esperar, lo ya conocido acerca del alto porcentaje sobre el total que supone un número muy reducido de elementos de alta frecuencia. En el caso concreto de la versión 0.83 del CORPES, los 21 lemas[16] que tienen frecuencia igual o superior al 0,5% del total, suponen conjuntamente nada menos que el 43,33% de las frecuencias de todos los lemas pertenecientes a este grupo. La cifra obtenida está próxima a los datos que pueden deducirse de los listados contenidos en Davies (2006): los 18 lemas con frecuencias superiores al 0,5% del corpus utilizado suponen conjuntamente el 41,1% del total. Los 22 lemas con esta misma característica en el corpus de algo más de dos millones de formas etiquetadas en Almela et al. (2005) suman en total el 39,92%. Son porcentajes congruentes entre sí y que se sitúan en la línea de lo esperado. El que sean claramente más altos que los que podemos obtener con el análisis de las frecuencias de formas se explica con rapidez. Trabajar con lemas supone una reorganización de las formas que altera la configuración de la zona más alta: las preposiciones y conjunciones no tienen cambios, pero sí se concentran las formas de los artículos, los pronombres, los verbos, etc. Así, frente a lo que se puede observar en los listados de formas, en el grupo de los que tienen frecuencia igual o superior al 0,5% del CORPES aparecen cuatro verbos: *ser* (1,53%), *haber* (0,96%), *ir* (0,53%) y *estar* (0,52%)[17]. En cuanto al aumento —no especialmente importante— con respecto a las frecuencias que se obtienen de los recuentos de Almela et al. (2005) y Davies (2006), hay que tener en cuenta que las estadísticas que usamos proceden de un conjunto que no contiene locuciones, entidades nombradas, etc., elementos que suponen un peso importante y que, si bien no influyen de modo apreciable en la frecuencia de los aquí considerados, sí pueden alterar su peso porcentual.

[14] Téngase en cuenta que, como hemos visto ya, los numerales considerados aquí son únicamente aquellos que aparecen escritos o transcritos íntegramente con caracteres alfabéticos (salvo los números romanos, que tienen su tipo propio). Los demás son incluidos, según los casos, en las clases NUMEX, etc. de la tabla 4.

[15] Utilizo estos términos como adaptaciones, ligeramente modificadas en su concepción, de los propuestos por Bybee (2007) *type frequency* y *token frequency*. Vid. Rojo (2011) y la bibliografía allí analizada para el desarrollo y justificación de este tratamiento.

[16] 19 lemas si no individualizamos las contracciones.

[17] Es necesario tener en cuenta que, al nivel en que nos movemos aquí, estos recuentos no diferencian entre los usos de los verbos como auxiliares o principales, frente a lo que se hace en la *Base de datos sintácticos del español actual* (BDS), que solo cuenta los usos como verbo principal. Los verbos que ocupan las primeras cuatro posiciones en la BDS son *ser*, *decir*, *estar* y *tener*, cf. Rojo (2001: 265).

Vayamos ahora al último punto de interés en esta aproximación a la configuración estadística de los corpus: el porcentaje de hápax. Como se ha señalado en el apartado 1, la cuestión es de importancia teórica y también práctica, puesto que la demostración de que el número de formas o lemas distintos desciende tanto a partir de un determinado volumen que la curva de crecimiento se hace plana podría llevarnos a la conclusión de que no tiene sentido tratar de reunir corpus altamente codificados con tamaños superiores a, por ejemplo, cien millones de formas. Incluso cabría pensar que, si bien se observa que el porcentaje de hápax sobre las formas diferentes (*types*) se mantiene estable con independencia del tamaño del corpus (cf. supra, apdo. 2), eso podría deberse a factores como, por ejemplo, la aparición constante de nuevos nombres propios o incluso al hecho de que pueden ir incorporándose formas de los tiempos menos frecuentes de los verbos (el futuro de subjuntivo, por ejemplo), pero que no se produce realmente la entrada de lemas nuevos.

No parece ser eso lo que sucede en realidad. Los datos obtenidos de la versión 0.83 del CORPES muestran que el porcentaje de lemas con frecuencia igual a 1 resulta considerablemente alto, aunque, como era de esperar, no tanto como el que se obtiene con las formas ortográficas. Trabajando, como en los apartados anteriores, sin tomar en cuenta nombres propios, locuciones ni cifras se llega a la conclusión de que el porcentaje de hápax entre los lemas restantes es del 33,45%. Es decir, nada menos que uno de cada tres lemas localizados en un corpus de unos doscientos millones de elementos de este tipo aparece solo una vez. No constituye, por supuesto, un resultado inesperado, pero sí una importante confirmación de lo que suponíamos acerca de la configuración de los textos también desde un enfoque considerablemente más abstracto que el habitual. Tampoco es de extrañar que el 99,67% de estos casos procedan de adverbios, adjetivos, verbos y sustantivos, es decir, elementos pertenecientes a clases abiertas. Por último, resulta también de interés comprobar que el porcentaje de hápax muestra oscilaciones según la clase de palabras. Como muestra la tabla 6, sustantivos y adjetivos son las clases en las que más abundan estos elementos, los verbos están 12 puntos por debajo y los adverbios ocupan una posición intermedia[18].

Clase de palabras	Porcentaje de *hápax*
Sustantivos	34,65
Adjetivos	34,53
Adverbios	28,84
Verbos	22,63

Tabla 6: Porcentaje de *hápax* en diferentes clases de palabras. Fuente: Real Academia Española (http://www.rae.es/recursos/banco-de-datos/corpes-xxi). Elaboración propia.

4. Conclusión

El análisis de los datos procedentes de la versión analizada del CORPES confirma y amplía los mostrados anteriormente (cf. Rojo 2008) a partir de los contenidos en el CREA. En primer lugar, la proporción de formas distintas con respecto al total de formas disminuye con el aumento del tamaño del corpus, pero queda demostrado que la curva de crecimiento de las formas distintas sigue ascendiendo, aunque lo haga de forma más suave. Este hecho se debe en buena parte a que la tasa de hápax con respecto al total de formas distintas se sitúa siempre en torno al 40% con independencia del tamaño del corpus. Ambos rasgos quedan demostrados al menos hasta corpus con un tamaño total de unos 240 millones de formas.

[18] Como es bien sabido, el CORPES constituye un recurso en construcción, lo cual afecta no solo a su composición, sino también a los procesos de segmentación y anotación. Es previsible, por tanto, que los resultados que se puedan obtener de versiones posteriores alteren ligeramente estos datos, aunque las oscilaciones no deberían ser de importancia.

En segundo término, estas características se confirman también cuando los elementos con los que se trabaja son lemas y no formas ortográficas. En un corpus de este tamaño, los lemas que tienen frecuencia igual a uno constituyen el 33% del total de lemas considerado en sentido estricto, lo cual permite pensar que su evolución será semejante a la que se puede observar con las formas, aunque, como es lógico, con un porcentaje inferior.

Relación de corpus y otros recursos electrónicos mencionados en el texto

BDS: *Base de datos sintácticos del español actual* (http://www.bds.usc.es).
CORPES: *Corpus del español del siglo XXI* (http://rae.es/recursos/banco-de-datos/corpes-xxi).
CREA: *Corpus de referencia del español actual* (http://rae.es/recursos/banco-de-datos/crea).
LIFCACH: Sadowsky, Scott, & Ricardo Martínez-Gamboa. 2012. LIFCACH 2.0: *Word Frequency List of Chilean Spanish (Lista de Frecuencias de Palabras del Castellano de Chile)*, version 2.0. Zenodo. http://doi.org/10.5281/zenodo.268043.

Referencias bibliográficas

Almela Pérez, Ramón, Pascual Cantos, Aquilino Sánchez, Ramón Sarmiento y Moisés Almela. 2005. Frecuencias del español. Diccionario y estudios léxicos y morfológicos, *Madrid: Universitas.*

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

Cantos, Pascual y Aquilino Sánchez. 2011. El inglés y el español desde una perspectiva cuantitativa y distributiva: equivalencias y contrastes, *Estudios ingleses de la Universidad Complutense*, 19: 15-44.

Capsada, Ramón y Joan Torruella. (En prensa). Métodos para medir la riqueza léxica de un texto. Revisión y propuesta. Aplicación en el Corpus Informatizado del Catalán Antiguo, *Verba*, 44.

Davies, Mark. 2006. *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*, New York, Routledge.

DLE. Real Academia Española y Asociación de Academias de la Lengua española. 2014. *Diccionario de la lengua española*, Madrid, Espasa. [en línea] Disponible en http://dle.rae.es

DUE. Moliner, María (1966-1967). 1998[2], 2007[3] y 2016[4]. *Diccionario de uso del español*, Madrid, Gredos.

Kučera, Henry. 1992. The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids, en Svartvik (Ed.) 1992: 401-420.

Rojo, Guillermo. 2001. La explotación de la *Base de datos sintácticos del español actual* (BDS), en Josse De Kock (ed.), *Lingüística con corpus. Catorce aplicaciones sobre el español* (= *Gramática española. Enseñanza e investigación* I.7), Salamanca, Univ. de Salamanca: 255-286.

Rojo, Guillermo. 2003. La frecuencia de los esquemas sintácticos clausales en español, en Francisco Moreno Fernández, Francisco Gimeno Menéndez, José Antonio Samper, M.ª Luz Gutiérrez Araus, María Vaquero y César Hernández (coords.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, Arco/Libros, Madrid, vol. I: 413-424.

Rojo, Guillermo. 2008. Lingüística de corpus y lingüística del español, ponencia plenaria en el XV congreso de la Asociación de Lingüística y Filología de América Latina (Montevideo, 18-21 de agosto de 2008). Montevideo. CD [ISBN 978-9974-8002-6-7]

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

Sánchez, Aquilino y Pascual Cantos. 1997. Predictability of Word Forms (types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the *Cumbre* Corpus: An 8-Millon-Word Corpus of Contemporary Spanish, *IJCL*, 2, 2: 259-280.

Svartvik, Jan (Ed.). 1992. Directions in Corpus Linguistics, *Proceedings of Nobel Symposium 82*, Berlin, Mouton de Gruyter.

Torruella, Joan y Ramón Capsada. 2013. Lexical Statistics and Typological Structures: a Measure of Lexical Richness, *Procedia. Social and Behavioral Sciences*, 95: 447-454.