Corpus léxico y diccionario: la estricta representatividad estadística

Hugo E. LOMBARDINI
Universidad de Bolonia

Silvia BIANCONCINI
Universidad de Bolonia

Resumen

Los corpus de grandes dimensiones proporcionan información léxica importante y completa, pero su análisis completo resulta prácticamente inabarcable, especialmente si se lo interroga con fines lexicográficos. Ahora bien, de tales corpus pueden extraerse subcorpus de dimensiones significativamente más pequeñas y, de este modo, superar las dificultades impuestas por las dimensiones. Pero la cuestión más significativa de este procedimiento de reducción es que esta debe no solo preservarse el número de acepciones del corpus total y sus respectivas frecuencias, sino, además, lograrlo con el número mínimo de ejemplos necesarios.

Aquí se demuestra que –ayudándonos con una metodología estadística– esto es posible, es decir, que se puede establecer el número mínimo de muestras que un subcorpus debe tener para reflejar fielmente cualitativa y cuantitativamente el corpus del cual procede. Para corroborar nuestros hallazgos, aplicamos nuestra fórmula a dos subcorpus del término *externo* extraídos del CREA y analizados en su totalidad (los de España y Argentina) y uno (el de México) del cual conocemos solo una reducción suya.

Palabras clave: corpus, lexicografía, estadística, diccionario, acepción lexicográfica.

Abstract

Corpuses of large dimensions provide important and complete lexical information, but their analysis can become cumbersome, particularly for lexicographic purposes. Sub-corpuses of significantly smaller dimensions could be extracted from the original corpus and analyzed to overcome such limitations. However, an important aspect is to define which is the optimal dimension for these selected sub-corpuses in order to preserve the main features of the original corpus, both qualitatively and quantitatively. We show how statistical methodologies can help in determining the optimal sample size. To corroborate our findings, we consider the corpus CREA (reference corpus of the current Spanish) and, as object of study, the adjective *externo* and its meanings. We show how the different meanings of this word are preserved and well-represented in a much smaller sub-corpus. This is shown for three different countries: Argentina, Spain and Mexico.

Keywords: corpus, lexicography, statistics, dictionary, lexicographical meaning.

0. INTRODUCCIÓN

¿Cómo reducir a un corpus "manejable" otro sensiblemente mayor con la finalidad de utilizar sus ejemplos léxicos para redactar una entrada lexicográfica[1] y cómo hacerlo sin que tal reducción suponga una pérdida "inaceptable" del número o calidad de las acepciones presentes? En otras palabras, ¿cómo confeccionar un subcorpus[2] que –tanto desde el punto de vista cualitativo como cuantitativo– represente estrictamente el corpus del cual procede? Esta es la idea central de este artículo.

Es de notar que, tanto el corpus como el término elegidos para resolver la cuestión deberán considerarse instrumentales, es decir, seleccionados para ejemplificar la secuencia lógica de nuestro razonamiento, pues la solución buscada deberá poder aplicarse a toda colección o corpus existente y a cualquier entrada léxica posible.

La hipótesis de base es que la clave para resolver el problema habrá de encontrarse entre los principios, métodos y fórmulas de la estadística.

Por lo que se refiere a la estructura de este artículo, además de la presente introducción (§ 0.), nuestro estudio constará de cuatro partes principales y unas breves conclusiones. En la primera parte (§ 1.), trataremos de definir la entrada *externo* a partir de lo propuesto en el *DRAE* y de dos corpus-n del CREA (el de Argentina y el de España). Tal estudio no deberá entenderse como necesario para aplicar nuestra fórmula –que deberá ser completamente independiente de cualquier estudio previo–, sino como instrumento para luego poder poner a prueba nuestra hipótesis. En la segunda parte (§ 2.), será el momento (i) de proponer una fórmula estadística que nos permita reducir el número de entradas de los corpus-n para crear, así, corpus-nr y (ii) de realizar los

[1] No consideraremos en este estudio la lexicografía bilingüe, pero –como se podrá observar– el tipo de solución propuesta podría aplicarse también a estos diccionarios. Para consideraciones amplias sobre la lexicografía en general, véanse los manuales clásicos de Seco (1987), Casares (1992), Lara (1997) y Porto Dapena (2002). Véanse, además, dos obras fundamentales de Castillo Peña, una sobre la definición sinonímica (1992 y 1993) y otra (2000-2001) sobre la obra antes mencionada de Lara.

[2] Se impone, antes de iniciar, una consideración terminológica. El término *corpus* puede interpretarse de muchas maneras, entre ellas, como colección de obras; así, el portal *Contrastiva.it* constituye una colección de diccionarios y de gramáticas, pero no es este el sentido que queremos darle aquí. En este estudio, llamamos *corpus* a cualquier colección de textos que permita a un lexicógrafo establecer las distintas acepciones de un lema. Por un lado, tales colecciones pueden contener un número extremadamente alto de textos o estar compuesto por una cantidad considerablemente pequeña de ellos. Por otro, puede que dos corpus se hayan originado en ambientes completamente distintos o que uno haya tenido origen de la reducción del otro. A todos los denominaremos *corpus*. El corpus de mayores dimensiones al cual haremos referencia en este estudio es el *Corpus de referencia del español actual* de la Real Academia Española y a él nos referiremos siempre con su sigla (*CREA*). Cuando seleccionemos el lema que utilizaremos para llevar a cabo nuestro estudio –que en nuestro caso será *externo, na*– haremos uso de dos corpus (o subcorpus) del *CREA*. El primero estará constituido por los ejemplos del lema *externo, na* cuya proveniencia diatópica está marcada por pertenecer a un país de habla española y lo llamaremos *corpus-n* ('corpus nacional'); el segundo, por una versión reducida del anterior y lo denominaremos *corpus-nr* ('corpus nacional reducido'). Por otra parte, cuando nos refiramos, en términos generales, a un corpus reducido cualquiera lo llamaremos *corpus-r* ('corpus reducido'). Como se sabe, la bibliografía sobre el uso de los corpus es amplísima, de algunas relaciones entre corpus y lexicografía destacamos Pérez Hernández (2002), Rojo (2009) y Gozález Ribao (2015).

cálculos –independientemente de lo estudiado en la primera parte– que nos llevarán a uno o más valores finales. En tercer lugar (§ 3.) se pondrán a prueba los valores estadísticos resultantes del § 2. gracias a los dos corpus-n estudiados íntegramente y, luego (§ 4.), se volverán a poner a prueba a partir de otro corpus-nr cuyo corpus-n de procedencia no ha sido estudiado.

1. LA DEFINICIÓN DE *EXTERNO*

Para esta parte, en términos prácticos:

1. Se ha decidido tomar como corpus general y de base el CREA (*Corpus de referencia del español actual)*. Tal elección se basa en que dicha colección de textos (i) es de grandes dimensiones, (ii) contiene representaciones léxicas –salvo pequeñas excepciones– de todas las áreas/naciones de habla hispánica, (iii) es de fácil consulta a través de un buscador cualquiera, (iv) constituye, quizás, el corpus léxico más utilizado en el ámbito de la lengua española y (v) goza de un prestigio indiscutible entre quienes se dedican a los estudios morfológicos, sintácticos o léxico-semánticos de español actual.
2. Se ha pensado abarcar el período 1975-2004 (la mayor amplitud temporal posible ofrecida por el CREA), si bien, para restringir algo el campo, se ha decidido delimitar la búsqueda solo al ámbito de la prensa, visto que buena parte de quienes se dedican a la norma lingüística del castellano consideran que la lengua de la prensa podría considerarse como modelo para la lengua culta del castellano.
3. Se ha seleccionado como término objeto de estudio el adjetivo *externo*, (i) por constituir una entrada léxica de variabilidad morfológica intermedia –si se la compara con la invariabilidad de un adverbio o la variabilidad extrema de un verbo–, (ii) por asegurar un número alto, pero de ningún modo excesivo de ejemplos –considérese, por ejemplo, el escaso número que presumiblemente se obtendría con un adjetivo como *desabido*, los innumerables casos de *grande* o los casi inabarcables de las preposiciones *a*, *de*, *para*, etc.– y (iii) por tratarse de un adjetivo que puede tener, además, un uso sustantivo.

Como se acaba de decir, en este apartado se tratará de definir la entrada *externo, na* a partir de lo que se propone en el *DRAE* y en dos corpus-n suficientemente numerosos del CREA (el de Argentina y el de España).

Por un lado, en la última versión en línea del *DRAE*, para la entrada *externo, na* se proponen dos acepciones y seis formas complejas:

> *externo, na*
>
> 1. adj. Dicho de una cosa: Que obra o se manifiesta al exterior, en comparación o contraposición con lo interno.
> 2. adj. Dicho de un alumno: Que solo permanece en el colegio o escuela durante las horas de clase. U. t. c. s.

ángulo externo
conducto auditivo externo
culto externo
economías externas
oído externo
otitis externa

Por otro, si se cuantifican –después de descartar algún ejemplo no pertinente– los casos de *externo*, *externa*, *externos* y *externos* presentes en el CREA[3], se obtienen los siguientes baremos:

país	*externo*	*externos*	*externa*	*externas*	total
Argentina	64	60	177	35	336
Bolivia	17	17	50	9	93
Chile	11	15	16	6	48
Colombia	59	42	54	17	172
Costa Rica	12	9	24	4	49
Cuba	24	10	33	8	75
Ecuador	8	3	20	1	32
El Salvador	6	3	10	6	25
EE. UU.	16	10	37	12	75
España	244	253	504	223	1224
Filipinas	0	0	0	0	0
Guatemala	31	24	43	12	110
Honduras	2	7	38	2	49
México	75	85	148	54	360
Nicaragua	9	12	33	9	63
Panamá	0	2	10	1	13
Paraguay	16	11	22	3	52
Perú	11	9	51	7	78
Puerto Rico	0	0	3	0	3
Rep. Dominicana	14	10	34	4	62
Uruguay	28	22	40	11	101
Venezuela	38	40	165	28	271
totales	**689**	**645**	**1521**	**455**	**3310**

Tabla 1: Casos de externo *(y variantes) en el CREA*

Con la finalidad de constatar si las dos acepciones consideradas por el *DRAE* eran adecuadas[4] y suficientes para interpretar correctamente todos los casos presentes

[3] La búsqueda de datos se realizó el 16/11/2018.

[4] Como se observará, más adelante, no tomaremos en consideración ningún tipo de significado figurado. Por lo que se refiere al concepto de polisemia léxica y acepción lexicográfica, véanse Cruce (1986), Vivanco Cervero (2003), Battaner (2008), Battaner y Torner Castells (2008) y, para la microestructura del diccionario, Garriga Escribano (2003). En nuestro caso, el procedimiento para evaluar la adecuación de las dos acepciones académicas se basó en la aplicación de la prueba de *sustitución* o *conmutación* (Medina Guerra 2003: 136-138), es decir, en el reemplazo del lema en cuestión por las acepciones propuestas.

en el CREA, se observó y sopesó individualmente cada uno de los casos presentes en dos de los corpus-n más numerosos (España, 1224 casos y Argentina, 336).

El resultado de esta constatación dio resultado negativo, pues solo en pocas ocasiones las definiciones del *DRAE* resolvían con agilidad la interpretación de dichos ejemplos. Y esto, según nuestro criterio, por dos razones principales y complementarias, pues algunas imprecisiones en las definiciones académicas impedían la perfecta interpretación de muchos ejemplos y, además, el número de las acepciones académicas no era suficiente para abarcar todas las significaciones presentes en los dos corpus-n observados.

Por lo que se refiere a las "imprecisiones" de las definiciones académicas, creemos que estas surgían de dos cuestiones problemáticas:

a. Una se relacionaba con la primera acepción del *DRAE* (*Dicho de una cosa: Que obra o se manifiesta al exterior, en comparación o contraposición con lo interno.*), pues, aunque era evidente que muchísimos casos del corpus orbitaban en su área de significación, a muy pocos de ellos les calzaba perfectamente. Esto, creemos, porque *obrar* presupone una significación activa y *manifestarse*, una pasiva y, tal como está redactada la definición, ambas perspectivas parecerían, en cierto sentido, rechazarse mutuamente. Además, mientras *manifestarse al externo* constituye un sintagma de interpretación clara ('manifestarse hacia el exterior'), *obrar al externo*, de ninguna manera lo es[5].

   Sopesando el significado de los ejemplos desde este punto de vista, se ha pensado que el aspecto clave para su interpretación se relaciona con los siguientes hechos:

   (i) *estar, obrar o manifestarse en el exterior* –y no ejercer ningún tipo de influencia evidente en el interior–,
   (ii) *obrar, manifestarse o proceder desde el exterior* –y ejercer dicha influencia–,
   (iii) *obrar, manifestarse o proceder hacia el exterior* –y ejercer alguna influencia que va del interior al exterior–.

   En otras palabras, se ha considerado que convendría desglosar la primera acepción académica en tres (sub)acepciones distintas:

   1a. *Dicho de algo o de alguien que está, obra o se manifiesta en el exterior.*
   1b. *Dicho de algo o de alguien que obra, se manifiesta o procede desde el exterior.*
   1c. *Dicho de algo o de alguien que obra, se manifiesta o procede hacia el exterior.*

b. La otra cuestión problemática se relaciona con la segunda acepción del *DRAE* (*Dicho de un alumno: Que solo permanece en el colegio o escuela durante las horas de clase*). Entre los 1560 ejemplos estudiados, ninguno podía interpretarse con dicha acepción. Por supuesto, tal circunstancia significa –pura y exclusivamente– que

---

Claramente, a la conmutación no se le exigía elegancia estilística sino, simplemente, adecuación sintáctica y semántica.

[5] Quizás el diccionario *Clave* trataba de resolver esta dificultad al proponer en sus páginas la siguiente reformulación de la definición académica: *Que está, actúa, se manifiesta o se desarrolla en el exterior*.

esa acepción no se halla entre los textos de periódicos españoles o argentinos incluidos en el CREA.

Ahora bien, se presenta un caso del corpus-n español que, por un lado, se acerca mucho al área de significación de esta segunda acepción académica y, por otro, no tiene ninguna relación con las restantes acepciones identificadas en el estudio y a las que haremos mención más adelante.

El ejemplo en cuestión es el siguiente:

[...] *será Chirac quien acapare los dividendos de haber elevado a los rumanos de la categoría de externos a la de mediopensionistas* (n. 56)[6].

Claro está, este ejemplo no puede interpretarse con la segunda definición académica, a no ser que se la module de la siguiente manera:

2. *Dicho de alguien cuya vivienda no coincide con su lugar de estudio o trabajo. U. t. c. s.*,

Es decir, solo si se ampliara un poco su extensión[7] –su alcance semántico– sería posible utilizarla para la interpretación del ejemplo 56.

En definitiva, de la observación y catalogación de los corpus-n de España y de Argentina se impone la conveniencia lexicográfica de reformular, como se acaba de indicar, las dos acepciones del *DRAE* de la siguiente manera[8]:

1a. *Dicho de algo o de alguien que está, obra o se manifiesta en el exterior.*
(ESP) "Como suele suceder, el mundo de las mujeres está lejos de la épica externa, confinado en las fronteras del matrimonio, la familia, la casa, los hijos." (n. 309)
(ARG) "El proyecto debe complementarse [...] con un sistema de iluminación y sonido externo en base a dos columnas metálicas laterales[...]." (n. 11)

1b. *Dicho de algo o de alguien que obra, se manifiesta o procede desde el exterior.*
(ESP) "Los resultados aconsejaron la extirpación por vía externa de la cuerda vocal." (n. 130)
(ARG) "[...] la pareja usa la hostilidad externa para descargar su propia hostilidad hacia afuera [...]" (n. 25)

1c. *Dicho de algo o de alguien que obra, se manifiesta o procede hacia el exterior.*
(ESP) "[...] la abadía de Samos finalizó el proyecto de una hospedería externa, con 16 habitaciones [...]" (n. 3)
(ARG) "La Municipalidad de la Ciudad de Buenos Aires consignó que [...] los hospitales municipales mantendrán una guardia similar a la de los fines de semana [...] sin atención en consultorios externos." (n. 8)

---

[6] Las indicaciones numéricas que aparecen de aquí en adelante junto a los ejemplos citados identifican tal ejemplo según la posición automática con que aparecen al consultarse el CREA. En este caso, por ejemplo, el texto citado aparece en quincuagésima sexta posición si se consulta el CREA con los siguientes filtros: *externos*, prensa, 1975-2004, España.

[7] En la segunda acepción del diccionario *Clave*, también se trata de ampliar la extensión y, por eso, se la reformula con los siguientes términos: *Referido a una persona, esp. un alumno, que no vive en el lugar en el que trabaja o en el que estudia.*

[8] Mantenemos el número 1 para las acepciones derivadas de la primera acepción académica y el 2 para la reformulación de la segunda. Para estas y para las siguientes acepciones mencionadas, daremos –siempre que contemos con uno– un ejemplo del corpus-n argentino y otro del español.

2. *Dicho de alguien cuya vivienda no coincide con su lugar de estudio o trabajo. U. t. c. s.*
(ESP) "[...] será Chirac quien acapare los dividendos de haber elevado a los rumanos de la categoría de externos a la de mediopensionistas." (n. 56)
(ARG) (sin ejemplos)

Ahora bien, de la observación y catalogación de los mencionados 336 ejemplos argentinos y 1224 españoles se impone una segunda constatación: las cuatro acepciones propuestas hasta aquí no son suficientes para cubrir interpretativamente todos los ejemplos obtenidos, serán necesarios al menos otras cinco para dar cuenta de la totalidad de dichos ejemplos.

Según nuestro criterio, las acepciones que deberían añadirse son las siguientes:

3. *Dicho de algo o de alguien, relacionado con el extranjero.*
(ESP) "El sucesor de ésta [...] dirigió recientemente una advertencia a Zia al señalar la «injerencia de intereses externos» en el fomento de los disturbios interiores indios." (n. 91)
(ARG) "[...] los títulos públicos [...] tuvieron un desarrollo favorable ante la perspectiva de una baja del riesgo país por la recompra de deuda externa." (n. 2)

4. *Dicho de algo o de alguien, ajeno a un determinado individuo, grupo social o institución.*
(ESP) "[...] será un grupo de expertos externos, es decir personal [...] ajeno a la institución docente, quienes [...] elaborarán un segundo informe sobre la calidad de la Universidad." (n. 148)
(ARG) "[...] la comisión asesora externa tiene [...] asignado un trabajo técnico [...]." (n. 56)

5. *En figuras con simetría bilateral, parte más alejada de su eje central.*
(ESP) "El lateral Mikel Lasa podría sufrir rotura de ligamentos cruzados externos y menisco, la temida 'triada'." (n. 172)
(ARG) "[...] el pibe [...] saca un zurdazo con la cara externa del pie –en un ángulo cerradísimo– para marcar el 1 a 0." (n. 166)

6. *En figuras geométrica (incluso tridimensionales), lo que está más cerca de un borde o de la superficie o lo que está en sus bordes o superficie.*
(ESP) "El preservativo, al no proteger los genitales externos del hombre, no logra impedir que éste transmita el HPV a la mujer." (n. 138)
(ARG) "En curva, la velocidad de la rueda externa es superior a la interna." (n. 41)

7. *Tipo de apuesta deportiva en la hípica. U. t. c. s.*
(ESP) (sin ejemplos)
(ARG) "Así llegamos al comienzo de la apuesta 5 y 6. Nos quedamos con Paca, porque la última vez que nos afirmamos con Congresista perdimos todas bancas [*sic*] en la externa." (n. 112)

Si pasamos ahora a las formas complejas[9], es necesario comentar que, a las seis formas complejas consideradas por el *DRAE* (a-f), podrían añadírseles al menos otras seis (g-l):

[9] Cabe aclarar que aquí por *formas complejas* entendemos colocaciones nominales, es decir, las asociaciones frecuentes de términos que forman un sintagma nominal y cuyo término común (en nuestro caso, *externo*) debe poder interpretarse según una cualquiera de sus acepciones. Si el último requisito no pudiera satisfacerse, es decir, si la interpretación del sintagma fuera más allá de la suma de sus componentes aislados, dicha secuencia léxica debería pertenecer a otra categoría fraseológica.

a. ángulo externo
b. conducto auditivo externo
c. culto externo
d. economía externa
e. oído externo
f. otitis externa

g. aspecto externo
h. auditoría externa
i. consulta externa
j. demanda externa
k. deuda externa
l. factor externo

La significación de *externo* o *externa* en estas formas complejas debería relacionarse con las acepciones propuestas de la siguiente manera: *ángulo externo* (acepción 1a), *economías externas* y *factores externos* (1b), *culto externo* y *aspecto externo* (1c), *demanda externa* y *deuda externa* (3 o 1b), *auditoría externa* y *consulta externa* (4) y *conducto auditivo externo*, *oído externo* y *otitis externa* (6).

Cabe, por último, un comentario sobre la frecuencia de uso de las acepciones de *externo* entre los ejemplos estudiados. La siguiente tabla propone la situación constatada entre los 1224 españoles y los 336 casos argentinos:

país		1a, 1b, 1c	2	3	4	5	6	7	total
España	(casos)	574	1	381	186	25	57	0	**1224**
	(%)	46,90	0,08	31,13	15,20	2,04	4,66	0,00	**100**
Argentina	(casos)	32	0	275	17	1	10	1	**336**
	(%)	9,52	0,00	81,85	5,06	0,30	2,98	0,30	**100**

Tabla 2: Número de casos y sus frecuencias de uso para las acepciones de externo *(y variantes)*

Existe una enorme diferencia entre las frecuencias de uso de las acepciones 1, 3 y 4 y las de las restantes (2, 5, 6 y 7), cuyos baremos –aunque difieran bastante entre sí– son todos marginales o, incluso, muy marginales.

Por otra parte, se observa que no hace falta que un término esté diatópicamente marcado para que su frecuencia de uso presente porcentajes dispares en distintos países de habla hispánica. En el caso de *externo* –un término sin marcas diatópicas o diastráticas extraído de dos corpus-n homogéneos– hay una evidente diversidad en la frecuencia de uso que se hace de las distintas acepciones en ámbito español y argentino.

Así, por un lado, la frecuencia de uso de la acepción 1 en España llega a un 46,90%, mientras que en Argentina apenas alcanza el 9,52% y, por otro, la frecuencia de la acepción 3 en Argentina es altísima (81,85%) y en España mucho más modesta (31,13%). Es bastante probable que tal "anomalía" pueda adjudicarse –al menos en parte– a que en los periódicos argentinos del período 1975-2004, la presencia de un tema relacionado con la *deuda externa*, por razones de situación histórica, haya sido

mucho más frecuente que en los españoles. En efecto, el término *deuda* aparece 147 veces en los 336 ejemplos de Argentina y solo 213 veces en 1224 ejemplos de España, es decir, en el 43,75% de los ejemplos argentinos y solo en el 17,40% de los españoles.

2. EL PROCEDIMIENTO ESTADÍSTICO UTILIZADO

En este apartado, como se ha dicho en la introducción del estudio, se buscará un procedimiento estadístico que permita –de un modo completamente independiente del estudio previo del corpus– reducir su número de entradas sin que tal reducción suponga una modificación significativa de los resultados que se obtendrían si se trabajara con el mismo corpus sin reducción. En otras palabras, se quiere responder a la siguiente pregunta: ¿cómo extraer un corpus-r que –tanto desde el punto de vista cualitativo como cuantitativo– represente lo más estrictamente posible el corpus del cual proviene?

Con la finalidad de lograr este objetivo, recurriremos a la estadística[10] para determinar el *número estrictamente representativo* de un corpus. Donde por *número estrictamente representativo* entendemos la cantidad mínima de ejemplos necesarios para reproducir sin variaciones significativas el conjunto de sus acepciones y sus correspondientes frecuencias de usos, siempre que estas frecuencias sean mayores que un porcentaje determinado. Todo esto, deberá ser posible con un elevado nivel de confianza y un limitado margen de error[11].

De esta manera –y por lo que se refiere a las frecuencias de uso–, extrayendo un subconjunto (o corpus-r) de casos, se quiere construir un intervalo de frecuencia que con una alta probabilidad sea capaz de contener en su interior las mismas frecuencias de uso que se habrían observado si se hubiera sido capaces de analizar todos los casos de un determinado corpus.

El procedimiento que hemos seguido para lograr este *número estrictamente necesario* tal como la acabamos de delimitar consta de una serie de fases: (i) se determina *a priori* el nivel mínimo de frecuencias que se quiere reproducir (p), se elige el margen de error (ε) que estamos dispuestos a tolerar (y que determinará la amplitud del intervalo) y se establece el nivel de confianza buscado ($1-\alpha$); (ii) se parte de la fórmula estadística del intervalo de frecuencias para determinar la fórmula de la cantidad estrictamente necesaria (n) para reproducir el nivel mínimo de frecuencia (p) en el corpus; (iii) se calcula n a partir de la adopción de un valor determinado para el *error* (ε).

[10] No estrictamente relacionados nuestro presente estudio, pero de gran interés para sus posibles desarrollos, véanse las consideraciones de Rojo (2017) sobre la relación entre corpus y estadísticas y las que presentan Torruella Casañas y Capsada Blanch (2013 y 2017) sobre riqueza léxica y estadísticas.

[11] Para cada acepción, se busca un intervalo de frecuencia cuya amplitud máxima sea del 5% o del 2% (márgenes de error tolerados) de modo que, con un 95% de probabilidad (nivel de confianza), este contenga la frecuencia que dicha acepción tiene en el corpus general. Para estos conceptos de la estadística y para cualquier otro concepto al que se haga referencia en el curso de este artículo, véase Piccolo (2004) u otro buen manual general.

(i) *Nivel mínimo de frecuencia reproducida, margen de error tolerado y nivel de confianza.*

Las elecciones realizadas *a priori* son las siguientes:

1. Son dos los niveles mínimos de frecuencias (*p*) que se quieren reproducir, uno es del 3% y otro del 1%[12].
2. El margen de error admitido es del 5% cuando *p* es igual al 3% y del 2% cuando *p* es igual al 1%[13].
1. El nivel de confianza se ha fijado en un 95% para ambos niveles mínimos de frecuencia[14].

(ii) *Fórmula del intervalo de frecuencia* (Piccolo, 2010)

Si se indica con $\hat{f}$ la frecuencia calculada en el subgrupo de acepciones seleccionadas, el procedimiento estadístico dará origen a un intervalo de estima de la probabilidad buscada, igual a

$$\hat{f} \pm z_{\alpha/2}\sqrt{\frac{\hat{f}(1-\hat{f})}{n}}$$

Ecuación 1: Intervalo de confianza para la frecuencia

donde $z_{\alpha/2}$ es un valor de las tablas normales estandarizadas determinado por el nivel de confianza seleccionado en la fase (i). Como consecuencia, el intervalo de frecuencia calculado en *Ecuación 1* deberá contener la probabilidad indicada en *Tabla 2* con un nivel de confianza igual a (1-α)%. Generalmente, α será un valor pequeño (1% o, más frecuentemente, 5%), de esta manera, el intervalo identificado tendrá una probabilidad elevada de reproducir la frecuencia de dicha *Tabla*.

Nuestro problema empírico se traduce en ser capaces de determinar el número de casos (*n* en la *Ecuación 1*) que asegure un cierta amplitud del intervalo (error) y que contenga la probabilidad desconocida (*p*) indicada por la *Tabla 2* con un nivel de confianza de (1-α)%. A partir de la *Ecuación 1* se obtiene la siguiente fórmula:

$$n \geq \frac{4 * z_{\alpha/2}^2(p(1-p))}{\varepsilon^2}$$

Ecuación 2: Determinación del número representativo

Al examinar la *Ecuación 2*, se observa que el número buscado depende exclusivamente de tres cantidades (*p*, *ε* y 1-*a*) fijadas en la fase (i) y que son

[12] Esto asegurará que, en corpus-r seleccionado, se reproduzcan las acepciones que en el corpus completo se presenten con una frecuencia a superior o igual al 3% o al 1%.
[13] Tales valores (5% o 2%) determinarán las amplitudes admitidas del intervalo de frecuencia.
[14] Este valor nos asegurará que en el 95% de los corpus-n posibles se reproduzcan los niveles mínimos de frecuencia buscados.

completamente independientes del tamaño del corpus original. Estas consideraciones son las que nos aseguran que el número resultante de esta fórmula será válido tanto para un corpus general de 500 casos como para uno de 5.000.000.

(iii) *Cálculo de número estrictamente representativo*

1. Si queremos determinar la cantidad de casos que nos asegure poder reproducir todas aquellas frecuencias de la *Tabla 2* que resulten iguales o superiores al 3% y, admitiendo para esto, un intervalo de frecuencia de amplitud máxima igual al 5%, esta es la fórmula resultante:

$$n \geq \frac{4 \times 1.96^2(0.03 \times 0.97)}{0.05^2} = 179$$

Fórmula 1: Número para frecuencias mayores o iguales al 3%

En nuestro caso concreto, esto significa que, en cambio de observar los 336 casos de Argentina o los 1224 de España, los corpus-n de ambos países podrán reducirse a 179 ejemplos cada uno y, aun así, representar correctamente todas las probabilidades de frecuencias cuyos valores sean mayores o iguales al 3% de la *Tabla 2*.

Claro está que, si los corpus-n no llegaran a 179 casos –como se constata en el § 5.2. para todos los países menos Argentina, España y México–, será necesario observar todos los casos disponibles y no podrán ser objeto de ningún tipo de reducción, pues el número requerido mínimo será siempre 179 o la totalidad de los casos.

2. Si quisiéramos mejorar la representación ofrecida por los 179 casos, tratando de reproducir todas aquellas acepciones cuyos significados presenten una frecuencia de al menos el 1% con un margen de error del 2%, la cantidad mínima de casos que deberían observarse sería igual a 381. Pues

$$n \geq \frac{4 \times 1.96^2(0.01 \times 0.99)}{0.02^2} = 381$$

Fórmula 2: Número para frecuencias mayores o iguales al 1%

También en este caso, si los corpus-n no llegaran a 381 casos –como se constata en el § 5.2. para todos los países menos España–, sería necesario observar todos los casos disponibles y dicho corpus no podría ser objeto de ningún tipo de reducción, pues el número mínimo requerido será siempre 381 o la totalidad de los casos de los que se dispone.

3. LAS PRUEBAS DE ESPAÑA Y ARGENTINA

En este apartado, se someterán a una prueba empírica las cantidades resultantes en el apartado anterior (§ 2.) y esto se hará a partir de los dos corpus-n íntegramente conocidos (España y Argentina) y ya estudiados en el primer apartado (§ 1.). El objetivo

de tal prueba será comprobar la validez de la fórmula propuesta y, sobre todo, la validez de la cantidad *n* resultante.

3.1. LA PRUEBA CON DOS CORPUS-NR DE 179 CASOS

Utilizando *R*[15], del corpus-n español (1224 casos totales) y del argentino (336 casos) se extrajeron aleatoriamente 1000 corpus-nr distintos de 179 casos para cada país y en ellos se analizó la presencia de acepciones (¿cuáles de las siete acepciones posibles estaban presentes?) y sus frecuencias en cada corpus-nr. Los resultados obtenidos se transcriben en las *Tablas 3* y *4*, tablas en las que se reemplaza las frecuencias realmente constatadas por sus frecuencias medias estimadas de los 1000 corpus-nr en su conjunto y, además, se añade la desviación estándar (o variabilidad relativa) de dichos 1000 corpus-nr con respecto a los corpus-n (de España o de Argentina) que les dieron origen. Los datos relacionados con España son los siguientes:

ESPAÑA

acepciones en corpus-n	**1a, 1b, 1c**	**2**	**3**	**4**	**5**	**6**	**7**	**total**
casos en corpus-n	574	1	381	186	25	57	0	**1224**
frec. en corpus-n (%)	46,90	0,08	31,13	15,20	2,04	4,66	0,00	**100**
frecuencias medias en los 1000 corpus-nr (%)	46,98	0,06	30,99	15,23	2,11	4,66	0,00	–
desviación estándar de los 1000 corpus-nr (%)	3,69	0,19	3,45	2,63	1,01	1,61	0,00	–
número de corpus-nr con significados requeridos	1000	132	1000	1000	979	1000	0	–

Tabla 3: El corpus-n de España y sus 1000 corpus-nr de 179 casos

Cabe señalar que, tal como se esperaba, al observarse solo 179 casos en lugar de los 1224 casos totales, las frecuencias de uso con porcentajes superiores o iguales al 3% del corpus-n (acepciones 1, 3, 4 y 6) se repiten correctamente en todos los corpus-nr de 179. Esto significa no solo que el 100% de los corpus-nr refleja todas las acepciones con frecuencia superior o iguales al 3%, sino que también las reproduce sin variaciones significativas[16].

Además, se puede observar que, incluso para el significado 5 –cuya frecuencia de uso en la población total de 1224 casos es de 2,04%, es decir, menor del 3%– su representación correcta se observa en un número muy elevado de corpus-nr (979 de 1000).

[15] Software libre que permite realizar análisis estadísticos complejos (R Development Core Team 2010).
[16] Para la acepción 1 se constata una frecuencia de uso del 46,90% en el corpus total y un 46,98% en la media de los corpus-r; para la acepción 3, los valores son 31,13% y 30,99% respectivamente; para la 4, 15,20% y 15,23%; y para la acepción 6, 4,66% en ambos casos.

Por otra parte, la desviación estándar[17] presentada en la *Tabla 3* muestra que las estimaciones obtenidas en los distintos corpus-nr son muy similares o iguales.

Los datos relacionados con Argentina, por su parte, son los siguientes:

ARGENTINA

acepciones en corpus-n	**1a, 1b, 1c**	**2**	**3**	**4**	**5**	**6**	**7**	**total**
casos en corpus-n	32	0	275	17	1	10	1	**336**
frec. en corpus-n (%)	9,52	0,00	81,85	5,06	0,30	2,98	0,30	**100**
frecuencias medias en los 1000 corpus-nr (%)	9,56	0,00	81,75	5,08	0,74	3,01	0,73	–
desviación estándar de los 1000 corpus-nr (%)	2,12	0,00	2,89	1,71	0,31	1,26	0,31	–
número de corpus-nr con significados requeridos	1000	0	1000	1000	399	995	427	–

Tabla 4: El corpus-n de Argentina y sus 1000 corpus-nr de 179 casos

En el caso de Argentina, al observarse solo 179 casos en lugar de los 336 casos totales, las frecuencias de uso con porcentajes superiores o iguales al 3% del corpus-n (acepciones 1, 3 y 4) se repiten correctamente en todos los corpus-nr extraídos. Y esto, como habíamos dicho para España, significa no solo que el 100% de los corpus-nr refleja todas las acepciones con frecuencia superior o iguales al 3%, sino también que reproduce dichas frecuencias sin variaciones significativas[18].

La acepción 6 merece un comentario aparte, pues esta, que presenta una frecuencia ligeramente menor al·3% en la población total de casos argentinos (2,98%), en las medias de los corpus-n llega a 3,01%. Esta marginalidad, este posicionamiento en el límite más bajo de nuestro rango de frecuencias hace que tal acepción se constate no en el 100% de los corpus-nr generados, sino "solo" en el 99,5% (995 presencias entre las 1000 posibles). Tal situación se resuelve a favor de nuestra propuesta si consideramos el concepto estadístico de *nivel de confianza*. El nivel de confianza (1-α)% describe la proporción de muestreos (corpus-r) que determina un intervalo de frecuencia y contiene la frecuencia *p* que se busca. Al contrario, tendremos 5% de muestras en las que tal intervalo estimado no contendrá la probabilidad buscada. En el ejemplo ilustrado en la *Tabla 4* hay 5 muestras de las 1000 generadas en las que el significado 6 no se ha observado. En otras palabras, desde el punto de vista estadístico, no debemos preocuparnos de que un valor marginal como el que estamos comentando se presente en el 99,5% de los casos y no en el 100% de ellos.

Los significados con una frecuencia inferior al 3% se observaron solo en menos de la mitad de los corpus-nr generados, pero esto era lo esperado, pues el método

[17] Con la desviación estándar se indica aquí cuánto varía la frecuencia de cada uno de los 1000 corpus-nr con respecto a las frecuencias medias indicadas en la tabla.

[18] Para la acepción 1 se constata una frecuencia de uso del 9,52% en el corpus total y un 9,56% en la media de los corpus-r; para la acepción 3, los valores son 81,85% y 81,75% respectivamente; para la 4, 5,06% y 5,08%; y para la 6, 2,89% y 3,01%.

aplicado garantiza la reproducción segura únicamente de los significados que tienen una frecuencia de uso superior o igual al 3%.

Por último, conviene resaltar que la desviación estándar presentada en la *Tabla 4* muestra que las estimaciones obtenidas en los distintos corpus-nr son muy similares en los 1000 corpus-nr observados.

3.2. LA PRUEBA CON UN CORPUS-NR DE 381 CASOS

Si se aplica al corpus-n español el mismo procedimiento detallado en el apartado anterior tanto a España como a Argentina, pero reduciéndolo no ya a 179, sino a 381 casos[19], a partir de otros 1000 nuevos corpus-nr generados se obtienen los siguientes datos:

ESPAÑA								
acepciones en corpus-n	**1a, 1b, 1c**	**2**	**3**	**4**	**5**	**6**	**7**	**total**
casos en corpus-n	574	1	381	186	25	57	0	**1224**
frec. en corpus-n (%)	46,90	0,08	31,13	15,20	2,04	4,66	0,00	**100**
frecuencias medias en los 1000 corpus-nr (%)	46,96	0,06	31,03	15,18	2,05	4,69	0,00	–
desviación estándar de los 1000 corpus-nr (%)	2,53	0,10	2,31	1,87	0,72	1,08	0,0	–
número de corpus-nr con significados requeridos	1000	271	1000	1000	999	1000	0	–

Tabla 5: El corpus-n de España y sus 1000 corpus-nr de 381 casos

Como se puede notar, todas las frecuencias mayores o iguales al 1% están representadas de modo más exacto y con menor variabilidad –menor desviación estándar– que con los corpus-nr españoles de 179 casos[20]. Por lo que se refiere a la acepción 5 –presente "solo" en 999 corpus-nr de los 1000 observados– vale, incluso con más fuerza, la argumentación propuesta para la acepción 6 de los corpus-nr argentinos, pues aquí se llega a cubrir el 99,9% delos corpus-nr generados.

4. LA PRUEBA DE MÉXICO

En este apartado se quiere, nuevamente, poner a prueba las fórmulas y los valores obtenidos en el § 2., no ya evaluándolos comparativamente con dos corpus-n conocidos íntegramente[21], sino utilizando un corpus-nr de 179 casos generado a partir de un

[19] Recordamos que la creación de corpus-nr de 381 casos no es aplicable al corpus argentino, pues este consta ya en su versión completa de 336 casos, es decir, menos de los 381 que aquí se requieren.

[20] La desviación estándar de la acepción 1 era de 3,69% en los corpus-nr de 179 casos y 2,53% en los de 381 casos; en la acepción 2, de 0,19% y 0,10% respectivamente; en la 3, de 3,45% y 2,31%; en la 4, de 2,63% y 1,87%; en la 5, de 1,01% y 0,72%; en la 6, de 1,61% y 1,08%; y en la 7, de 0% en ambos casos.

[21] Nos referimos a los de España y Argentina que hemos estudiado en el § 1. y que nos han servido para las comprobaciones del § 3.

corpus-n del cual conocemos solo el número de casos que lo constituyen. Esto con la finalidad de comprobar si las acepciones y frecuencias resultantes de esta segunda prueba coinciden con lo observado para el lema *externo* en España y Argentina.

En primer lugar, se ha extraído aleatoriamente del CREA un corpus-nr de *externo* (y variantes) con 179 casos a partir de corpus-n de México que cuenta con 360 casos totales y luego se le ha adjudicado una acepción a cada uno de tales 179 casos. El número de casos resultante para cada acepción y sus frecuencias relativas son las siguientes:

MÉXICO

acepciones en corpus-nr	**1a, 1b, 1c**	**2**	**3**	**4**	**5**	**6**	**7**	**total**
casos en corpus-nr	31	1	128	14	1	4	–	**179**
frec. en corpus-nr (%)	17,32	0,56	70,95	8,28	0,56	2,24	–	**100**

Tabla 6: El corpus-nr de 179 casos para México

Para poder de comparar aquí más fácilmente estos datos –especialmente la presencia de acepciones y sus frecuencias de uso– con los obtenidos a partir de los corpus-n de España (1224 casos) y Argentina (336 casos) proponemos ahora la *Tabla 7*, en la que se retoman parcialmente los datos ya considerados en la *Tabla 2*:

	1a, 1b, 1c	**2**	**3**	**4**	**5**	**6**	**7**	**total**
frec. en corpus-nr/Esp. (%)	46,90	0,08	31,13	15,20	2,04	4,66	–	**100**
frec. en corpus-nr/Arg. (%)	9,52	0,00	81,85	5,06	0,30	2,98	0,30	**100**

Tabla 7: Frecuencia de uso de acepciones de externo *(y variantes) en España y Argentina*

Si se comparan ambas tablas, se puede observar, no solo que los datos obtenidos a partir de este corpus-nr mexicano coinciden con los obtenidos a partir de los corpus-n de España y Argentina[22], sino también que en él se constatan casos que, en principio podrían no haberse hallado, visto sus frecuencias inferiores al 3% (acepciones 2, 5 y 6).

Ahora bien, los datos que acabamos de mencionar no son, por ahora, significativos desde el punto de vista estadístico, aunque intuitivamente parezca que sí. Para justificar estadísticamente esta intuición, es decir, para que se nos permita generalizar los datos obtenidos y asegurar que el corpus-nr mexicano de 179 casos representa estrictamente el corpus-n del que procede, o sea, que en sus 179 casos se reflejan (con frecuencias similares) todas las acepciones que el corpus-n de 360 casos contiene con frecuencias mayores o iguales al 3%, es necesario aplicar un test estadístico que ponga a prueba tales datos: el denominado test binomial (Piccolo, 2010). Con este test lo que se hace es confirmar una *hipótesis* (normalmente denominada *nula*) que, en nuestro caso, podría expresarse como sigue: *en la población de referencia –el corpus-n de 360 casos mexicanos– la frecuencia de cada una de sus acepciones no diferirá (o no estará muy lejos) de las frecuencias observadas en el corpus-nr de 179 casos proveniente del corpus-n de 360 casos antes mencionado.*

[22] Es decir, están presentes todas las acepciones con frecuencias superiores o iguales al 3% (las 1, 3 y 4) y dichas frecuencias son comparables con las de Argentina y México.

Al aplicar el test a las seis acepciones presentes en el corpus-nr mexicano de 179 casos se han obtenido los siguientes resultados:

acepción	hipótesis nula	hipótesis alternativa	p-value	intervalo de confianza al 95%	frecuencias en corpus-nr
1	frecuencias en los 360 casos no igual al 17%	frecuencias en los 360 casos distinta del 17%	0,9207	[12,08-23,67]	17,32
2	frecuencias en los 360 casos igual a 0,1%	frecuencias en los 360 casos distinta del 0,1%	0,1640	[0,01-3,07]	0,56
3	frecuencias en los 360 casos igual al 70%	frecuencias en los 360 casos distinta del 70%	0,8074	[63,71-77,48]	70,95
4	frecuencias en los 360 casos igual al 10%	frecuencias en los 360 casos distinta del 10%	0,5346	[4,77-13,44]	8,38
5	frecuencias en los 360 casos igual al 0,1%	frecuencias en los 360 casos distinta del 0,1%	0,1640	[0,01-3,07]	0,56
6	frecuencias en los 360 casos igual al 2%	frecuencias en los 360 casos distinta del 2%	0.7852	[0,61-5,62]	2,24

Tabla 8: Resultados del test binomial aplicado a un corpus-nr mexicano de 179 casos

Como se puede observar, las seis hipótesis nulas[23] se confirman en las cuatro acepciones cuyas frecuencias son mayores o iguales al 3%, pues sus *p-value* son mayores de 0,5, esto es, se acercan a 1 (0,9207 para la acepción 1; 0,8074 para la 3; 0,5346 para la 4 y 0.7852 para la 6) y, por tanto, las frecuencias de estas acepciones en el corpus-n mexicano de 360 casos pueden considerarse cercanas o idénticas a las observadas en el corpus-nr de 179 casos. La hipótesis nula, en cambio, no se confirma para las acepciones 2 y 5, pues sus *p-value* son menores de 0,5 (0,1640 tanto para la acepción 2 como para la 5); pero tales acepciones tienen frecuencias inferiores al 3% y, por consiguiente, ya desde el inicio quedaban fuera del alcance previsto para los corpus-nr de 179 casos.

Por último, observamos que los valores de todas las frecuencias constatadas en el corpus-nr mexicano[24] se encuentran comprendidas entre los valores de los intervalos de confianza estimados según un nivel de confianza del 95%, lo que hace que podamos considerar plenamente representativos lo constatado en el corpus-nr de 179 casos. En otras palabras, se puede asegurar que las frecuencias observadas en los 179 casos analizados proporcionan una representación significativa de la distribución real del corpus-n mexicano de 360 casos.

[23] Creemos conveniente dedicar, para quien lo necesite, unas pocas palabras sobre qué se entiende, en general, por test estadístico y cómo deberán interpretarse sus resultados. Estos test contraponen una *hipótesis nula* a otra *hipótesis* (denominada *alternativa*) para aceptar una y rechazar la otra. Estas dos hipótesis deben ser alternativas, es decir, deben excluirse mutuamente. En nuestro caso la *hipótesis nula* es que *un subcorpus-r de 179 casos no difiere del subcorpus de cual procede* y la *alternativa* consiste en que *dicho subcorpus-r difiere de su subcorpus de proveniencia*. Dos son los valores principales que se obtienen a partir de estos test: un *valor de p* y un *intervalo de confianza*. El *valor de p* (o *p-value*) va de 0 a 1 y si se acerca a 1 considera aceptable la hipótesis nula, mientras que, si se acerca a 0, la rechaza. El *intervalo de confianza* indica los límites entre los cuales se debería ubicar un determinado valor para que se considere alta la probabilidad de acierto.

[24] Véase la última línea de la Tabla 6.

5. CONCLUSIONES

En el presente trabajo, se ha demostrado que, si se quiere redactar –a partir de un corpus preexistente de textos– un artículo lexicográfico en el que se reflejen, no solo las acepciones más representativas de un lema (esto es, las más frecuentes), sino también, sus frecuencias relativas, será suficiente definir las acepciones de un subcorpus de 179 casos extraídos aleatoriamente de un corpus mayor o de 381 casos. Se tratará, en ambas posibilidades, de un corpus-r, pero estricta y estadísticamente representativo de la totalidad de los casos presentes en el mayor del cual proviene.

En otras palabras, trabajando con un subcorpus de 179 casos tomados aleatoriamente de otro corpus más amplio, el lexicógrafo está estadísticamente seguro de hallar todas aquellas acepciones cuyas frecuencias de uso sean mayores o iguales al 3% y de poder asignar a cada una de dichas frecuencias un valor muy cercano al que habría encontrado si hubiera observado la totalidad de los casos de que disponía. Esta situación se da independientemente de la magnitud concreta del corpus mayor del que se disponga, pues 179 casos reflejarán con la misma exactitud un corpus general de 180 casos como uno sensiblemente mayor. Por lo que se refiere a las acepciones con frecuencias más pequeñas (del 3% o muy cercanas al 3%) lo dicho anteriormente se confirmará en el 95% de los subcorpus extraíbles, pues –como se ha fijado un nivel de confianza del 95%– existe la posibilidad de que estas acepciones con frecuencias más pequeñas no se encuentren presentes en el 5% de los subcorpus extraíbles. El lexicógrafo, por supuesto, también podrá hallar en los subcorpus acepciones cuyas frecuencias estén por debajo del 3%, pero no podrá considerarse estadísticamente seguro de que esto deba suceder como tampoco de que, en el caso de que suceda, las acepciones halladas constituyan la totalidad de las menos frecuentes ni de que las frecuencias con las que se asocian sean representativas. Por último, el lexicógrafo estará autorizado a reducir solo los corpus que sobrepasen los 179 casos, pues no existe posibilidad estadística de reducir corpus menores y obtener con tal operación el mismo grado de detalle.

Análogamente, podrá confirmarse una situación muy similar para los subcorpus de 381 casos. Las únicas diferencias dignas de mención se deben al hecho de que el estar constituidos por un mayor número de casos asegura la inclusión en dichos subcorpus de todas aquellas acepciones cuyas frecuencias de uso sean mayores o iguales al 1% (y no ya al 3%). Además, como en el subcorpus de 179 casos, las acepciones con frecuencias más bajas de las previstas (del 1% o muy cercanas a 1%) podrán no aparecer y no se podrá tener seguridad estadística sobre ninguna acepción menor del 1% que pueda aparecer "casualmente" en dichos corpus-r.

BIBLIOGRAFÍA

BATTANER, M. Paz (2008): "El fenómeno de la polisemia en la lexicografía actual: otra perspectiva", *Revista de Lexicografía*, XIV, pp. 7- 25.

BATTANER, M. Paz; TORNER CASTELLS, Sergi (2008): "La polisemia verbal que muestra la Lexicografía", en Azorín Fernández, Dolores *et al.* (coord.) *El diccionario como puente entre las lenguas y culturas del mundo. Actas del II Congreso Internacional de Lexicografía Hispánica*, pp. 201-216.

CASARES, J. (1992): *Introducción a la lexicografía moderna*, Madrid: CSIC.

CASTILLO PEÑA, Carmen (1992): "La definición sinonímica y los círculos viciosos", *Boletín de la Real Academia Española*, 72, 257, pp. 463-566.

___ (1993): "La definición sinonímica y los círculos viciosos (continuación)", *Boletín de la Real Academia Española*, 73, 258, pp. 131-213.

___ (2000-2001): "La naturaleza del diccionario: (A propósito de la Teoría de diccionario Monolingüe, de Luis Fernando Lara)", *Revista de lexicografía*, 7, pp. 201-224.

Clave. Diccionario de uso del español actual (2012): Madrid: Ediciones S. M. [en línea] <http://clave.smdiccionarios.com/app.php>.

CRUCE, Alan (1986): *Lexical Semantics,* Cambridge: Cambridge University Press.

GARRIGA ESCRIBANO, Cecilio (2003): "La microestructua del diccionario: las informaciones lexicográficas", en Medina Guerra, Antonia M. (coord.): *Lexicografía española*, Barcelona: Ariel Lingüística, pp. 103-126.

GOZÁLEZ RIBAO, Vanessa (2015): "Sobre algunos conflictos de la 'pre'-lexicografía: la selección de corpus para la elaboración de un diccionario contrastivo alemán-español", en Domínguez Vázquez, María José; Gómez Guinovart, Xavier; Valcárcel Riveiro, Carlos (eds.): *Lexicografía de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva*, Berlin/München/Boston: de Gruyter, vol. II, pp. 247-270.

LARA, Luis Fernando (1997): *Teoría del diccionario monolingüe*, México: El Colegio de México.

MEDINA GUERRA, Antonia M.ª (2003): "La microestructua del diccionario: la definición", en Medina Guerra, Antonia M. (coord.): *Lexicografía española*, Barcelona: Ariel Lingüística, pp. 126-146.

PÉREZ HERNÁNDEZ, M. Chantal (2002): "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento", *ELiEs*, 18 <http://elies.rediris.es/elies18/index.html>.

PICCOLO, Domenico (2010): *Statistica*, 23 ed., Bolonia: Il Mulino.

PORTO DAPENA, J. A. (2002): *Manual de técnica lexicográfica*, Madrid: Arco/Libros.

R DEVELOPMENT CORE TEAM (2010): *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Viena, 7.

REAL ACADEMIA ESPAÑOLA (2001): *Diccionario de la lengua española* (versión electrónica 23.2.). < https://dle.rae.es/>.

REAL ACADEMIA ESPAÑOLA: *Corpus de referencia del español actual*, Banco de datos (CREA) [en línea] < http://corpus.rae.es/creanet.html>.

ROJO, Guillermo (2009): "Sobre la construcción de diccionarios basados en corpus", *Tradumatica* (revista electrónica), 7/2009 .

ROJO, Guillermo (2017): "Sobre la configuración estadística de los corpus textuales", *Lingüística*, 33/1, pp. 121-134.

SAN VICENTE, Félix (dir.): *Contrastiva. Portal de gramática y de lingüística contrastiva español-italiano* <http://www.contrastiva.it/wp/>.

SECO, M. (1987): *Estudios de lexicografía española*, Madrid: Paraninfo.

TORRUELLA CASAÑAS, Joan; CAPSADA BLANCH, Ramon (2013): "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness", *Procedia - Social and Behavioral Sciences*, 95, pp. 447-454.

___ (2017): "Métodos para medir la riqueza léxica de los textos. Revisión y propuesta", *Verba*, 44, pp. 347-408.

VIVANCO CERVERO, Verónica (2003): *Homonimia y polisemia, teoría semántica y aplicación lexicográfica*, Buenos Aires: Ediciones del Sur.