Mark Davies

Uso del *Corpus del Español* y los corpus relacionados para la lexicografía histórica española

Resumen: El *Corpus del Español* original de 100 millones de palabras (siglos XIII–XX) permite a los investigadores llevar a cabo investigaciones avanzadas que incluyen colocativos, n-gramas, sinónimos, así como comparaciones a través de períodos históricos, que no son posibles con otros corpus como *CORDE*. Además, los datos del siglo XX permiten comprender la variación en el léxico basada en el género textual, lo que no es posible con otros corpus como *CREA*. En 2016, el *Corpus del Español* se amplió considerablemente para incluir dos mil millones de palabras de 20 países de habla hispana, lo que permite que la investigación sobre la variación regional del léxico sea mucho más profunda que con otros corpus como *CORPES*. Finalmente, el próximo corpus *NOW-Spanish* (que se lanzará a finales de 2018), con seis mil millones de palabras en español, es completamente original en la forma en que los investigadores pueden rastrear los cambios léxicos en curso (incluso semana por semana).

Palabras clave: Corpus Léxico, Histórico, Géneros, Dialecto

Abstract: The original 100 million word *Corpus del Español* (1200s–1900s) allows researchers to carry out advanced research involving collocates, n-grams, synonyms —as well as comparisons across historical periods— in ways that are not possible with other corpora such as *CORDE*. In addition, the data from the 1900s allows insight into genre-based variation in lexis, which is not possible with other corpora like *CREA*. In 2016 the *Corpus del Español* was greatly enlarged to include two billion words of data from 20 Spanish-speaking countries, and it allows research on regional variation in lexis that is much more powerful than with other corpora like *CORPES*. Finally, the upcoming *NOW-Spanish* corpus (to be released in late 2018) is completely unique in the way that researchers can track ongoing changes in lexis (even week-by-week) in a six billion-word corpus of Spanish.

Keywords: Corpus, Lexis, Historical, Genres, Dialect

1 Introducción

Hasta hace unos veinte años, la lexicografía histórica solía llevarse a cabo mediante minuciosos y laboriosos análisis de textos impresos. Sin embargo, a partir de la llegada de grandes corpus en línea, la tarea de quienes estudian el

cambio y la variación léxica se ha convertido en algo mucho más fácil. Los investigadores pueden lograr ahora en solo unos minutos lo que antes podría haber necesitado días o incluso semanas.

En este trabajo, consideraremos con cierto detalle cómo el *Corpus del Español* se puede utilizar para examinar la variación y el cambio léxico[1]. Nos centraremos en las dos partes del corpus: el corpus de 100 millones de palabras «histórico/basado en el género» que se lanzó en 2002, y el corpus de dos mil millones de palabras «basado en web/dialectal» que se lanzó en 2016. También trataremos brevemente un corpus de cinco mil millones de palabras que se lanzará en 2018, y que permitirá a los investigadores examinar los neologismos y el cambio léxico actual con increíble detalle. A medida que analicemos estos diferentes corpus, también trataremos brevemente los corpus similares de la Real Academia Española: *CORDE* (textos históricos), *CREA* (principalmente textos de finales del siglo XX) y *CORPES* (textos de principios del siglo XXI).

2 Los corpus textuales

El *Corpus del Español* se compone de dos partes diferentes, ambas diseñadas para dar respuesta a diferentes preguntas sobre el cambio y la variación del lenguaje. La primera parte es el corpus «histórico/género». Se completó en 2002 y fue revisado a fines de 2007. Está compuesto por aproximadamente 100 millones de palabras, desde el español antiguo hasta finales del siglo XX, tiene aproximadamente 18 millones de palabras del siglo XIII al XV, 42 millones de palabras del siglo XVI al XVIII, y alrededor de 40 millones de palabras de los siglos XIX y XX. Está compuesto de textos de una amplia gama de géneros, incluidos más de cinco millones de palabras de transcripciones de conversaciones habladas de finales del siglo XX. Para el siglo XX, están divididos en partes iguales entre los siguientes géneros: hablado, ficción, prensa y académico. Los detalles completos sobre cada uno de los casi 14 000 textos se pueden encontrar a través del enlace «Textos» en el sitio web del corpus, y los usuarios pueden descargar un archivo de Excel que enumera todos esos textos.

La segunda parte del *Corpus del Español* fue lanzada en 2016. Está orientada a estudios sobre la variación dialectal en español o que necesitan un corpus mucho más grande que el corpus original de 100 millones de palabras. El corpus contiene más de dos mil millones de palabras, lo que lo hace 100 veces más grande que la porción del siglo XX del corpus original. Además, los usuarios pueden

1 Para estudios similares sobre el inglés, vid. Davies en prensa a y b.

realizar comparaciones entre los 21 países del corpus (España, México, Colombia, Argentina, etc.) para ver la frecuencia por dialecto y buscar directamente todas las palabras que son más frecuentes en un país que en otro.

En este artículo, compararemos la parte histórica (y basada en el género) del *Corpus del Español* con el *CORDE* (*Corpus diacrónico del español*) y el *CREA* (*Corpus de referencia del español actual*) de la Real Academia Española. A continuación, en las Secciones 3–11, compararemos la funcionalidad de la parte histórica del *Corpus del Español* con el *CORDE*, por lo que daremos una visión general del *CORDE*.

El *CORDE* se creó a finales de los años 90 del siglo XX y fue el primer gran corpus histórico del español. Está compuesto por aproximadamente 250 millones de palabras de texto, con buena representación a lo largo de los diferentes períodos históricos, y un buen equilibrio entre géneros, que incluyen poesía, escritos históricos, literatura, materiales didácticos, etc. Con 250 millones de palabras, es de dos a tres veces más grande que el *Corpus del Español*. Sin embargo, como veremos, el tamaño no es todo. Sin el tipo correcto de arquitectura e interfaz de corpus, los datos textuales están, en esencia, «atrapados» dentro del corpus y no están disponibles para los usuarios finales.

3 Descripción general del uso de corpus para la lexicografía histórica

Los corpus históricos, para ser útiles, deberían permitir a los usuarios realizar investigaciones sobre varios aspectos diferentes del cambio léxico[2]. Estos incluyen lo siguiente:

- Encuentra resultados representativos. En el nivel más básico, los usuarios pueden buscar una palabra o frase, encontrar la primera aparición de la palabra o frase y ver todas las ocurrencias en contexto.
- Frecuencia (más avanzada). Los usuarios pueden ver fácilmente la frecuencia de una palabra o frase a lo largo del tiempo, con frecuencias normalizadas[3].
- Frecuencia (aún más avanzada). En lugar de tener que decirle al corpus qué palabras o frases específicas buscar, el corpus puede generar una lista de palabras cuya frecuencia coincide con ciertos criterios, como los nombres que

2 Para estudios similares con corpus históricos del inglés, vid. Davies 2012a y 2012b.

3 En otras palabras, frecuencia por millar o por millón de palabras de texto, para tener en cuenta el diferente tamaño del corpus en diferentes períodos históricos.

se incorporaron al léxico en el siglo XVII o todas las palabras que se usan al menos cinco veces más en el siglo XIII que en el XIV

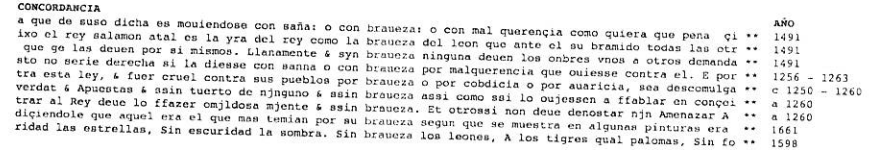
- Forma de palabra (morfológica). Los usuarios deben poder buscar por prefijos, sufijos y raíces, y ver la frecuencia de cada forma coincidente en los diferentes períodos históricos, así como la frecuencia general de todas las formas en cada período histórico.
- Significado de la palabra (semántica): simple. Los usuarios pueden encontrar las colocaciones más frecuentes (palabras cercanas) de una palabra o frase determinada, lo que obviamente proporciona una muy buena comprensión del significado de la palabra. Prácticamente cualquier arquitectura e interfaz de corpus permite a los usuarios ver las palabras cercanas caso por caso, pero los corpus realmente útiles resumen toda esta información sobre las colocaciones para todas las ocurrencias de una palabra o frase determinada.
- Semántico (más avanzado). Suponiendo que el corpus pueda encontrar colocaciones, debería ser posible compararlas a través de períodos históricos o entre diferentes géneros. Los cambios en las colocaciones a lo largo de períodos sirven a menudo como marcadores de cambio semántico.
- Semántica (aún más avanzada). En lugar de simplemente buscar palabras y frases, los usuarios pueden buscar por campo semántico. P. ej., si un repertorio está integrado en el corpus, o si los usuarios pueden crear listas personalizadas de palabras, entonces podrían crear una búsqueda donde cualquier palabra en un campo semántico es parte de la consulta. Un ejemplo de esto podría ser [miembro de la familia] seguido de [sinónimo de *pedir*] seguido de [sinónimo de *limpiar*], o [hora del día] cerca de [sinónimo de *lúgubre*]. Del mismo modo, se podría comparar la frecuencia de todas las palabras o frases en un campo semántico completo, y comparar la frecuencia y la distribución de cada miembro a lo largo del tiempo.

En los siguientes apartados, proporcionaré ejemplos concretos de cómo estos dos grandes corpus históricos, el *Corpus del Español* y el *CORDE*, pueden usarse (o no) para investigar la amplia gama de fenómenos enumerados anteriormente. (Para obtener una visión general previa de la funcionalidad del *Corpus del Español* con arquitectura e interfaz más antiguas, v. Davies 2002, 2005a, 2005b, 2008, 2010). Algunos lectores pueden, así, comenzar a obtener una perspectiva completamente nueva de lo que se puede hacer con los corpus históricos. Si han utilizado corpus con arquitecturas e interfaces limitadas, quizá los hayan usado solo para encontrar las ocurrencias de una palabra o frase específica. Sin embargo, una vez que una persona ha utilizado un corpus de los que permiten una amplia gama de consultas, se da cuenta de que hay inmumerables temas de lingüística histórica que podrían estudiarse con un corpus completo.

4 Encontrar y mostrar resultados representativos

En el nivel más básico, un corpus debería permitir al investigador buscar una palabra o frase, encontrar la primera ocurrencia de la palabra o frase o ver todas los resultados en contexto (y quizás limitar las ocurrencias a un periodo histórico determinado). Los programas para realizar tales búsquedas son abundantes y bastante rápidos: 1–2 segundos (en el caso de la búsqueda simple) incluso para un corpus de 100 millones de palabras.

El *CORDE* puede hacer estas búsquedas básicas bastante bien. P. ej., supongamos que el usuario quiere encontrar todas las apariciones de la palabra *braueza*. Después de enviar la búsqueda, el usuario ve que hay 273 ocurrencias en 86 documentos. Al hacer clic en «Obtención de ejemplos», el usuario ve entradas de «Palabra clave en contexto» (KWIC, *Keyword in Context*) como las siguientes:



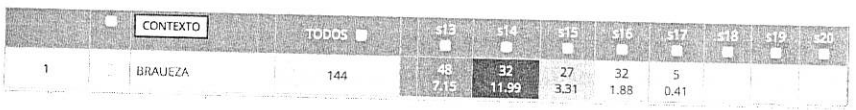
CONCORDANCIA

		AÑO
a que de suso dicha es mouiendose con saña: o con braueza: o con mal querençia como quiera que pena çi	**	1491
ixo el rey salamon atal es la yra del rey como la braueza del leon que ante el su bramido todas las otr	**	1491
que se las deuen por si mismos. Llanamente & syn braueza ninguna deuen los onbres vnos a otros demanda	**	1491
sto no serie derecha si la diesse con saña o con braueza por malquerencia que ouiesse contra el. E por	**	1256 - 1263
tra esta ley, & fuer cruel contra sus pueblos por braueza o por cobdicia o por auaricia, sea descomulga	**	c 1250 - 1260
verdat & Apuestas & asin tuerto de njnguno & asin braueza assi como asi lo ouiessen a ffablar en conçei	**	a 1260
trar al Rey deue lo ffazer omjldosa mjente & asin braueza. Et otrossi non deue denostar njn Amenazar A	**	a 1260
diçiendole que aquel era el que mas temian por su braueza segun que se muestra en algunas pinturas era	**	1661
ridad las estrellas, Sin escuridad la sombra. Sin braueza los leones, A los tigres qual palomas, Sin fo	**	1598

Gráfico 1: Palabra clave en la visualización de contexto con el *CORDE*

Sin embargo, una restricción importante a los usuarios del *CORDE* es que limita la visión de la palabra en contexto solo cuando aparece menos de mil veces en el corpus. De esta manera, cuando se encuentran miles de palabras no hay una manera fácil de verlas en contexto.

Una búsqueda básica de palabras y frases simples funciona de manera similar con el *Corpus del Español*. P. ej., después de hacer la búsqueda, el usuario verá lo siguiente:



	CONTEXTO	TODOS	≤13	≤14	≤15	≤16	≤17	≤18	≤19	≤20
1	BRAUEZA	144	46 7.15	32 11.99	27 3.31	32 1.88	5 0.41			

Gráfico 2: Listado de frecuencia con el *Corpus del Español*

Se muestra aquí la frecuencia absoluta de la palabra en cada siglo (p. ej., 32 ocurrencias en el siglo XIV), así como la frecuencia normalizada por siglo (p. ej.,

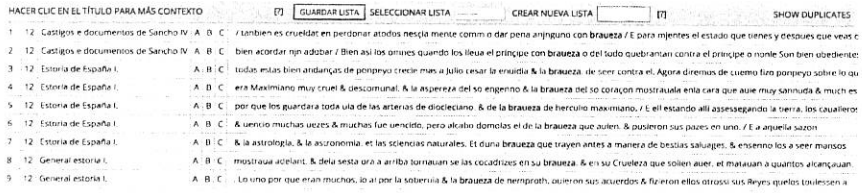


Gráfico 3: Pantalla de palabras clave en contexto con el *Corpus del Español*

un pico de alrededor de 12 ocurrencias por millón de palabras en el siglo XIV). Además, uno puede ver la palabra clave en contexto para cualquier palabra, no solo aquellas con baja frecuencia (como es el caso del *CORDE*).

En el *Corpus del Español*, al hacer clic en los números en cualquier columna se mostrará la palabra clave en el contexto de ese siglo, o se pueden ver todas las entradas a la vez haciendo clic en TOTAL. Luego se puede hacer clic en otras entradas para ver el contexto extendido (aproximadamente 200 palabras en total).

La visualización de palabra clave en contexto en el *Corpus del Español* permite algunas funcionalidades importantes que no son posibles con el *CORDE*. Primero, los usuarios pueden guardar ocurrencias (en tres grupos diferentes de ocurrencias) haciendo clic en A, B o C. Posteriormente, pueden organizar estas listas de ocurrencias (lo que incluye moverlas entre diferentes listas), lo que facilita en gran medida el almacenamiento y análisis de datos.

Hasta este punto, entonces, las búsquedas en los dos corpus son bastante similares. El *CORDE* tiene la ventaja de ser el corpus más grande, mientras que el *Corpus del Español* tiene la ventaja de mostrar la frecuencia en cada siglo y de mostrar la palabra clave en contexto para todas las palabras, independientemente de la frecuencia.

5 Frecuencia de palabra: datos básicos

Sin embargo, además de obtener simplemente todas las apariciones de una palabra o frase determinada, los usuarios a menudo quieren saber cuán frecuente eran en diferentes siglos o en determinados períodos históricos. Es en este punto donde el *CORDE* comienza a mostrar algunas debilidades serias. P. ej., después de buscar *braueza* y luego seleccionar «Ver estadística», el usuario visualiza:

Estadísticas

Año	%	Casos
1627	20.95	35
1547	20.35	34
1610	17.96	30
1632	8.98	15
1566	4.79	8

País	%	Casos
ESPAÑA	86.44	236
MÉXICO	10.98	30
PERÚ	1.46	4
BOLIVIA	0.36	1
COLOMBIA	0.36	1

Tema	%	Casos
22.- Verso narrativo	26.37	72
12.- Prosa narrativa	20.51	56
19.- Prosa histórica	17.58	48
21.- Verso lírico	12.45	34
16.- Prosa de sociedad	7.69	21

Gráfico 4: Frecuencia (por año) con el *CORDE*

Esta tabla nos dice los años específicos en los cuales la palabra o frase es más común, pero es imposible ver la frecuencia por década o por siglo. No sirve de nada mostrar que la palabra fue la más frecuente en 1627 si, de hecho, es mucho menos común en el siglo XVII que en el XIII o el XIV. El otro problema importante es que las cifras no están relativizadas. En otras palabras, vemos la frecuencia absoluta por año, pero una palabra o frase puede ser más común en ese año simplemente porque hay más palabras para ese año en el corpus. Cualquier comparación seria de frecuencia requiere que los resultados se relativicen a lo largo de períodos históricos para que podamos tener en cuenta los diferentes tamaños del corpus en diferentes periodos históricos, y ver la frecuencia de la palabra o frase por millón de palabras.

El *Corpus del Español* permite este tipo de búsqueda con bastante facilidad. P. ej., con *braueza* en el *Corpus del Español*, podemos ver la «visualización de tabla» (como en la tabla 2 anterior), o una visualización en gráfico:

SECCIÓN	s13	s14	s15	s16	s17	s18	s19	s20
OCURRENCIAS	48	32	27	32	5	0	0	0
POR MILL	7.15	11.99	3.31	1.88	0.40	0.00	0.00	0.00

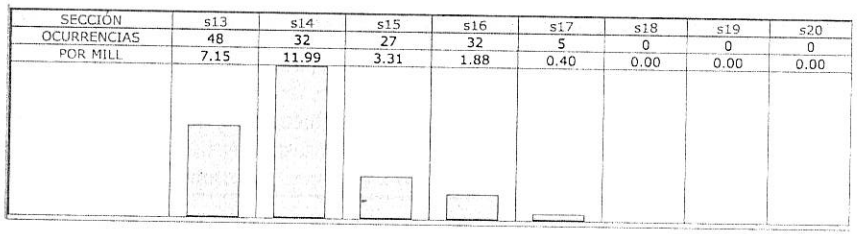


Gráfico 5: *Corpus del Español*: Frecuencia de palabra y frase por siglo

Se nos muestra aquí la frecuencia absoluta (p. ej., 32 ocurrencias en siglo XIV), así como la importante frecuencia relativa («por millón»), que tiene en cuenta el tamaño de la sección en millones de palabras. P. ej., hay 32 ocurrencias en los 3 millones de palabras del siglo XIV, o 10.8 ocurrencias por millón de palabras. Un gráfico como este es la única manera de ver realmente los cambios en la frecuencia de una palabra, frase o construcción, y solo es posible con el *Corpus del Español*.

6 Frecuencia de palabra: comparación de períodos históricos

En lugar de tener que decirle al corpus qué palabras o frases específicas buscar, un corpus con arquitectura e interfaz bien diseñadas generaría una lista de palabras cuya frecuencia coincide con ciertos criterios. P. ej., podría encontrar todos los sustantivos que entraron en la lengua del siglo XVII o las palabras que se usaron al menos cinco veces más en el siglo XIII que en el XIV. Tal consulta es completamente imposible con el *CORDE*. Todo lo que se puede hacer es buscar palabras y frases específicas. Si el *CORDE* reconoce la frecuencia de todas las palabras y frases en todos los períodos históricos, ciertamente no permite a los investigadores usar esa información mediante una consulta.

Con el *Corpus of Spanish*, por otro lado, tales consultas son bastante sencillas. P. ej., se puede buscar simplemente [nn *] (sustantivos) y seleccionar [siglo XIII] (1200–1299) «SECCIÓN 1» para comparar con [siglo XIV] (1300–1399) «SECCIÓN 2». En uno o dos segundos, el usuario ve la siguiente lista[4]:

Tabla 1: *Corpus del Español*. Comparación de la frecuencia de palabras por siglo (todas las palabras a la vez)

Siglo XIII	Siglo XIV
capitolo, ascendente, ladeza, saturnus, orizon, morauedis, roque, xaque, acendent, armella, murcia, baldouin, ascendent, segonda, gudufre, significador, fferrando, corualan, alffil, zonte, dond, iudga, tiemplo, boymonte, catamiento, infortunas, uinie, caput, camyaron, sacrifficio, declinacion, sacrifficios, yguador, juppiter, algarue, linnas, deillos, hierusalem, decima.	armadas, osso, ome, avia, paris, ahe, collado, camjno, hercoles, ley, çima, rrayzes, armada, elena, falcon, yuierno, verano, gonçales, encarnaçion, pase, bjen, venja, avras, falcones, jnfante, façer, puerco, ynfanta, ynfante, ynperio, vyno, venjdo, sembrar, fojas,ençima, talante, mjel, menalao, syenpre, dolençia, ssiete, avedes, castilla, muria, aujdo, peça, arroyo, vuas, çiençia, termjno, tenjan.

4 Téngase en cuenta que en la versión web hay frecuencias –en bruto y normalizadas– para cada palabra, así como enlaces para ver la palabra en contexto, como se muestra en el Gráfico 2. En la Tabla 1, hemos simplificado la visualización.

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	SECTOR	2540	0	111.3	0.0	11,129.5
2	TELEVISIÓN	2119	0	92.8	0.0	9,284.8
3	SECTORES	1360	0	59.6	0.0	5,959.1
4	FÚTBOL	1208	0	52.9	0.0	5,293.1
5	LÍDER	1119	0	49.0	0.0	4,903.1
6	PROTEÍNAS	753	0	33.0	0.0	3,299.4
7	IMPACTO	746	0	32.7	0.0	3,268.7
8	AEROPUERTO	721	0	31.6	0.0	3,159.2
9	NARCOTRÁFICO	705	0	30.9	0.0	3,089.1
10	INFLACIÓN	704	0	30.8	0.0	3,084.7
11	LÍDERES	666	0	29.2	0.0	2,918.2
12	CAMPEONATO	657	0	28.8	0.0	2,878.8

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	VENTURA	1237	25	64.1	1.1	58.5
2	NÚM	1025	21	53.1	0.9	57.7
3	RODRIGO	697	19	36.1	0.8	43.4
4	VULGO	662	20	34.3	0.9	39.1
5	PAJE	611	21	31.7	0.9	34.4
6	MATILDE	669	24	34.7	1.1	33.0
7	APOSENTO	1174	44	60.8	1.9	31.6
8	HONRA	1652	63	85.6	2.8	31.0
9	MENESTER	1545	60	80.1	2.6	30.5
10	MARGARITA	554	23	28.7	1.0	28.5
11	ERMITAÑO	553	23	28.7	1.0	28.4
12	CONDESA	1829	78	94.8	3.4	27.7

Gráfico 6: *Corpus del Español*: Comparación de la frecuencia de las palabras (sustantivos en siglo XIX/siglo XX)

Obviamente, solo algunas de las palabras de esta lista son significativas. Muchas palabras son simplemente variantes ortográficas, otras son sustantivos propios que pueden aparecer en un puñado de textos de un siglo, pero no de otro.

Por supuesto, tales búsquedas no están limitadas solo al español antiguo, también se pueden llevar a cabo para períodos históricos más recientes. La siguiente tabla muestra (a la izquierda) nombres que son comunes en el siglo XX, pero no en el XIX y (a la derecha) aquellos que son comunes en el siglo XIX pero no en el siglo XX. Esta tabla muestra todas las palabras que son más comunes en un período que en otro, incluso si no aparecen en el segundo período. P. ej., no hay resultados de televisión o aeropuerto en el siglo XIX (lo que probablemente no sea sorprendente).

Para «comparar manzanas con manzanas», podría ser útil indicar que una palabra debe darse con al menos una frecuencia determinada en cada uno de los dos períodos. P. ej., la tabla 7 muestra los nombres que aparecen al menos diez veces en los siglos XIX y XX, pero en los que hay un aumento significativo a lo largo del tiempo[5].

En resumen, debido a la arquitectura del *Corpus del Español*, donde el corpus «reconoce» la frecuencia de cada palabra y frase en cada período histórico, tales comparaciones son bastante simples. En cambio, con el *CORDE* el corpus aparentemente no reconoce la frecuencia de palabras y frases en cada sección

5 Téngase en cuenta que, debido a que el etiquetado no es siempre perfecto, hay algunas entradas extrañas, como *lic* en el siglo XX o el nombre propio de *Matilde* en el siglo XIX, pero la mayoría de las palabras son relevantes.

SEC 1 (1900s): 22,822,256 PALABRAS							SEC 2 (1800s): 19,297,249 PALABRAS						
	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	DÉCADA	3647	16	159.8	0.8	192.7	VENTURA	1237	25	64.1	1.1	58.5
2	DÓLARES	3081	19	135.0	1.0	137.1	NÚM	1025	21	53.1	0.9	57.7
3	FINALES	2644	19	115.9	1.0	117.7	RODRIGO	697	19	36.1	0.8	43.4
4	CÉLULAS	1693	16	74.2	0.8	89.5	VULGO	662	20	34.3	0.9	39.1
5	POSIBILIDADES	1350	15	59.2	0.8	76.1	PAJE	611	21	31.7	0.9	34.4
6	DÉCADAS	1112	14	48.7	0.7	67.2	MATILDE	669	24	34.7	1.1	33.0
7	CINE	1613	21	70.7	1.1	64.9	APOSENTO	1174	44	60.8	1.9	31.6
8	UTILIZACIÓN	758	10	33.2	0.5	64.1	HONRA	1652	63	85.6	2.8	31.0
9	LIC	2004	27	87.8	1.4	62.8	MENESTER	1545	60	80.1	2.6	30.5
10	DEPORTE	712	10	31.2	0.5	60.2	MARGARITA	554	23	28.7	1.0	28.5
11	EQUIPO	3185	46	139.6	2.4	58.5	ERMITAÑO	553	23	28.7	1.0	28.4
12	OBJETIVOS	809	13	35.4	0.7	52.6	CONDESA	1829	78	94.8	3.4	27.7

Gráfico 7: *Corpus del Español*: Comparación de la frecuencia de las palabras (sustantivos en siglo XIX/siglo XX)

(hasta que se buscan, una palabra específica cada vez), por lo que tal listado sería completamente imposible.

7 Búsqueda de formas de palabras (morfológicas)

Idealmente, los usuarios deberían poder ir más allá de las palabras y frases exactas y buscar por prefijos, sufijos y raíces. Esto les permitiría ver la frecuencia de cada forma coincidente en los diferentes períodos históricos, así como también la frecuencia general de todas las formas en cada período histórico.

El *CORDE* tiene serios problemas en búsquedas orientadas morfológicamente porque el motor de búsqueda no fue diseñado para ser utilizado en investigaciones de orientación gramatical. En el mejor de los casos, el corpus produce resultados, aunque no son demasiado útiles. P. ej., supongamos que un usuario busca *des*m?ento* en los siglos del XIII al XV. El corpus indica que hay «797 casos en 81 documentos». Luego, puede recorrer todas las 797 ocurrencias, una por una, y contar manualmente el total de cada forma diferente (*desfazimiento*, *desaffiamiento*, etc.) y ver con qué frecuencia aparece cada una, pero esto llevaría una o dos horas. Se podría seleccionar «Recuperar/agrupaciones» para ver las cadenas de dos, tres y cinco palabras más frecuentes (*destroymiento de*, *destruimiento de*, etc.), y se podría tardar una hora, aproximadamente, en encontrar las palabras más frecuentes que coincidan con este patrón. Estas búsquedas, además, solo funcionan cuando el número total de ocurrencias para una forma determinada aparece mil veces o menos en el corpus. Para una búsqueda como **azo* (*puñetazo*, *portazo*, etc.), el corpus simplemente afirma que «No se pueden ver estadísticas. Demasiados documentos». Una vez más, esto se debe a que su motor de

búsqueda fue diseñado para permitir a los usuarios encontrar y leer documentos completos (como con Google) y (en este caso, al menos) no es adecuado para la investigación gramatical.

Con el *Corpus del Español*, sin embargo, las búsquedas orientadas morfológicamente son fáciles y rápidas. P. ej., un usuario que desee encontrar las formas más frecuentes para *des*m?ento* entre los siglos XIII–XVII, aproximadamente en un segundo, verá lo siguiente:

		CONTEXTO	TODOS	s13	s14	s15	s16	s17
1		DESCUBRIMIENTO	864			13	515	336
2		DESABRIMIENTO	160			2	80	78
3		DESTRUYMIENTO	121	89	17	15		
4		DESTROYMIENTO	97	94	2	1		
5		DESTRUYMJENTO	87	14	45	28		
6		DESVANECIMIENTO	77			2	15	60
7		DESFALLESÇIMJENTO	74		3	71		
8		DESTERRAMIENTO	65	53	10	2		
9		DESAGRADECIMIENTO	57			3	40	14
10		DESCOMEDIMIENTO	46				33	13
11		DESCOMULGAMIENTO	42	41		1		
12		DESTERRAMJENTO	42	4	14	24		

Gráfico 8: *Corpus del Español*: frecuencia de formas verbales (*des*m?ento* en los siglos XIII–XVII)

La interfaz muestra la frecuencia de cada forma en cada siglo (p. ej., 53 casos de *desterramiento* en el XIII), así como en el total de los siglos seleccionados (en este caso, del XIII al XV) en la columna de la derecha. Los usuarios pueden seleccionar las formas y los siglos que sean de interés, y luego hacer clic para ver las palabras en contexto.

Además de recuperar las frecuencias individuales para todas las formas coincidentes (como en el Gráfico 8), también es posible ver el total agregado para todas las formas coincidentes en cada siglo, como hemos indicado en el Gráfico 5. Finalmente, como se describió en la sección 6, también podemos comparar la frecuencia de las formas en diferentes secciones del corpus. P. ej., supongamos que un usuario del *Corpus del Español* quiere ver qué palabras que terminan en **ismo* son más comunes en el siglo XX (izquierda) y el siglo XIX (derecha), respectivamente. En menos de un segundo, vería lo siguiente:

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	PROTAGONISMO	175	0	7.7	0.0	766.8
2	METABOLISMO	148	0	6.5	0.0	648.5
3	URBANISMO	127	0	5.6	0.0	556.5
4	SURREALISMO	118	0	5.2	0.0	517.0
5	ATLETISMO	111	0	4.9	0.0	486.4
6	MARXISMO	108	0	4.7	0.0	473.2
7	FASCISMO	108	0	4.7	0.0	473.2
8	CUBISMO	105	0	4.6	0.0	460.1
9	HINDUISMO	104	0	4.6	0.0	455.7
10	NERVIOSISMO	97	0	4.3	0.0	425.0
11	EXPRESIONISMO	79	0	3.5	0.0	346.2

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	PAUPERISMO	71	0	3.7	0.0	367.9
2	CACIQUISMO	123	14	6.4	0.6	10.4
3	DESPOTISMO	293	38	15.2	1.7	9.1
4	ABISMO	1030	186	53.4	8.1	6.5
5	PAGANISMO	88	16	4.6	0.7	6.5
6	FANATISMO	281	65	14.6	2.8	5.1
7	PATRIOTISMO	330	78	17.1	3.4	5.0
8	FATALISMO	50	19	2.6	0.8	3.1
9	SENTIMENTALISMO	57	22	3.0	1.0	3.1
10	CATOLICISMO	304	138	15.8	6.0	2.6
11	CATECISMO	147	68	7.6	3.0	2.6

Gráfico 9: *Corpus del Español*: Comparación de formas (**ismo* en siglos XIX/XX)

Esto muestra, p. ej., que *despotismo* se encuentra 293 veces en el siglo XIX, pero solo 38 veces en el XX, y que *fascismo* se encuentra 108 veces el siglo XX, pero ninguna en el XIX. La capacidad de comparar formas de palabras en diferentes siglos es una característica de gran alcance del *Corpus del Español*, pero no es posible con el *CORDE*.

8 Significado de la palabra (semántica): colocaciones básicas

Como les gusta señalar a los lingüistas de corpus, «se puede decir mucho sobre una palabra con las otras palabras con las que se junta» (Firth 1957). A veces, las colocaciones (palabras cercanas) simplemente confirman lo que ya sabemos. P. ej., las colocaciones nominales más comunes (palabras cercanas) para *selva* son *árboles*, *vegetación*, *sierra*, *bosque*, etc. Para una palabra menos concreta, a menudo es necesaria más perspicacia. P. ej., los sustantivos más comunes que aparecen con formas de *lúgubre* son *acento*, *voz*, *silencio*, *noche*, *eco*, *gemido*, etc.

La clave del significado, entonces, se suele encontrar en las colocaciones o en las palabras cercanas. Prácticamente cualquier arquitectura e interfaz de corpus permite a los usuarios buscar una palabra y luego ver esa palabra en contexto. El usuario del corpus siempre puede recorrer los ejemplos uno por uno, tomar notas sobre palabras cercanas comunes y luego tratar de usar estas colocaciones para discernir el significado. Sin embargo, esto puede consumir mucho tiempo para palabras comunes. Un enfoque más productivo consistiría en hacer que el corpus encuentre todas las colocaciones por sí mismo y luego presentarlas al usuario en orden de frecuencia.

En términos de cambio histórico, lo deseable sería poder encontrar las colocaciones de una palabra dada en diferentes períodos históricos. Al observar los

cambios en las colocaciones a lo largo del tiempo, podemos obtener información sobre los cambios en el significado y el uso de la palabra.

Consideremos brevemente cómo el *CORDE* y el *Corpus del Español* permiten a los usuarios encontrar y procesar las colocaciones para obtener una idea del significado de las palabras. En el caso del *CORDE*, supongamos que queremos examinar las casi 38 000 colocaciones de todas las formas de *duro* (*dura*, *duros*, etc.). Suponiendo que a un usuario le toma unos 20 segundos encontrar cada ocurrencia en contexto y anotar (lo que supone que son) las palabras cercanas relevantes, dedicaría alrededor de 26 días (a ocho horas diarias) a repasar todos los ejemplos relevantes; y esto, suponiendo que el usuario no decidiera cambiar el ancho de la «ventana de colocaciones» o buscase un tipo diferente de colocación, en cuyo caso tendría que emplear, más o menos, otro mes.

El *CORDE* permite a los usuarios ver «agrupaciones» para una palabra determinada, como las de la forma simple *duro* en el siglo XIII:

De 2 palabras	%	Casos
duro la	6.97	30
duro el	6.27	27
duro en	5.34	23
duro fasta	3.95	17
duro esta	3.48	15

De 3 palabras	%	Casos
duro en el	2.55	11
duro la batalla	1.86	8
duro punto nado	1.16	5
duro esta batalla	1.16	5
duro fasta el	1.16	5

Gráfico 10: *CORDE*: *duro* + colocaciones

Pero tal listado es de poco valor, porque como el *CORDE* no tiene ninguna forma de discriminar qué palabras son relevantes, nos da colocaciones como *dura la*, *dura en*, etc. (ya que *la*, *en*, etc., concurren con frecuencia con casi cualquier término), que proporcionan poca o ninguna comprensión del significado de la palabra. En cualquier caso, solo lista las nueve colocaciones más frecuentes, lo que no es suficiente para obtener información completa sobre el significado. Estas búsquedas son mucho más fáciles con el *Corpus del Español*. Los usuarios simplemente introducen la «palabra del nodo» (p. ej., *duro*, o *lúgubre*, o *selva*) y pueden, opcionalmente, seleccionar la categoría gramatical de las colocaciones;

en aproximadamente dos o tres segundos tienen todas las colocaciones en orden. P. ej., supongamos que un usuario quiere encontrar colocaciones relacionadas con el concepto '*duro*' en español antiguo. Después de introducir [= duro] (*duro, duras*, etc.,) y esperar unos dos segundos, aparece una lista como la siguiente:

		CONTEXTO	TODOS	s13	s14	s15	s16	s17
1		PIEDRA	188	33	7	52	56	40
2		QUEBRANTAR	91	89			2	
3		MIENTRAS	84			1	48	35
4		PIEDRAS	81	10	3	23	28	17
5		DURA	77	13		15	30	19
6		GOLPES	75		12	1	53	9
7		FUERT	71	71				
8		HIERRO	53			3	36	14
9		PEÑA	48			2	21	25
10		MÁRMOL	47				26	21
11		HADO	42				40	2
12		PEÑAS	42		1	1	24	16

Gráfico 11: *Corpus del Español*: colocaciones de *duro* en siglos XIII–XVIII

Esta tabla muestra la frecuencia de cada colocación en cada siglo (aquí solo se muestran los siglos XIII al XVII). P. ej., *piedra* ocurre cerca de [*duro*] 33 veces en el siglo XIII y 7 veces más en el XV. Hay 188 apariciones totales de piedra entre los siglos XIII y XV, por lo que los 40 casos cerca de [*duro*] son aproximadamente el 1,8 % de todos los resultados. Esto se traduce en una puntuación de información mutua de 3.53, que muestra que la relación entre las dos palabras es significativa. Por lo tanto, con el *Corpus del Español*, podemos hacer en 2–3 segundos lo mismo que llevaría hacer con el *CORDE* un mes o más.

9 Cambio semántico: comparación de colocaciones en diferentes períodos históricos

Si tenemos una arquitectura e interfaz de corpus que nos permite encontrar fácilmente colocaciones (lo que es posible, como se ha visto, con el *Corpus del Español*), podemos utilizar esta información de manera ingeniosa para examinar el cambio semántico. La idea básica es que si las palabras «cercanas» a una

determinada palabra cambian con el tiempo, puede deberse a que la misma palabra ha cambiado de significado (o al menos se está utilizando de una manera diferente). P. ej., la siguiente tabla muestra (a la izquierda) los sustantivos que aparecen con [*duro*] en el siglo XX, pero que no son muy comunes en el XIX, y (a la derecha) en el XIX, pero no en el XX:

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	CRÍTICAS	27	0	1.2	0.0	118.3
2	MADERAS	19	0	0.8	0.0	83.3
3	LÍNEA	18	0	0.8	0.0	78.9
4	REPRESIÓN	16	0	0.7	0.0	70.1
5	DISCO	13	0	0.6	0.0	57.0
6	REGALO	10	0	0.4	0.0	43.8
7	LUCHA	10	1	0.4	0.1	8.5
8	COMPETENCIA	15	2	0.7	0.1	6.3
9	BATALLA	12	2	0.5	0.1	5.1
10	GOLPE	51	12	2.2	0.6	3.6
11	MESES	12	3	0.5	0.2	3.4
12	FORMA	12	3	0.5	0.2	3.4
13	GOBIERNO	13	4	0.6	0.2	2.7

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	MILES	30	0	1.6	0.0	155.5
2	RENTA	24	0	1.2	0.0	124.4
3	MILLONES	20	0	1.0	0.0	103.6
4	TRANCE	18	0	0.9	0.0	93.3
5	ART	18	0	0.9	0.0	93.3
6	CARÁCTER	17	0	0.9	0.0	88.1
7	ENTRAÑAS	14	0	0.7	0.0	72.5
8	DUROS	13	0	0.7	0.0	67.4
9	REALES	13	0	0.7	0.0	67.4
10	PESOS	12	0	0.6	0.0	62.2
11	PUÑADO	10	0	0.5	0.0	51.8
12	CANTIDAD	10	0	0.5	0.0	51.8
13	ALMA	10	0	0.5	0.0	51.8

Gráfico 12: *Corpus del Español*: comparación de las colocaciones de *duro*, siglos XIX/XX

P. ej., *disco* se produce cerca de *duro* 13 veces en el siglo XX, pero no hay resultados (lo que no sorprende) en el XIX. *Entrañas*, por otro lado, aparece 14 veces con [*duro*] en el XIX, pero no hay resultados en el siglo XX. Si *disco* está en el corpus en el siglo XIX y *entrañas* en el siglo XX, ¿por qué su frecuencia como colocación con *duro* cambia tanto de un siglo a otro? ¿Es porque el uso de *duro* puede haber cambiado en algo? Mostremos otro ejemplo: la siguiente es una lista parcial de las colocaciones adjetivales de *mujer* (y *mujeres*) en los siglos XX (izquierda) y XIX (derecha):

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	INTERNACIONAL	17	0	0.7	0.0	74.5
2	SEXUAL	15	0	0.7	0.0	65.7
3	CUBANAS	15	0	0.7	0.0	65.7
4	EMBARAZADAS	14	0	0.6	0.0	61.3
5	IMPORTANTE	13	0	0.6	0.0	57.0
6	PROFESIONALES	13	0	0.6	0.0	57.0
7	LABORAL	12	0	0.5	0.0	52.6
8	EX	12	0	0.5	0.0	52.6
9	ÁRABE	10	0	0.4	0.0	43.8
10	DERECHOS	10	0	0.4	0.0	43.8
11	DIFERENTES	15	1	0.7	0.1	12.7
12	MADURA	15	1	0.7	0.1	12.7
13	SEXUALES	12	1	0.5	0.1	10.1

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	HONRADA	58	0	3.0	0.0	300.6
2	HONRADAS	35	0	1.8	0.0	181.4
3	DIGNA	24	0	1.2	0.0	124.4
4	LIVIANA	21	0	1.1	0.0	108.8
5	INFAME	19	0	1.0	0.0	98.5
6	DESVENTURADA	19	0	1.0	0.0	98.5
7	PERDIDAS	18	0	0.9	0.0	93.3
8	VANO	18	0	0.9	0.0	93.3
9	MALDITA	16	0	0.8	0.0	82.9
10	SEMEJANTE	15	0	0.8	0.0	77.7
11	DESDICHADA	14	0	0.7	0.0	72.5
12	LIVIANAS	13	0	0.7	0.0	67.4
13	INFELICES	13	0	0.7	0.0	67.4

Gráfico 13: *Corpus del Español*: comparación de colocaciones de *mujer*, siglos XIX/XX

Obsérvese cómo los adjetivos del siglo XIX (derecha) se refieren a las «virtudes morales» de las mujeres, que están casi por completo ausentes en el siglo XX (izquierda); en este siglo, por otro lado, son mucho más prosaicos y se refieren a clasificaciones sobre la nacionalidad, el empleo, etc. En este caso, los datos del corpus proporcionan información interesante del cambio en la forma de ver a las mujeres en estos dos siglos; podemos obtener esta útil información con una simple búsqueda de 1–2 segundos en el corpus.

Aplicado al español antiguo o en los textos de los siglos XVI al XVIII, se podría adoptar un enfoque similar. Usando la interfaz para el *Corpus del Español,* simplemente es necesario indicar qué palabras o conceptos son de interés, especificar el tipo de colocación (sustantivo, verbo, etc., si corresponde), y luego hacer clic una o dos veces más para mostrar qué dos períodos históricos deben ser comparados. En dos o tres segundos, se recopilan y resumen todos los datos relevantes. Usando el *CORDE*, en cambio, las búsquedas como esta serían muy difíciles o imposibles, ya que la arquitectura del *CORDE* no está diseñada para encontrar colocaciones.

10 Cambios léxicos en un campo semántico: sinónimos

Con la arquitectura de corpus adecuada, los usuarios podrían buscar por campos semánticos, en lugar de simplemente buscar palabras y frases. Así, en el *Corpus del Español*, los índices onomasiológicos se integran en la arquitectura del corpus; esto nos permite, en el nivel más básico, encontrar la frecuencia histórica de todas las palabras relacionadas con un concepto en particular. P. ej., los usuarios pueden introducir = *mujer* y ver la frecuencia de todos los sinónimos de *mujer* a lo largo del tiempo (y, p. ej., en diferentes géneros en el siglo XX).

Esta lista parcial de resultados muestra, p. ej., que *doncella* y *moza* han disminuido desde el siglo XVI (por cada millón de palabras), mientras que *chica* y *muchacha* han aumentado desde el siglo XVIII hasta el XIX/XX. En cuanto a los períodos medievales, lo que obviamente se necesitaría es algún tipo de «índice histórico» que por ahora no está disponible. Pero, en la medida en que lo estuviera, la arquitectura del corpus podría incorporarlo fácilmente.

Además de buscar la frecuencia de palabras sueltas, la información semántica de los repertorios o las listas de palabras personalizadas y definidas por el usuario se pueden integrar directamente en la sintaxis de la consulta. P. ej., en el *Corpus del Español*, es posible que los usuarios creen (a través de la interfaz web) listas personalizadas de palabras de un campo de interés semántico particular, como términos navales, palabras relacionadas con las emociones, una lista de términos relacionados con la estructura familiar o una lista de palabras relacionadas con

	CONTEXTO	TODOS	s13	s14	s15	s16	s17	s18	s19	s20	ACAD	PER	FIC	ORAL
1	MUJER [S]	35395	19	6	687	5707	8233	1962	11836	6945	478	803	3740	1924
2	SEÑORA [S]	32812	161	955	1584	9385	8635	1704	7465	2923	43	314	1819	747
3	PERSONA [S]	24445	430	215	2637	5576	3430	2044	5368	4745	787	854	786	2318
4	JOVEN [S]	11639			4	238	465	607	7193	3132	362	729	1408	633
5	ESPOSA [S]	9400	121	36	219	1712	2809	365	2694	1444	196	480	572	196
6	DAMA [S]	8287	3	2	133	1412	3879	332	2020	506	46	134	266	60
7	MUCHACHA [S]	2266				115	135	132	994	890	13	49	663	165
8	DONCELLA [S]	3025	3	1		1003	762	224	967	65	8	11	42	4
9	CHICA [S]	1725	29	43	95	97	59	96	390	916	19	98	406	393
10	SEÑORITA [S]	1534			1	1	7	134	962	429	9	28	261	131
11	HEMBRA [S]	1397	1	3	261	311	144	97	241	339	207	22	89	21
12	MOZA [S]	1592	3		1	326	584	116	538	24	1	1	20	2

Gráfico 14: *Corpus del Español*: comparación de sinónimos de *mujer*

un concepto teológico particular. Esta lista personalizada de palabras se puede usar como parte de la sintaxis de la consulta. P. ej., si un usuario [andrés.gómez] crea una lista de 100 palabras relacionadas con 'emociones' en español antiguo y otra lista de 70 palabras relacionadas con 'relaciones familiares' (*padre, hermanastro, nuera*, etc.), podría encontrar cada vez que aparezca una palabra en la Lista 1 cerca de la Lista 2. De esta forma, se pueden llevar a cabo búsquedas semánticas de gran alcance en el corpus.

El *Corpus del Español* puede incorporar este tipo de consultas orientadas semánticamente debido a la arquitectura subyacente del corpus, que se basa en bases de datos relacionales. Con este tipo de bases de datos relacionales es posible agregar cualquier cantidad de conjuntos de datos nuevos (repertorios, listas de palabras definidas por el usuario, etc.) y luego integrarlos sin problemas en la sintaxis de la consulta. La arquitectura para el *CORDE*, en cambio, no es «abierta» y no admite la incorporación de otros conjuntos de datos. Solo se puede buscar palabras o frases individuales, pero nada que se aproxime a un campo semántico completo o algo similar.

11 Desvío sincrónico # 1: género

Hasta este punto, hemos discutido los cambios históricos en el léxico y el significado, pero apenas hemos hecho alusión al género y a cómo este afecta a la frecuencia y al significado de las palabras. Sin embargo, obviamente, hay que considerar la importancia del género. P. ej., analícense los siguientes tres cuadros, que muestran la frecuencia de tres palabras del español moderno que son más comunes en el género de ficción (*borracho*), académico (*proporcionar*) y el marcador de discurso *bueno* en español hablado (como en «lo haremos. Bueno, pero no es tan fácil»):

borracho proporcionar bueno

ACAD	PER	FIC	ORAL
1207	281	127	76
241.40	56.61	26.63	17.95

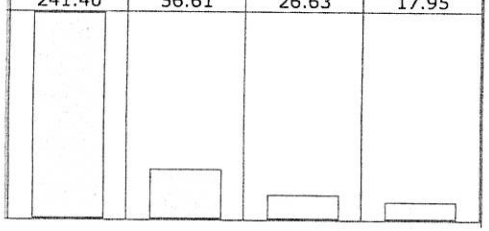


Gráfico 15: Frecuencia de palabras por género en español moderno

Estadísticas

Año	%	Casos
1996	10.08	124
2000	8.86	109
2001	8.78	108
1995	7.39	91
1997	7.07	87
2002	5.93	73

País	%	Casos
ESPAÑA	64.81	818
MÉXICO	14.18	179
ARGENTINA	8.00	101
CHILE	2.77	35
PERÚ	2.45	31
EE. UU.	1.66	21

Tema	%	Casos
6.- Salud.	22.59	289
1.- Ciencia y Tecnología.	20.95	268
2.- Ciencias sociales, creencias y pensamiento.	15.71	201
5.- Ocio, vida cotidiana.	11.64	149
4.- Artes.	10.47	134
3.- Política, economía, comercio y finanzas.	10.39	133

Gráfico 16: Frecuencia por género/dominio en *CREA*

Afortunadamente, el *Corpus del Español* nos permite encontrar rápida y fácilmente la frecuencia de cualquier palabra, frase, morfema o construcción sintáctica en los diferentes géneros del español moderno (siglo XX). Los corpus de la Real Academia Española no proporcionan cifras de frecuencia útiles por género. El *CREA* es el corpus «sincrónico» de la RAE y, de forma similar a lo que hemos visto en el Gráfico 4 para el corpus diacrónico *CORDE*, no proporciona frecuencias «normalizadas» por género. P. ej., muestra que el 22.59 % de las ocurrencias de *borracho* proviene de textos de 'salud' y que 11.64 (aproximadamente la mitad de Salud) proviene de textos que tratan de 'ocio, vida cotidiana'.

Pero si el sub-corpus para 'salud' tiene dos veces el tamaño del corpus 'ocio', entonces cualquier cosa ocurriría aproximadamente el doble en 'salud' y, por tanto, los porcentajes no tendrían sentido. Un corpus tiene que mostrar la frecuencia normalizada (p. ej., por millón de palabras) para que los resultados adquieran algún valor.

Al igual que con la parte histórica *del Corpus del Español* (v. Gráfico 8), también podemos encontrar la frecuencia por género en el *Corpus del Español* para

	CONTEXTO	TODOS	s13	s14	s15	s16	s17	s18	s19	s20	ACAD	PER	FIC	ORAL
1	[HERMOSO] [S]	22329	559	432	1249	6328	6059	1392	4919	1391	117	226	869	179
2	[BELLO] [S]	12404	39	160	261	1632	3375	1366	3844	1727	315	516	631	265
3	[VALIENTE] [S]	6714	62	156	448	2256	1818	353	1397	224	15	46	129	34
4	[LINDO] [S]	4742	61	30	293	857	1143	213	1125	1020	11	133	431	445
5	[BONITO] [S]	2549			10	93	58	63	731	1594	18	120	338	1118
6	[GALÁN] [S]	3887			30	713	2038	169	795	142	6	60	71	5
7	[ELEGANTE] [S]	2423		1	58	215	119	238	1169	628	126	92	322	83
8	[GALLARDO] [S]	2856		14	2	606	1173	238	704	119	8	83	22	6
9	[VALEROSO] [S]	2878			63	1465	774	181	349	46	10	8	27	1
10	[ATRACTIVO] [S]	1501				26	40	227	498	710	151	300	154	105

Gráfico 17: Frecuencia por sinónimo

SEC 1 (FICT): 4,769,873 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	[CANDADO]	9	0	1.9	0.0	188.7
2	[CUELLO]	8	0	1.7	0.0	167.7
3	[RELOJ]	6	0	1.3	0.0	125.8
4	[CASA]	5	0	1.0	0.0	104.8
5	[BICICLETA]	5	0	1.0	0.0	104.8
6	[RUIDO]	4	0	0.8	0.0	83.9
7	[BOTELLA]	3	0	0.6	0.0	62.9
8	[AMOR]	3	0	0.6	0.0	62.9
9	[CERROJO]	3	0	0.6	0.0	62.9
10	[MANO]	3	0	0.6	0.0	62.9
11	[CRUZ]	2	0	0.4	0.0	41.9

SEC 2 (ACAD): 4,999,945 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	[PARTE]	12	0	2.4	0.0	240.0
2	[FORMA]	11	0	2.2	0.0	220.0
3	[CORDILLERA]	10	0	2.0	0.0	200.0
4	[ELECTRÓN]	8	0	1.6	0.0	160.0
5	[POLIMERASA]	8	0	1.6	0.0	160.0
6	[CARBONO]	7	0	1.4	0.0	140.0
7	[ÁCIDO]	7	0	1.4	0.0	140.0
8	[SERIE]	7	0	1.4	0.0	140.0
9	[TRANSPORTE]	7	0	1.4	0.0	140.0
10	[AMINOÁCIDO]	6	0	1.2	0.0	120.0
11	[SISTEMA]	6	0	1.2	0.0	120.0

Gráfico 18: Colocaciones por género

los sinónimos de una palabra determinada. P. ej., el gráfico 17 muestra la frecuencia de sinónimos de *guapo* en los cuatro géneros principales (académico, periodístico, ficción y oral). Podemos observar que algunas palabras (como *lindo* y *bonito*) son mucho más frecuentes en los géneros informales (especialmente hablados) que en los géneros más formales, como el académico.

Finalmente, el uso y el significado de una palabra también se ve afectado por el género. P. ej., el gráfico 18 muestra las colocaciones de *cadena* en ficción (a la izquierda) y académica (a la derecha). Obsérvese que, en la ficción, *cadena* generalmente se refiere a una cadena física (es decir, algo que uno puede sostener en la mano), mientras que en el género académico su uso es más metafórico y, a menudo, se refiere a sustantivos abstractos ('serie', 'sistema', etc.).

Como podemos ver, el género constituye un factor crucial en términos de frecuencia y significado. *El Corpus del Español* nos permite mostrar los resultados por género de manera fácil y rápida, búsqueda que no es posible con los corpus *CREA* y *CORDE*.

Tabla 2: Frecuencia de palabras de *-idad* en el *Corpus del Español* original (2002)

> 100 resultados	NONE
50–100	sonoridad, inmovilidad, expresividad, obscuridad
20–49	morbilidad, unicidad, posmodernidad, adaptabilidad, cotidianeidad, discrecionalidad, selectividad, salubridad, elegibilidad, insensibilidad, disconformidad, mensualidad, afectividad, deformidad
10–19	perpetuidad, receptividad, morosidad, anormalidad, hipersensibilidad, disparidad, extremidad, interioridad, permeabilidad, corresponsabilidad, perversidad, susceptibilidad, viscosidad, emotividad, natividad, plasticidad, asiduidad, promiscuidad, salinidad, virilidad, homogeneidad, inutilidad, frivolidad, perplejidad, voracidad
< 10	empleabilidad, trazabilidad, centralidad, consanguinidad, invisibilidad, inhabilidad, estadidad, habitabilidad, alteridad, sociabilidad, probidad, banalidad, materialidad

12 El nuevo *Corpus del Español* (Web/Dialectos)

Si bien es muy útil el componente original «Histórico/Género» de 100 millones de palabras, el *Corpus del Español* presenta algunas limitaciones. La primera es el tamaño: solo hay 20 millones de palabras del siglo XX y esta cifra resulta cuantitativamente escasa para una investigación en profundidad del léxico. La segunda es que el *Corpus del Español* original termina en 1999, no contiene textos del siglo XXI. Y la tercera es que el corpus original no permitía a los usuarios comparar el léxico en diferentes países.

Para abordar estas tres limitaciones creamos una gran extensión del *Corpus del Español* en 2014–2015. El nuevo corpus alcanza un tamaño de dos mil millones de palabras, lo que significa que es 100 veces el tamaño de la parte del español del siglo XX en el *Corpus del Español* original (2002). Todos los textos basados en la web se recopilaron en 2014–2015, por lo que el corpus representa mejor el español reciente. Finalmente, permite a los usuarios comparar frecuencias léxicas en 21 países diferentes de habla hispana.

En términos de tamaño, un corpus que es 100 veces más grande proporciona datos mucho más ricos y mayor información sobre la variación léxica. P. ej., considérese la Tabla 2, que muestra aquellas palabras que terminan en *-idad*, que tienen una frecuencia de entre 2000 y 3000 ocurrencias en el nuevo corpus de dos mil millones de palabras. La tabla muestra cuántas veces aparecen en el corpus de más de 100 millones de palabras (donde 20 millones de palabras son del siglo XX).

Solo unas pocas palabras tienen una frecuencia de al menos 50 ocurrencias en el corpus anterior, la mayoría ocurre menos de 20 veces. Con un número tan limitado de apariciones, es muy poco lo que se puede hacer para investigar el significado y uso de las palabras o investigar su frecuencia en todos los géneros. Además, en algún punto hay tan pocos resultados que la misma palabra *frecuencia* se vuelve problemática. P. ej., una voz que aparece solo 15 veces puede aparecer en solo 3–4 textos diferentes, y la palabra podría no encontrarse en absoluto en un corpus que tuviera una composición textual ligeramente diferente. En algún punto, los términos con baja frecuencia son simplemente «ruido».

Además de la frecuencia de voces aisladas, el nuevo (2016) corpus de dos mil millones de palabras también proporciona datos mucho más ricos para palabras en contexto con otras palabras. P. ej., en el *Corpus del Español* original, solo hay 67 adjetivos diferentes que aparecen cinco veces en el contexto gente + ADJ (*gente joven, pobre, rica... ponderosa, popular*), mientras que hay aproximadamente 1445 adjetivos diferentes en el corpus nuevo de dos mil millones de palabras (p. ej., *visionaria, desmotivada, sumisa, enfadada, gritona, caritativa*). Del mismo modo, solo hay 24 nombres diferentes que aparecen cinco veces o más antes de una forma de *rico* en el corpus anterior (*países, comerciantes, alimentos*), mientras que hay alrededor de 650 nombres diferentes en el corpus de dos mil millones de palabras (*vocabulario, parientes, subsuelo, guion, piedras*). Como ejemplo final, solo hay unos diez sustantivos que aparecen al menos cinco veces en cuatro palabras después de una forma de *destrozar* en el corpus anterior (p. ej., *corazón, alma, guerra, árboles, puerta*), pero hay más de 570 nombres en el nuevo corpus (p. ej., *canción, confianza, películas, cadenas, pedacitos, hacha*). En la mayoría de los casos, hay 50–60 veces más colocaciones en el nuevo corpus de dos mil millones de palabras, lo que, por supuesto, proporciona a los investigadores datos mucho más detallados y útiles sobre el significado y el uso de estas palabras.

13 Búsqueda y estudio de neologismos

Desde el punto de vista del cambio léxico, el corpus nuevo también proporciona datos mucho más ricos para los neologismos. Esto es consecuencia del incremento del corpus (100 veces más grande que la parte del siglo XX del *Corpus del Español* original), pero también se debe a que el nuevo corpus contiene textos muy recientes. Los dos millones de textos fueron extraídos de la Web en 2014–2015, una década y media después de haber sido introducidos los últimos textos en el *Corpus del Español* original.

Hay muchos miles de palabras que aparecen al menos 500 veces en el corpus más reciente, pero menos de diez veces (en muchos casos nada) en el corpus

más antiguo y pequeño. Por supuesto, muchas de estas son palabras relacionadas con la tecnología (p. ej., *blog, web, internet, celular, navegador, correo electrónico, clic, tweet*), pero otras son palabras relacionadas con otros campos (p. ej., *documental, implementación, fiscalía, inversionista, globalización, operativo, migrante, sostenibilidad, biodiversidad*).

Como una prueba más de la extensión de estos neologismos, los sustantivos que terminan en *-idad* y que se anotan a continuación aparecen al menos 700 veces en el nuevo corpus, pero menos de cinco (generalmente 0 veces) en el corpus antiguo: *interoperabilidad, catolicidad, viralidad, masividad, proactividad, escalabilidad, colonialidad, transversalidad, transexualidad, digestibilidad, sincronicidad, ciberseguridad, emocionalidad, accidentalidad, inexequibilidad, tipicidad, ruralidad.* Las palabras que terminan en *-ismo* incluyen: *kirchnerismo, chavismo, fujimorismo, extractivismo, multiculturalismo, uribismo, massismo, madridismo, emprendedurismo, agnosticismo, biomagnetismo, ateísmo, vaginismo, veganismo, ventajismo, macrismo, bruxismo, cateterismo.*

Las palabras que terminan en *-ción* y aparecen al menos 700 veces en el corpus nuevo y más amplio, pero que prácticamente desaparecen en el corpus más antiguo y reducido son: *autenticación, fidelización, virtualización, mercantilización, victimización, geolocalización, encriptación, dispensación, precarización, suplementación, judicialización, visibilización, abducción, desafección, compartición, cosificación, procrastinación, remediación, iteración, resignificación, disrupción, deslocalización, feminización, extranjerización, redirección, evicción, invisibilización, bancarización, demonización, gamificación, previsualización, propiciación.*

Esto no quiere decir que todas estas palabras constituyan neologismos, solo que rara vez aparecen en el corpus más antiguo y pequeño y que, por el contrario, resultan frecuentes en el corpus más moderno y de mayor dimensión. Los lexicógrafos experimentados, por supuesto, querrían investigar tales palabras con más detalle. Pero el punto principal es que el nuevo corpus de dos mil millones de palabras proporciona, además, los datos para la investigación de voces donde ha habido un relativo aumento de frecuencia en los últimos 15–20 años.

14 Desvío sincrónico # 1: dialectos

Además de identificar neologismos, una de las ventajas del nuevo corpus es que permite a los investigadores obtener la frecuencia de cada palabra o frase en 21 países diferentes de habla hispana[6]. Para algunos países, como España, el nuevo

6 Para datos similares de un corpus de inglés análogo, vid. Davies/Fuchs 2015.

corpus contiene hasta 459 millones de palabras (con 261 para México, 183 para Argentina y 180 para Colombia). Incluso entre los países con menor representación cuantitativa en el corpus, cada uno tiene al menos 30 millones de palabras (p. ej., 39 para El Salvador, 39 para Honduras, 37 para la República Dominicana, 36 para Puerto Rico, 35 para Nicaragua, 33 para Paraguay y 32 para Costa Rica).

Las siguientes son solo algunas de las palabras que se registran con una frecuencia más alta en un país que en los otros[7]:

- Caribe
  Puerto Rico *ay bendito, chavos, chiringa, mahones, habichuela* (+DR), *zafacón* (+DR); Cuba *guajiro, jimaguas, babalao, bitongo, pedir botella*; Rep Dom *mangú, fucú, tutumpote, mangulina, mofongo* (+PR).
- México y América Central
  México *ándale, híjole, órale, güero, (muy) padre, chamaco* (CAm/Car), *pinche, popote, charola* Guatemala *huipil, canche, muchá, patojo, chafa* (+HN), *chirmol, canche*; El Salvador *cipote, chero, pupusa, cuilio, bayunco, piscucha*; Honduras *catracho, papada*; Nicaragua *chavalo, maje* (+CAm), *pinol, pinolillo, chigüín, vigorón, gallo pinto* (+CR), *idiay* (+CR) Pánama *fulo, chombo, guandul*; Costa Rica *chinear, guila, chunche.*
- América del Sur
  Colombia *cachaco, cachifo, verraquera, estar mamado, guandoca, biche*; Venezuela *bojote, coroto, catire, gafo, macundales, arepa, cachapa, cambur, caraotas, jojoto*; Ecuador *chumar, chulla, montuvio, omoto*; Perú *anticucho, jebe, chupe, pisco, jora, chompa* (+CL/EC), *choclo* (+CL/EC); Bolivia *opa, colla, chuño, lagua*; Chile *pololo*, pololear, achuntar, bencina, bacán, fome, huaso*; Paraguay *ñembo, ñanduti, karai, yopará, mitai*; Uruguay *tropero, hacer * sota, con fritas*; Argentina *pibe, fiaca, morfar, falopa, sobre el pucho, falluto, cafishi.*
- España *ordenador, aparcar, enfadar, gafas, zumo, chulo, guay, coger, bolígrafo, patata, melocotón, echar de menos, vale.*

Además de poder analizar la frecuencia de una palabra o frase específica en los 21 países, también es posible hacer que el corpus produzca una lista de todas las palabras que son más comunes en un país (o conjunto de países) que en otro. P. ej., la siguiente tabla muestra palabras de *-ismo* que son más comunes en Venezuela (izquierda) que en Colombia, México, Argentina o España (derecha):

7 Téngase en cuenta que los países y la región entre paréntesis indican que la palabra también tiene una alta frecuencia relativa en estas otras zonas. La lista de palabras se corresponde con los datos léxicos aportados por Lipski 1996.

SEC 1 (Venezuela): 98,170,248 WORDS							SEC 2 (Argentina, Colombia, España...): 1,008,426,240 WORDS						
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	JALABOLISMO	44	1	0.4	0.0	452.0	1	MASSISMO	1249	1	1.2	0.0	121.6
2	ESCUALIDISMO	21	1	0.2	0.0	215.7	2	PRIISMO	362	0	0.4	0.0	35.9
3	MADURISMO	141	10	1.4	0.0	144.8	3	EMPRENDEDORISMO	251	0	0.2	0.0	24.9
4	GOMECISMO	41	3	0.4	0.0	140.4	4	GARANTISMO	254	1	0.3	0.0	24.7
5	VENTAJISMO	812	63	8.3	0.1	132.4	5	PARKINSONISMO	219	0	0.2	0.0	21.7
6	PUNTOFIJISMO	165	14	1.7	0.0	121.1	6	NEOPLATONISMO	218	1	0.2	0.0	21.2
7	RIQUISMO	22	3	0.2	0.0	75.3	7	CHARRISMO	204	1	0.2	0.0	19.9
8	HIPISMO	295	41	3.0	0.0	73.9	8	SOCORRISMO	196	0	0.2	0.0	19.4
9	CASTRO-COMUNISMO	55	8	0.6	0.0	70.6	9	LIBERTARISMO	184	0	0.2	0.0	18.2
10	ECO-SOCIALISMO	27	4	0.3	0.0	69.3	10	CRISTINISMO	557	3	0.6	0.0	18.1
11	VENEZOLANISMO	27	4	0.3	0.0	69.3	11	COPERNICANISMO	173	0	0.2	0.0	17.2
12	CASTROCOMUNISMO	58	9	0.6	0.0	66.2	12	MOTOTAXISMO	171	0	0.2	0.0	17.0

Gráfico 19: Comparación por dialecto en el corpus Web/Dialecto

15 Comparación con el *CORPES*

Al analizar los datos históricos del *Corpus del Español* original, los comparamos con los datos del corpus *CORDE* y, cuando discutimos los datos sincrónicos del *Corpus del Español* original, lo comparamos con el corpus *CREA*. Existe un tercer corpus de la Real Academia Española que es relevante, en términos de una comparación con la parte más moderna de dos mil millones de palabras «Web/Dialectos» del *Corpus del Español*, que hemos discutido en las Secciones 12–14; este tercer corpus es el *CORPES* (*Corpus del Español del Siglo XXI*).

En cuanto a posibilidades de estudios sobre léxico, existen algunas diferencias importantes entre el *CORPES* y la nueva extensión de dos mil millones de palabras del *Corpus del Español* (que llamaremos *CDE-2* en esta sección). La primera diferencia importante es el tamaño: el *CORPES* tiene aproximadamente 175 millones de palabras, que es menos del 10 % del tamaño del *CDE-2*. Esto es, en caso de que una palabra tenga 200 ocurrencias en *CDE-2* (probablemente un número suficiente para muchos tipos de investigación), en el *CORPES* probablemente tendría menos de 20 ocurrencias, lo que resulta mucho más problemático. En segundo lugar, los materiales que integran el *CDE-2* son más recientes, lo que resulta importante para detectar neologismos. Solo alrededor del 17 % del *CORPES* es posterior a 2010, mientras que el 100 % de los datos del *CDE-2* corresponde a ese período (los textos se recopilaron en 2014–2015).

También hay diferencias importantes en cuanto a la arquitectura e interfaz del corpus. Aunque el *CORPES* puede generar gráficos de frecuencia útiles para palabras individuales o para una frase exacta, no es posible encontrar la frecuencia de cadenas coincidentes en una búsqueda como «*menos * que*» (p. ej., *menos valor/intenso/arriesgado/sano que*) o «adjetivo + *ojos*» (p. ej., *ojos bellos/nublados/*

cristalinos/cariñosos). En relación con esto, con el *CORPES* no es posible obtener la frecuencia de colocaciones significativas (es decir, es típico del *CORPES* mostrar artículos o preposiciones como las colocaciones más importantes de una palabra, como es también el caso del *CREA* y el *CORDE*, v. el Gráfico 10). El *CORPES* tampoco permite la comparación de colocaciones entre palabras diferentes (p. ej., *potente* y *poderoso*, o *iluminar* y *alumbrar*), aunque puedan ser búsquedas útiles para poner de manifiesto las diferencias de significado. Y, por último, el *CORPES* no permite comparaciones entre todas las palabras en diferentes países para encontrar las que son más comunes en un país que en otro (v. Gráfico 19). Por tanto, aunque el *CORPES* tiene una arquitectura e interfaz más avanzada que el *CORDE* o el *CREA*, todavía resulta bastante limitada en los tipos de búsquedas que permite.

16 Datos muy recientes y datos continuamente actualizados

En 2016 lanzamos un corpus de textos en inglés llamado *NOW* («Noticias en la web»). Este corpus crece automáticamente en tamaño en alrededor de 5–6 millones de palabras por día y está basado en aproximadamente 10 000 nuevas URL diarias extraídas de Google News. Con este corpus, pueden rastrearse los cambios en la frecuencia y el uso de palabras en el transcurso de meses, semanas e incluso días, lo que es obviamente muy útil para observar cambios extremadamente recientes en el lenguaje. P. ej., los investigadores pueden rastrear la frecuencia en el tiempo de las palabras «nuevas» en el idioma (que han surgido a partir de 2007), como *Brexit, manspreading, makerspace, gig economy, dadbod, momager, swatting, walkscore, trigger warning, mommy porn, normcore, listicle, sufferfest, catfishing, sapiosexual, nomophobia, omnishambles, humblebrag, FOMO, precariat, filter bubble, range anxiety, collaborative consumption, churnalism, birther, truther, staycation, glamping, locavore, voluntourism, freegan.*

Hemos recopilado los textos para un corpus similar en español (que llamaremos aquí *NOW-Español*), que se lanzará en mayo de 2018. Cuando se publique contendrá aproximadamente 4800 millones de palabras desde enero de 2012 hasta mayo de 2018, y luego crecerá en tamaño alrededor de 140 millones de palabras cada mes (alrededor de 1600 millones de palabras cada año). De la misma manera que ya es posible para el inglés, los usuarios de *NOW-Español* podrán seguir la frecuencia de cualquier palabra, frase o construcción sintáctica a lo largo del tiempo (incluso a nivel de meses y semanas), y también tendrán las listas de neologismos en los textos que generará el corpus automáticamente.

Aunque *NOW-Español* aún no está disponible, podemos utilizar algunos datos preliminares para mostrar cómo *NOW-Español* se compara con la extensión de

dos mil millones de palabras en el *Corpus del Español* que, como hemos indicado, contiene textos recopilados de 2014–2015. P. ej., las siguientes palabras que terminan en *-idad* aparecen al menos 10 veces en *NOW-Español*, pero no figuran en los datos del *Corpus del Español* de 2014–2015: *poliautoinmunidad, bancabilidad, promovilidad, superelasticidad, sucrosidad, sucreñidad, inexportabilidad, teleseguridad.* Las palabras terminadas en *-ismo* incluyen *narcouribismo, uribestialismo, cadivismo, cuñadismo, carnismo, larretismo, raspacupismo, chaveztialismo, urbanudismo, figurinismo*. Y las palabras terminadas en *-ción* incluyen *narcorevolución, uberización, multihabitación, posdesmovilización, incoración, cibervictimización, kilometración, deaferentación, electrorreducción, multilateración, desdiabolización, termonebulización, retermalización, hipsterización*. Una vez más, no hay garantía de que todas estas palabras sean neologismos (es decir, que sean novedades de los últimos 3–4 años), solo de que aparecen en el español actual y no figuran en el *Corpus del Español* de dos mil millones de palabras. Los lexicógrafos tendrán acceso a esta rica información para observer los cambios más recientes en el lenguaje.

17 Conclusión

Los grandes corpus accesibles mediante Internet han ayudado a revolucionar el campo de la lexicografía histórica. Con solo unos pocos clics del ratón, los investigadores pueden buscar corpus que contengan cientos de millones (y ahora miles de millones) de palabras de miles (y ahora millones) de textos.

Cada corpus tiene sus propias fortalezas y debilidades. Los corpus *CORDE* y *CREA* de la Real Academia Española son bastante sólidos en términos de «corpus textual», pero tienen arquitecturas e interfaces muy limitadas y obsoletas que limitan el acceso de los investigadores a estos datos. Por otro lado, la parte «histórica/de género», original del *Corpus del Español* (que fue lanzado en 2002) constituye un corpus de dimensiones más reducidas que el *CORDE* o el *CREA*, pero permite varios tipos de investigación que no pueden llevarse a cabo con estos dos corpus, lo que incluye recuperar la frecuencia de palabras, frases y subcadenas (p. ej., prefijos o sufijos) por siglo o género, seguir los patrones de cambio de colocaciones para observar las modificaciones en el significado y uso, y utilizar el índice integrado y las listas de palabras personalizadas para ver cómo las voces compiten por el «espacio semántico» a lo largo del tiempo (o en diferentes géneros).

Uno de los usos más interesantes de los grandes corpus en línea es poner de relieve los cambios muy recientes en el lenguaje, como puede hacerse con la nueva extensión del *Corpus del Español*, que contiene dos mil millones de palabras de

21 países diferentes de habla hispana de 2014–2015. A partir de mayo de 2018, será posible utilizar *NOW-Español* para analizar los cambios que se producen virtualmente en «tiempo real», como ya es posible para *NOW-English*. En algún momento, incluso, sería posible crear corpus que monitoricen continuamente trillones de palabras de las redes sociales para rastrear con increíble precisión cómo las palabras y frases se propagan a lo largo del tiempo a través de distintas comunidades de habla y dialectos[8].

En definitiva, los grandes corpus en línea nos permiten realizar muchos tipos de investigaciones que apenas eran imaginables hace 15 o 20 años y, con las mejoras y avances en la tecnología, podemos darnos cuenta de que solo estamos en el comienzo de lo que se puede llegar a hacer.

Referencias bibliográficas

Davies, Mark (2002): «Un corpus anotado de 100.000.000 de palabras del español histórico y moderno», en *Actas del XVIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Valladolid: SEPLN, 21–27.

Davies, Mark (2005a): «Advanced research on syntactic and semantic change with the Corpus del Español», en Claus Pusch *et al.* (eds.), *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tubinga: Guntar Naar, 203–214.

Davies, Mark (2005b): «The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation», *International Journal of Corpus Linguistics* 10, 301–328.

Davies, Mark (2008): «Spanish and Portuguese corpus linguistics», *Studies in Hispanic and Lusophone Linguistics* 1, 149–186.

Davies, Mark (2010): «Creating useful historical corpora: A comparison of *CORDE*, the Corpus del Español, and the Corpus do Português», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus*. Fráncfort/Madrid: Vervuert/Iberoamericana, 137–166.

Davies, Mark (2012a): «Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English», *Corpora* 7, 121–157.

8 Véase la Sección 7 de Davies 2015.

Davies, Mark (2012b): «Examining Recent Changes in English: Some Methodological Issues», en Terttu Nevalainen y Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*. Oxford: Oxford Univ. Press, 263–287.

Davies, Mark (2015): «Corpora: An introduction», en Douglas Biber y Randi Reppen (eds.), *Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 11–31.

Davies, Mark (en prensa a): «Using large online corpora to examine lexical, semantic, and cultural variation in different dialects and time periods», en Eric Friginal *et al.* (eds.), *Corpus-Based Sociolinguistics*. Londres: Routledge.

Davies, Mark (en prensa b): «Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design», en Carla Suhr, Terttu Nevalainen e Irma Taavitsainen (eds.), *From data to evidence in English language research* (Digital Linguistics). Leiden: Brill.

Davies, Mark/Robert Fuchs (2015): «Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (*GloWbE*)», *English World-Wide* 36, 1–28.

Firth, J. R. (1957): *Papers in Linguistics 1934–1951*. Londres: Oxford University Presiglos.

Lipski, John (1996): *El español de América*. Madrid: Cátedra.

Virginia Bertolotti y Concepción Company Company

El corpus para América: *CORDIAM*

Resumen: El *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) fue creado para poder historiar el español de América y para hacer lingüística histórica general con datos de muchas variedades del español. Estos datos, exclusivamente americanos, se procesan a través de una interfaz amigable y eficiente, producto del trabajo conjunto de informáticos e investigadores en lingüística histórica e historia de la lengua. Este trabajo muestra las características lingüísticas y textuales del *CORDIAM*, sus características informáticas y da cuenta, además, del proceso de toma de decisiones para la creación de cada uno de los tres subcorpus que conforman el *CORDIAM*: *CORDIAM-Documentos*, *CORDIAM-Prensa* y *CORDIAM-Literatura*. Muestra también cómo los textos incluidos representan facetas diversas de la cultura en América.

Palabras clave: Lingüística de corpus, Corpus escrito, Lingüística histórica, Español en América

Abstract: The *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) is presented in this paper. It was created to make the history of Spanish in America and historical linguistics with data from many dialects of Spanish. These exclusively American data are processing by a friendly and efficient interface, which is the result of the joint work of computer engineers and investigators both in historical linguistics and in language history. This work shows the linguistic and textual characteristics of *CORDIAM*, its computing characteristics and, in addition, explains the decisions underlying process for the creation of each of the three subcorpus of *CORDIAM*, which are *CORDIAM-Documents*, *CORDIAM-Press*, and *CORDIAM-Literature*. It also shows how the included texts exhibit different aspects of the American culture.

Keywords: Corpus linguistics, Written corpus, Historical linguistics, Spanish in America

1 Presentación y objetivos

Los *corpus lingüísticos*, escritos y orales, diacrónicos y sincrónicos, las *ediciones críticas de textos*, antiguos y modernos, así como los *atlas lingüísticos*, son, como es sabido, los tres tipos de productos fundamentales que constituyen infraestructura para la investigación lingüística, y son, por ello, soportes esenciales para abrir nuevos horizontes de investigación, hallar nuevas evidencias, descriptivas