Citius, maius, melius: del CREA al CORPES XXI

Guillermo Rojo

1. Introducción

En 1995, la Real Academia Española tomó la decisión de acometer la construcción del *Corpus de referencia del español actual* (CREA) para lograr así mayor amplitud y seguridad en los materiales utilizados en la confección de su diccionario. Vistos los primeros resultados obtenidos, se decidió ampliar el banco de datos que comenzaba a formarse para incorporar también el español de períodos anteriores y, de acuerdo con los mismos objetivos generales, construir el *Corpus diacrónico del español* con el fin de disponer de materiales mejores y mucho más voluminosos para la redacción del *Diccionario histórico del español*. Las primeras versiones de ambos corpus fueron publicadas en 1998 y ampliadas y mejoradas hasta la finalización de ambos proyectos en 2008. A lo largo de todos esos años y hasta la actualidad, la RAE y todas las Academias que forman con ella la Asociación de Academias de la lengua española (ASALE) han basado en el CREA y el CORDE todas las obras que han ido publicando. Pero el impacto de estos dos corpus ha sido considerablemente mayor, puesto que han supuesto una modificación radical también en los modos de trabajo de cuantos se dedican a la investigación de la lengua española.

Aunque siguen siendo útiles, tanto el CREA como el CORDE tienen un diseño que, dado que fueron concebidos hace casi veinte años, no resulta congruente con las prácticas actuales, un tamaño insuficiente para buena parte de las necesidades que se plantean en la investigación y una aplicación de búsqueda rica y flexible, pero un tanto envejecida. Como consecuencia de todo ello, las Academias de ASALE decidieron, en 2007, acometer la creación del *Corpus del español del siglo xxi* y encargar su realización a la Real Academia Española. La primera versión beta del CORPES se presentó en el Congreso internacional de la lengua española (CILE) celebrado en Panamá en noviembre de 2013 y se publicó como versión 0.6. en diciembre de ese mismo año. En abril de 2015 se publicó la versión 0.8. del CORPES, que acaba de entrar en su segunda fase, cuya finalización está prevista en diciembre de 2018.

El propósito de este trabajo es analizar las novedades que supone el CORPES en la lingüística española por un lado y en la lingüística de corpus por otro. Para ello, en el apartado 2 se analizan las características de sus antecedentes (fundamentalmente el CREA), enmarcados en el contexto de la época, y también algunas cuestiones generales referidas al lugar que ocupan los corpus de referencia en la lingüística de corpus actual. El apartado 3 se centra en las características que tiene el CORPES, con especial atención a lo que supone novedad con respecto al CREA y el CORDE y también a otros corpus de español.

2. Antecedentes inmediatos: el CREA y el CORDE

Como es sobradamente conocido, en 1995 la Real Academia Española tomó la decisión de emprender la construcción de un banco de datos electrónico del español contemporáneo, el *Corpus de referencia del español actual* (CREA). La intención básica del proyecto era proporcionar a la RAE y a todas las demás integrantes de la Asociación de Academias de la lengua española (ASALE) un recurso gracias al cual fuera posible documentar con mayor seguridad los usos lingüísticos reales y, como consecuencia de ello, basar mejor las decisiones de carácter normativo

Borrador final. Pendiente de aparición en Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín: de Gruyter. En prensa (2016).

que estas instituciones han de adoptar continuamente. Pero desde el principio quedó claro que se trataba de construir un banco de datos al que pudieran acceder y del que pudieran beneficiarse también todas las personas interesadas en el conocimiento de las características del español actual, con finalidades investigadoras (no solo en lingüística), de aplicación a la producción de materiales de diferentes tipos, documentación, etc. Vistos los primeros resultados, muy pocos meses después la Academia decidió crear otro corpus textual, el *Corpus diacrónico del español* (CORDE), cuya finalidad era reunir en formato electrónico una gran cantidad de textos en español correspondientes al período comprendido entre los orígenes de la lengua y el punto de arranque del CREA. El proyecto fue desarrollado conjuntamente a un ritmo bastante alto gracias a la financiación parcial del Ministerio de Educación en los primeros años de trabajo.

Ambos corpus, pues, constituyen en realidad un proyecto único que se escinde en dos subproyectos en atención a las características básicas de los que se consideraban sus ámbitos de trabajo principales: el CREA iba a ser la fuente básica de datos para el español contemporáneo y el CORDE serviría fundamentalmente para los estudios de carácter diacrónico. En su diseño inicial, el CREA comprendía textos de los más diversos tipos y géneros, con un 10 % del total formado por transcripciones de textos orales, procedentes de todos los países hispánicos y con una distribución general que asignaba el 50 % a textos producidos en España y el otro 50 % a textos producidos en América. Tendría un volumen total de 125 millones de formas correspondientes a los 25 años comprendidos entre 1975 y 1999. Se estructuraba en cinco quinquenios, a cada uno de los cuales correspondía un porcentaje que, siguiendo una línea muy utilizada en aquel momento, iba aumentando desde los más antiguos a los más modernos (10 %, 15 %, 20 %, 25 % y 30 %, respectivamente). Por su parte, el CORDE fue proyectado para reunir trescientos millones de formas procedentes de los más variados tipos y géneros, de todos los países hispánicos (incluida Filipinas) desde los orígenes de la lengua hasta 1974.

Los treinta años transcurridos desde la aparición del *Brown Corpus* hasta el arranque del proyecto de la RAE pueden hacer pensar que la decisión fue tomada con un retraso notable con respecto a la marcha general de la lingüística de corpus (LC). La distancia temporal es innegable, pero hay que tener en cuenta que durante esos años no se habían producido muchos corpus textuales y, por supuesto, muy pocos con el alcance y el volumen que tienen el CREA y el CORDE. Aunque la visión dominante de la historia de la LC se refiere sistemáticamente a un período inicial muy difícil, en un contexto hostil dominado por la pujante y novedosa orientación chomskyana, lo cierto es que esa caracterización es válida solo para los Estados Unidos, mientras que en países como Inglaterra, Noruega, Suecia y, en menor medida, Francia, Alemania o Italia la LC tuvo en esa época un desarrollo creciente y progresivo desde sus arranques respectivos.[1]

En 1995, el momento en que la Academia decide emprender la construcción de CREA y CORDE, las referencias fundamentales están en el inglés, lengua en la que al corpus conocido como Lancaster-Oslo-Bergen (LOB) y el COBUILD han seguido otros y, sobre todo, el *British National Corpus* (BNC), constituido por cien millones de formas y que es, sin duda, el modelo en el que basamos las características del CREA. En el ámbito hispánico, dejando a un lado los que podemos considerar proyectos de transición,[2] en la época inmediatamente anterior a la planificación de

1 Tienen gran importancia en esta fase los corpus construidos para uso en proyectos lexicográficos. Son, en general, proyectos de alto coste económico y también organizativo, que implican un cambio en la práctica lexicográfica que tardará algún tiempo en consolidarse y emprender el camino que lleva a la situación actual. Para detalles, vid. Rundell (2012, 18).

2 Es la denominación que empleo en Rojo (2015) para los que se sitúan en las cercanías de la LC, pero sin llegar a emplear recursos electrónicos, como el *Proyecto de estudio coordinado de la norma lingüística culta*, y los que suponen la preparación de textos ya en formato electrónico, pero sin llegar a constituir un corpus en sentido estricto, como los materiales reunidos en el *Hispanic Seminar of Medieval Studies* para la redacción del *Dictionary of Old Spanish*

CREA y CORDE pueden encontrarse corpus insertos en cinco bloques diferentes. En primer lugar, los de tamaño pequeño (incluso para los estándares de la época), como el corpus de Lovaina o las dos entregas de ENTREVIS. El segundo bloque es el constituido por los que se construyen, siguiendo el modelo del COBUILD, para servir a propósitos lexicográficos, como el *Vox-Biblograf*, el CUMBRE o el *Corpus del español mexicano contemporáneo* (CEMC). El tercer grupo está constituido por corpus de tamaño pequeño que se desarrollan en el marco de proyectos europeos, como CRATER, NERC o PAROLE. El cuarto bloque lo forman varios corpus de carácter general y volumen reducido, como los dirigidos por Francisco Marcos Marín en diversas acciones patrocinadas por la Sociedad Estatal del Quinto Centenario[3] o el corpus LEXESP. Por fin, en la dimensión diacrónica, hay que mencionar el proyecto ADMYTE, cuyos responsables son Francisco Marcos Marín, Charles Faulhaber, Ángel Gómez Moreno y Antonio Cortijo Ocaña.[4]

El retraso con que partieron los corpus de la RAE tuvo algunos efectos beneficiosos. En primer lugar, la evolución de las computadoras, con el enorme incremento de capacidad y velocidad experimentado en aquellos años, hacía posible pensar en emprender la confección de corpus de cientos de millones de formas, siguiendo y superando el modelo establecido por el BNC. Al tiempo, los procedimientos utilizables para la digitalización de textos habían avanzado considerablemente, de modo que la posibilidad de usar escáneres y programas de reconocimiento óptico de caracteres aliviaba mucho el penoso trabajo de conversión de texto impreso en texto electrónico.[5] Por otra parte, el desarrollo de la *Text Encoding Initiative* (TEI) establecía un modelo y un estándar de codificación adaptable de forma no excesivamente complicada a cualquier proyecto de corpus. La conjunción de estos tres factores (mayor capacidad y velocidad, facilidad en la digitalización y sistema estándar de codificación en SGML) produjo un cambio enorme en la propia concepción de los corpus, que dejaron de ser conjuntos consultables solo de forma integral, con lo que ello supone para la representatividad de los materiales incluidos y su equilibrio, y pasaron a constituir complejos textuales en los que era posible construir, de forma dinámica, subcorpus virtuales configurados mediante la selección de diferentes valores en los parámetros de construcción (soporte, tipo de texto, año, características sociolingüísticas, país, etc.). Por último, el desarrollo de Internet, aunque todavía muy reducido en aquel momento, permitía ya pensar directamente en un modelo cliente-servidor que hiciera posible la consulta cómoda y sencilla de los corpus desde cualquier parte del mundo, con cualquier máquina, cualquier sistema operativo y cualquier navegador.

El CREA y el CORDE surgieron, pues, en un contexto favorable a la creación de corpus de referencia y encajaban perfectamente en los estándares del momento en cuanto a tamaño, codificación, estructuración y sistema de recuperación de datos. Además, presentaban algunas características adicionales de especial interés. El CORDE fue proyectado con un tamaño (300 millones de formas) difícilmente alcanzable por un corpus de carácter diacrónico.[6] El CREA tenía un diseño que lo situaba a caballo entre los corpus cerrados (al estilo del BNC), que se terminan

Language, los textos periodísticos informatizados en la Universidad de Göteborg o las obras de teatro informatizadas por Hiroto Ueda. Vid. Rojo (2015, apdo. 2) para detalles sobre estos proyectos y las referencias bibliográficas correspondientes.

3 Son el *Corpus de referencia de la lengua española contemporánea*, el *Corpus lingüístico de referencia de la lengua española en Argentina* y el *Corpus lingüístico de referencia de la lengua española en Chile*.

4 Para detalles sobre todos estos y algunos otros proyectos y las referencias bibliográficas pertinentes, vid. Rojo (2015, apdo. 3).

5 Salvo en los textos de prensa, con los que el trabajo de conversión a formato electrónico seguía presentando muchas dificultades. Curiosamente, la situación dio un giro radical con la generalización de la prensa digital. Cf. Rojo / Sánchez (2010, cap. 4). para una perspectiva panorámica de la evolución experimentada.

6 La excepción más notable es, sin duda, el *Corpus of Historial American English* (COHA), construido por Mark Davies, que contiene unos cuatrocientos millones de formas procedentes de textos editados entre 1810 y 2009.

Borrador final. Pendiente de aparición en Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín: de Gruyter. En prensa (2016).

cuando alcanzan el tamaño previsto, y los abiertos, que añaden textos de forma continua, con los efectos fácilmente imaginables sobre la estabilidad de los resultados obtenidos en las consultas. Fue proyectado en realidad como el corpus textual de los últimos veinticinco años de historia del español, de modo que al período abarcado en su configuración inicial (1975-1999) se irían añadiendo luego quinquenios posteriores (2000-2004, 2005-2009, etc.). La entrada de cada quinquenio nuevo supondría la retirada del más antiguo, para mantener así siempre un tramo general de veinticinco años. Y, dado que ambos corpus constituían un proyecto conjunto, el tramo retirado del CREA se integraría en el CORDE, que iría de este modo ampliando su período de actuación. Por esa razón, los ficheros del CREA llevan en su cabecera los rasgos clasificatorios que les corresponderían en el CORDE cuando se produjera su integración en este corpus.

El diseño tenía un punto débil: el hecho de que los diferentes lustros integrados en el CREA tuvieran porcentajes diferentes sobre el total hacía que la operación de reajuste resultara especialmente complicada: con el paso al CORDE del tramo 1975-1979, el tramo siguiente (1980-1984) debería perder los textos necesarios (y hacerlo de forma congruente con el diseño general para mantener el equilibrio), dejar de suponer el 15 % del total y pasar a ser únicamente el 10 %. Como es lógico, habría que aplicar una operación equivalente sobre los demás tramos. Además de la enorme complejidad de una remodelación de este tipo, los textos retirados para ajustar los porcentajes no podrían pasar todavía a formar parte del CORDE, que no habría llegado a esos años. En consecuencia, ese ajuste no se produjo nunca y el CREA amplió su ámbito al período 2000-2004 con un tamaño similar al del inmediatamente anterior (unos 37,5 millones de formas), con lo que, cuando se cerró en 2008, el CREA había llegado a tener en torno a 160 millones de formas, las mismas que pueden ser consultadas en la actualidad.[7]

La publicación, a partir de 1998, de varias versiones provisionales del CREA y el CORDE dio lugar a un fuerte cambio en la metodología aplicada por las Academias de la lengua española en la preparación de las obras publicadas desde ese momento,[8] así como en los recursos disponibles para los investigadores del español de todas las épocas y variedades. Desde su cierre, en 2008, ambos corpus han permitido mejorar considerablemente nuestros conocimientos sobre el español.[9]

Pero, a pesar de su importancia y utilidad, es evidente que estos corpus son el resultado de un proyecto que tiene ya más de veinte años de antigüedad, con lo que ello supone, en una disciplina de desarrollo tan acelerado como la LC, para el diseño, los procesos intermedios y la recuperación de datos. El CREA, por ejemplo, posee una enorme flexibilidad para la creación de subcorpus virtuales y la consiguiente recuperación selectiva de datos, pero la versión pública no está lematizada y la aplicación de consulta presenta inconvenientes en recuperaciones de cierta complejidad.

En efecto, a lo largo de todos estos años se han producido importantes modificaciones en el proceso de diseño, construcción y explotación de corpus que hacían necesario el replanteamiento de las características que deben tener los corpus de referencia del estilo del CREA y el CORDE. En términos generales, este tipo de corpus se sitúa actualmente en una zona comprendida entre dos tendencias muy diferentes entre sí. En un extremo, los corpus pequeños (unos pocos millones de formas) que suponen una edición muy cuidada y sometida a unos principios únicos de los textos que

7 En el proceso de organización de los materiales para el *Nuevo diccionario histórico del español*, la mayor parte de los textos que componen el CORDE y el CREA han sido integrados en el *Corpus del NDHE* (CDH).

8 En efecto, todas las obras publicadas por las Academias desde 1998 (las ediciones vigésima segunda (2001) y vigésima tercera (2014) del DRAE, el *Diccionario Panhispánico de Dudas* (2005), el *Diccionario del Estudiante* (2005 y 2011), el *Diccionario Esencial de la lengua española* (2006), la *Nueva gramática de la lengua española* (2009) y la *Ortografía de la lengua española* (2010)) se han beneficiado de los datos contenidos en el CORDE y, sobre todo, el CREA.

9 En febrero de 2015, el CREA recibió casi 100.000 consultas. El CORDE, algo menos de la mitad.

los componen. Frente a la dependencia que tienen los textos que integran el CORDE con respecto a los diferentes criterios utilizados en las ediciones integradas en él, las que componen proyectos como CODEA, *Biblia medieval* o CORDIAM[10] tienen unas directrices muy marcadas y todos los textos responden estrictamente a ellas. Son, además, textos transcritos específicamente para los proyectos respectivos y pueden integrar diferentes presentaciones del mismo «texto» (por ejemplo, una edición paleográfica al lado de una edición crítica y la imagen del manuscrito). Como es lógico, ese cuidado exquisito tiene como contrapartida el limitadísimo tamaño que se puede conseguir y también la habitual restricción a un cierto tipo de textos.

Estos corpus son «small and tidy», para usar la expresión utilizada por Mair (2006). Al otro lado, el constituido por los que resultan «big and messy» y que pueden tener el *Bank of English* como su modelo inicial, se encuentran actualmente los que resultan de la tendencia conocida como 'Web as Corpus', que produce conjuntos obtenidos de modo oportunista a partir de lo que se encuentra ya en la red. En una formulación estricta, estos conjuntos textuales carecen de diseño y, por tanto, no encajan realmente en lo que se exige para que puedan recibir la consideración de corpus,[11] pero es preciso reconocer que permiten construir, con unos plazos y unos costes muy reducidos, conjuntos textuales formados por miles de millones de formas[12] y que los filtros automáticos para seleccionar los textos, evitar repeticiones, excluir las zonas escritas en lenguas diferentes, etc. han mejorado considerablemente desde los utilizados hace unos años. Cerca de este segundo tipo se encuentran también otros corpus de gran tamaño y, al menos de entrada, mucho más homogéneos, que se construyen directamente con alguno de los recursos globales existentes del estilo de la Wikipedia, las intervenciones en el Parlamento europeo, en la ONU, etc.

Los corpus de referencia no pueden competir en cuidado con los pequeños ni en tamaño con los grandes. Con unos costes elevados, pero asumibles para instituciones de cierta importancia, consiguen reunir cientos de millones de formas con un determinado diseño en el que quede garantizada la representatividad, la presencia de textos de los más diversos tipos en proporciones adecuadas y un nivel de codificación que permita la recuperación selectiva de la información a partir de los rasgos pertinentes en cada caso (país, época, tipo de texto, características de los hablantes, etc.). Constituyen, pues, un recurso intermedio que, sin negar la necesidad de los otros tipos para ciertas clases de análisis o aplicaciones, aúna un tamaño que garantiza la fiabilidad y generalidad de los datos que se pueden obtener de su análisis con el detalle de la codificación añadida, que permite una selección de datos muy fina y estructurada de modo acorde con las características propias de los textos.

Por otro lado, el aumento en la capacidad de las computadoras y la reducción de sus costes, unidos a la mayor facilidad existente para la obtención, codificación y anotación automática de textos, permiten que los corpus de referencia puedan superar la división tradicional entre corpus cerrados al estilo del *Corpus del español* construido por Mark Davies o el BNC (estables, pero condenados a una pronta obsolescencia) y corpus abiertos, como el *Bank of English* (siempre actualizados, pero sistemáticamente inestables). Un corpus de referencia puede ser concebido como el resultado de la incorporación año tras año de una determinada cantidad de formas distribuidas de acuerdo con unos

10 El *Corpus de documentos españoles anteriores a 1700* (CODEA) contiene unos 1500 documentos transcritos hasta el momento según las directrices seguidas en el proyecto *Corpus hispánico y americano en la red: textos* antiguos (CHARTA). El proyecto *Biblia medieval*, constituido por traducciones de la Biblia al castellano tiene una enorme gama de posibilidades de recuperación de datos y consta de unos cinco millones de formas. El *Corpus diacrónico y diatópico del español de América* (CORDIAM), cuya publicación está prevista para finales de 2015, contendrá la transcripción de unos 3000 documentos, con un total de unos cuatro millones de formas. Para detalles, cf. Rojo (2015, apdo. 4).

11 Cf. Sinclair (2005, 15).

12 El corpus *EsTenTen*, construido por Adam Kilgarrif tenía, en diciembre de 2013, algo más de 8300 millones de formas, etiquetadas, procedentes de todos los países hispánicos. Cf. Kilgarriff / Renau (2013).

principios constantes, de modo que está cerrado y es estable en los años que ya han sido terminados, pero está abierto y actualizado en tanto que va añadiendo nuevos textos a medida que pasa el tiempo.

3. El Corpus del español del siglo XXI (CORPES)

Esta es precisamente la línea en la que se inscribe el CORPES. Pretende ser un recurso lingüístico en el que se aúnen la riqueza y variedad de datos que solo pueden aparecer como consecuencia del aumento del tamaño del corpus y la finura que se alcanza en los corpus pequeños, aunque, como es lógico, no en el mismo grado ni con el mismo detalle. La proyección de las formulaciones generales sobre la situación y características actuales del español configura una serie de parámetros que vertebran la codificación de los textos por una parte y la recuperación de la información por la otra.
En el congreso que celebraron en Medellín (Colombia) en marzo de 2007, las Academias de la lengua española decidieron encomendar a la Real Academia Española la confección de un corpus textual que respondiese a las características actuales de la LC en todos los aspectos. Y en ese proyecto ha venido trabajando la RAE desde entonces, con el asesoramiento y la colaboración de las demás Academias de la lengua, el patrocinio de Banco Santander, la colaboración de grupos editoriales y autores de todo el mundo hispánico y la participación de equipos de codificación pertenecientes a distintas instituciones españolas y americanas, dirigidas y coordinadas por un equipo central radicado en Madrid.[13]
La configuración general del CORPES XXI consiste en la incorporación de 25 millones de formas gráficas por año, lo cual supondrá un total de 400 millones al final de la segunda fase del proyecto (años 2001 a 2016). Esos 25 millones anuales se reparten de modo que el 30 % corresponde a textos editados o producidos en España y el 70 % restante se distribuye entre todos los demás países, tomando en cuenta rasgos como la población, el volumen de su producción editorial y su integración en alguna de las áreas lingüísticas con las que las Academias han venido trabajando tradicionalmente.[14] Por ejemplo, al área constituida por México y los países centroamericanos le corresponde el 21 % del total de cada año.
Para cada país y año, los textos se distribuyen con diferentes pesos según los distintos parámetros que entran en la configuración del corpus: medio (oral / escrito), bloque (ficción / no ficción), soporte (internet / libro / miscelánea / prensa), área temática (actualidad / artes / ciencia y tecnología / ciencias sociales / política y economía / salud) y, en los textos de ficción, género (novela / teatro / relato / guion). A estas caracterizaciones, muy parecidas a las que se utilizan en el CREA, el CORPES añade, para cada texto, una indicación tipológica en función del grupo al que pertenece; así, por ejemplo, los textos de prensa son caracterizados como noticia, reportaje, entrevista, carta al director, etc. Todos esos rasgos son combinables entre sí y también, por supuesto, con zona, país, año e incluso con autor y obra si tal grado de especificación es deseable.
Así pues, el CORPES posee una configuración estable tanto en lo referente al volumen de formas para cada año como a su distribución según los diferentes parámetros tenidos en cuenta en su construcción. La estabilidad en el volumen correspondiente a cada año y su distribución interna hace que pueda ser considerado como un recurso que combina las características de los corpus

13 Los equipos externos que han colaborado en el proyecto hasta 2014 son la Academia Argentina de Letras, la Academia Puertorriqueña de la lengua española, la Fundación Comillas, la Universidad de Alcalá de Henares, la Universidad Autónoma de Barcelona, la Universidad de Salamanca, la Universidad de León, la Universidad de Santiago de Compostela y la Universidad de Valencia. Para más detalles sobre la colaboración de grupos editoriales y autores, cf. http://www.rae.es/recursos/banco-de-datos/corpes-xxi.
14 Son las siguientes: Chile, Río de la Plata, zona andina, Caribe continental, México y Centroamérica, Antillas y Estados Unidos, a las que se añaden Filipinas y Guinea Ecuatorial.

abiertos y los corpus cerrados. Es abierto en tanto que irá aumentando en 25 millones de formas por cada año transcurrido. Es cerrado en tanto que los años y los quinquenios ya completados se harán fijos y proporcionarán la estabilidad en los resultados propia de este tipo de corpus.

Como se ha indicado, la distribución interna por países, tipos de texto, bloques, etc. obedece a un reparto que se considera razonable, proporcionado a lo que se persigue habitualmente en la investigación y con unos costes elevados, pero asumibles. Es evidente que el tan discutido problema de la representatividad ha estado mal planteado. Todo corpus es una muestra extraída de una población cuyas características desconocemos, de modo que el objetivo real es que esté equilibrado (es decir, que contenga textos con un volumen suficiente para cada uno de los corpus virtuales que se puedan obtener de forma dinámica mediante la selección de valores en los diferentes parámetros de consulta). Es igualmente claro que la representatividad es un problema de gran importancia en corpus de tamaño pequeño (digamos, inferiores a diez millones de formas), que, además, habitualmente solo admiten consultas globales. Un corpus con esas características no debería dar resultados generales sesgados como consecuencia de su constitución. Pero lo que se persigue habitualmente en los corpus de referencia no es el resultado global, sino la comparación entre los resultados que arroja un cierto corpus virtual y los que se obtienen en otro. El uso de las frecuencias normalizadas permite establecer una base de comparación segura entre volúmenes de formas y textos dispares.[15]

Esta recuperación selectiva de la información es posible gracias a que todos los textos, independientemente de sus características y procedencias, han sido codificados en XML mediante un esquema común para todos ellos. La experiencia obtenida en el desarrollo del CREA y el CORDE nos ha llevado a organizar un procedimiento que, sin dejar de estar basado en las indicaciones generales de la TEI, reduce fuertemente su complejidad en todos aquellos aspectos que no son de interés para la recuperación de la información practicada para la investigación lingüística. Con esta simplificación se consigue también que el manejo y la extracción de los casos relevantes de un conjunto de cientos de millones de formas se haga con unos tiempos muy razonables a pesar de la considerable cantidad de parámetros que pueden entrar en juego en una consulta.

Los textos del CORPES han sido anotados, lematizados y desambiguados automáticamente mediante un complejo conjunto de programas desarrollados, lo mismo que la aplicación de consulta, en el departamento de informática de la RAE.[16] Es evidente el progreso que la adición de esta información supone para las búsquedas léxicas, puesto que no será necesario ya recurrir a la utilización de expresiones regulares que remeden la estructura morfológica de, por ejemplo, un verbo para obtener todas las formas vinculadas al lema (del tipo *lleg** para las formas del verbo *llegar* y similares) y, por otro lado, permitirá la recuperación correcta de los casos en los que hay formas homógrafas que deben ser vinculadas a lemas distintos (del tipo *casa*, *desarrollo* o *vino*). Sin embargo, lo más interesante de esta característica radica en el enorme avance que supone para la obtención de materiales necesarios para estudios gramaticales. En efecto, el sistema de búsqueda ha sido diseñado de modo tal que admite la petición de elementos que tienen una determinada característica gramatical con independencia del lema al que pertenezcan, por ejemplo los que han sido etiquetados como pertenecientes al futuro de subjuntivo de cualquier verbo. Además, dado que la aplicación admite la incorporación de varios elementos en la búsqueda (tanto en secuencia inmediata como en una ventana de proximidad), es posible, por ejemplo, localizar casos de un sustantivo seguido inmediatamente por dos adjetivos (del tipo *situación política actual*) o bien de un verbo cualquiera seguido de la conjunción *que* y otro verbo en modo subjuntivo. En definitiva, la estructura de la información incorporada a los textos y las formas del CORPES permite una gran

15 Para un análisis más amplio de estas cuestiones, cf. Rojo (2014, 376 y ss.).

16 El etiquetario de la versión 0.82 consta de aproximadamente 330 etiquetas.

riqueza de recuperación de fenómenos léxicos y gramaticales, siempre con la posibilidad de restringir los resultados a un cierto subconjunto del corpus (es decir, un cierto país, un tipo de texto determinado, etc.).
El CORPES permite, pues, una auténtica recuperación selectiva de la información. Como es bien sabido, la mayor riqueza de un corpus, sea cual sea su tipo, consiste en la posibilidad de establecer corpus virtuales de forma dinámica y permitir así establecer la comparación entre las características que presenta un determinado fenómeno en un cierto subconjunto (por ejemplo, noticias de prensa referentes a economía publicadas en periódicos colombianos en 2008) con las que tiene en otro (por ejemplo, de un año y un país distintos, un área temática diferente, etc.). Esta posibilidad es, por cierto, la que permite superar el viejo problema de la representatividad y su repercusión sobre los diferentes pesos que en un conjunto como el CORPES deberían tener los diferentes países o zonas, áreas temáticas, medios, etc. Además, la aplicación de consulta facilita sistemáticamente tanto la frecuencia general como la frecuencia normalizada para los datos de cada corpus virtual, lo cual hace posible realizar las comparaciones pertinentes con carácter inmediato y obtener las conclusiones oportunas.
Las búsquedas pueden hacerse, como es de esperar a partir de lo anterior, por formas (que pueden estar constituidas por varias palabras gráficas) o por lemas. Es posible también exigir la grafía original o bien tolerar el tratamiento indiferenciado de caracteres habitual en estos casos (con y sin tilde, mayúsculas y minúsculas).
Dada la complejidad originada por los numerosos parámetros con respecto a los cuales se ha caracterizado cada texto, la aplicación de consulta se organiza sobre un sistema de ventanas desplegables que van mostrando niveles a medida que se va haciendo la selección y que, por tanto, no tienen más valores que los válidos en cada uno de los parámetros en los que es posible hacer la elección. Este sistema se aplica tanto en los valores clasificatorios como en la selección de las categorías y subcategorías gramaticales. Una vez se ha optado por el rasgo «verbo» en la clase de palabras, aparecen las ventanas correspondientes a modo, tiempo, número y persona, con los valores seleccionables en cada una de ellas. Este procedimiento, más largo y pesado en su desarrollo informático, evita a quienes consulten el CORPES la necesidad de profundizar en la organización de, por citar el caso más complejo, las etiquetas gramaticales que recibe cada elemento.
Las búsquedas de ejemplos, con posibilidad de restringirlos a subconjuntos del CORPES, admiten dos grandes tipos de salida. La más general contiene la estadística de resultados, que se puede ir especificando por zonas, países, tipos de texto, etc., siempre con indicación de frecuencia general y frecuencia normalizada. La segunda proporciona las concordancias en el formato habitual, con indicación de la procedencia de cada ejemplo y la posibilidad de obtener un contexto más amplio si es necesario. Ambas salidas están interconectadas, de modo que pulsando la zona correspondiente a la frecuencia de un elemento en un cierto país se accede a los ejemplos correspondientes.
La búsqueda por formas, lemas o rasgos gramaticales admite la concatenación de cualesquiera de esos rasgos en un contexto próximo, como se ha indicado antes. Las condiciones de la búsqueda pueden referirse a un contexto de cierta longitud a cualquiera de los lados del que se utiliza como central o bien a elementos que se sitúen a una distancia determinada del primero. Así, por ejemplo, cabe plantear la búsqueda de casos del verbo *dudar* seguido inmediatamente por la preposición *de* o bien con la preposición *de* en un margen de, por ejemplo, tres elementos a la derecha (para cubrir casos del tipo *dudaba muy intensamente de su sinceridad*). Utilizando las caracterizaciones gramaticales pueden recuperarse los casos de *ir* seguido inmediatamente de *a* y luego de un verbo cualquiera en infinitivo, los de cualquier verbo seguido inmediatamente por cualquier otro en infinitivo, los de un verbo seguido a una distancia no superior a cinco elementos por una preposición, etc.

Borrador final. Pendiente de aparición en Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín: de Gruyter. En prensa (2016).

Además de la estadística y los ejemplos correspondientes al elemento seleccionado, la aplicación de búsqueda permite obtener aquellos otros elementos que coaparecen con el seleccionado en un contexto que abarca por defecto cinco elementos a cada lado. Como es bien sabido, las coapariciones[17] han ido ganando importancia en los más diversos estudios, de modo que disponer de un recurso que permita identificar las que se dan con respecto a un determinado elemento supone una importante vía de acceso a sus características léxicas y gramaticales. En la versión 0.82, la aplicación trabaja directamente con lemas (no con formas) y permite la indicación de la clase de palabras a la que pertenece. El resultado muestra los elementos que, de acuerdo con la frecuencia general y tres estadísticos distintos (información mutua, verosimilitud (*log-likelyhood simple*) y distribución *t* (*t-score*)),[18] coaparecen con el seleccionado en un cierto grado de importancia. Esos lemas llevan también la indicación de la clase de palabras a la que pertenecen, de modo que es sencillo seleccionar, por ejemplo, los adjetivos que coaparecen con un determinado sustantivo.

Por otro lado, siempre en la línea de la creación dinámica de corpus virtuales, la aplicación permite trabajar con únicamente los ejemplos de un determinado país o área temática (o ambos factores al mismo tiempo), lo cual supone una importantísima mejora en la calidad y profundidad de los datos obtenidos. Así, por ejemplo, los cinco términos que coaparecen con el índice de información mutua (MI) más alto con el lema *saco* son *terrero*, *yute*, *arpillera*, *tweed*, *amniótico*. La aparente incongruencia de estos resultados se aclara al hacer las búsquedas diferenciadas y comprobar que en América son *yute*, *tweed*, *abotonar*, *corbata* y *solapa*, mientras que en textos procedentes de España son *terrero*, *arpillera*, *cemento*, *romper* y *patata*. Es evidente que solo la recuperación diferenciada (en la que sería posible profundizar todavía más) permite entender lo que sucede con las dos grandes acepciones que tiene esta palabra en los diferentes países hispánicos. El CORPES, pues, va en este punto bastante más allá de lo que es habitual en un corpus de referencia.

De acuerdo con el diseño inicial, un 10 % de los materiales del CORPES estará constituido por transcripciones de textos orales. Por causas de diferentes tipos, el porcentaje que estos materiales suponen en la versión 0.82 (noviembre de 2015) es todavía muy inferior al previsto y procede de una única fuente original: el corpus CORALES, construido por la RAE en paralelo a la última etapa del CREA. Comprende algo menos de un millón de formas gráficas correspondientes a textos orales de diversos tipos y de todos los países hispánicos producidos entre los años 2001 y 2004. Su característica más llamativa consiste en que el texto de la transcripción está alineado con el sonido correspondiente. La información, por tanto, se recupera, como en todos los casos, a través de la versión textual, pero ofrece la posibilidad de obtener el sonido vinculado a la zona devuelta por la concordancia. Esta posibilidad, que será utilizada también en materiales de otras procedencias, abre una vía del mayor interés para estudios en los que el análisis directo del componente fónico (y no su traducción, más o menos detallada, a marcas añadidas del texto) resulta importante. Dado que, además, las búsquedas textuales incluyen ya la posibilidad de localizar signos de puntuación, el análisis de, por ejemplo, aspectos relacionados con la entonación en secuencias interrogativas, exclamativas o parentéticas se hace algo relativamente sencillo.

En la versión siguiente (la 0.83, prevista para junio de 2016), el CORPES incorporará textos procedentes del proyecto PRESEEA,[19] con cuyos responsables firmó la Academia un acuerdo de

17 A mi modo de ver, el término *colocación* no es el adecuado en español, que ha atribuido un significado diferente a las palabras procedentes de esa raíz latina. Quizá *conlocación* podría servir, pero parece mucho más razonable adoptar una expresión perfectamente reconocible para cualquier hablante de español con un significado general que resulte congruente con el que se le atribuye en lingüística.

18 Para una explicación general de las características de estos tres estadísticos puede verse la ayuda que se despliega en la página de resultados de la aplicación de consulta del CORPES-XXI.

19 En el *Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA), dirigido por Francisco Moreno, participan en la actualidad cerca de 40 equipos de todo el mundo hispánico. Para más información,

cesión de materiales en 2008. Dado que PRESEEA tiene un ámbito de actuación que comprende todo el mundo hispánico, los materiales de este proyecto enriquecerán considerablemente la variedad de la parte oral del CORPES. No obstante, en la nueva fase del proyecto (entre 2015 y 2018) será necesario dedicar una atención muy especial a la incorporación de textos orales, tarea todavía muy complicada y costosa a pesar de los importantes avances de estos años en el tratamiento de los materiales sonoros. Con un importante porcentaje de los textos orales con sonido alineado, el CORPES se situará en el camino de la integración de diferentes capas y perspectivas sobre los textos.

A partir de 2016, la aplicación de búsqueda, con las características que tiene ya en este momento (noviembre de 2015) y algunas adicionales, tendrá a su lado la posibilidad de consulta de la nómina de textos por cualquier combinación de los parámetros de configuración. Habrá también una lista de lemas y formas asociadas con sus frecuencias generales y normalizadas. Las consultas seguirán siendo realizables únicamente mediante el sistema clásico de las concordancias de longitud restringida, con posibilidad de cierta ampliación de contexto. Esta limitación, inevitable en los corpus de referencia, es compensada con creces por las ventajas de los más diversos tipos que la restricción de la longitud proporciona. Por citar únicamente la más importante, hace posible que el CORPES contenga multitud de textos de gran interés lingüístico que, por cuestiones legales, no podrían ser incluidos en condiciones diferentes de consulta y descarga. Por todo ello, el CORPES representa, en el conjunto de los corpus textuales del mundo hispánico, un recurso que va más lejos, tiene mayor volumen y es mejor que los anteriores.

4. Relación de corpus y otros recursos electrónicos mencionados en el texto

Bank of English (http://www.titania.bham.ac.uk/docs/svenguide.html).

Biblia medieval (http://www.bibliamedieval.es/index.php).

BNC: *British National Corpus* (http://www.natcorp.ox.ac.uk/).

Brown Corpus: *The Standard Corpus of Present-Day Edited American English* (http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/).

CDH: *Corpus del nuevo diccionario histórico del español* (http://www.rae.es/recursos/banco-de-datos/cdh).

CE: *Corpus del español* (http://www.corpusdelespanol.org/).

CEMC: *Corpus del español mexicano contemporáneo* (http://www.corpus.unam.mx:8080/cemc/).

CHARTA: *Corpus hispánico y americano en la red: textos antiguos* (http://www.charta.es).

COBUILD: *Collins Birmingham University International Language Database* (http://www.collins.co.uk/category/English+Language+Teaching/COBUILD+Reference).

CODEA: *Corpus de documentos españoles anteriores a 1700* (http://demos.bitext.com/codea/).

COHA: *Corpus of Historical American English* (http://corpus.byu.edu/coha/).

CORDE: *Corpus diacrónico del español* (http://rae.es/recursos/banco-de-datos/corde).

CORDIAM: *Corpus diacrónico y diatópico del español de América* (http://http://www.cordiam.org/).

CORPES: *Corpus del español del siglo XXI* (http://rae.es/recursos/banco-de-datos/corpes-xxi).

cf. http://preseea.linguas.net.

CRATER: *Corpus Resources and Terminology Extraction* (http://ucrel.lancs.ac.uk/projects.html#crater).

CREA: *Corpus de referencia del español actual* (http://rae.es/recursos/banco-de-datos/crea).

Es-Ten-Ten (http://www.sketchengine.co.uk/documentation/wiki/Corpora/esTenTen).

LOB: *The Lancaster-Oslo/Bergen Corpus* (http://www.helsinki.fi/varieng/CoRD/corpora/LOB).

PRESEEA: *Proyecto para el estudio sociolingüístico del español de España y de América* (http://preseea.linguas.net/).

5. Referencias bibliográficas

Mair, Christian, *Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora*, in: Renouf, Antoinette / Kehoe, Andrew, *The changing face of corpus linguistics*, Amsterdam, Rodopi, 2006, 355-376.

Rojo, Guillermo, *Hispanic Corpus Linguistics*, in: Lacorte, Manel (ed.): *The Routledge Handbook of Hispanic Applied Linguistics*, Nueva York, Routledge, 2014, 371-387.

Rojo, Guillermo, *Los corpus textuales del español*, in: Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*, Nueva York, Routledge. En prensa (2015).

Rojo, Guillermo / Sánchez, Mercedes, *El español en la red*, Madrid / Barcelona: Fundación Telefónica / Ariel, 2010.

Sinclair, John, *Corpus and text. Basic principles*, in: Wynne, Martin (ed.), *Developing Linguistic Corpora. A Guide to Good Practice*, Oxford, Oxbow Books, 2005, 1-16.

Kilgarrif, Adam / Renau, Irene, *EsTenTen, a Vast Web Corpus of Peninsular and American Spanish*", in *Procedia - Social and Behavioral Sciences*, 95 (2013), 12–19. Descargable de *http://www.sciencedirect.com/science/article/pii/S1877042813041372*.

Rundell, Michael: *The road to automated lexicography: An editor's viewpoint*, in: Granger, Silviane / Paquot, Magali (eds.): *Electronic Lexicography*, Oxford University Press, 2012, 15-30.