EL BANCO DE DATOS DE LA REAL ACADEMIA ESPAÑOLA: *CREA* Y *CORDE*

MERCEDES SÁNCHEZ SÁNCHEZ
CARLOS DOMÍNGUEZ CINTAS
Departamento de Banco de Datos
Real Academia Española

El Banco de Datos de la Real Academia Española está constituido por dos grandes corpus: el *Corpus de Referencia del Español Actual* (*CREA*) y el *Corpus Diacrónico del Español* (*CORDE*). Ambos conjuntos son complementarios, de manera que el *CREA* contiene textos pertenecientes a los últimos treinta años de historia del español, mientras que el *CORDE* se ocupa de todo lo demás. El carácter integrado de ambos corpus se refleja en la previsión de que los textos pertenecientes a períodos que, por el paso del tiempo, vayan quedando fuera del ámbito de *CREA* pasarán a formar parte de *CORDE*.

Los corpus que integran el Banco de Datos constituyen la principal fuente de documentación de la que se sirve el Instituto de Lexicografía de la Real Academia Española para preparar los materiales que se discuten en Comisión y argumentar, de este modo, las propuestas. Entre otras fuentes, el Banco de Datos ha servido para la redacción de los últimos trabajos de la Real Academia Española: tanto el *Diccionario del Estudiante* como el *Diccionario Panhispánico de Dudas*, o más recientemente, el *Diccionario Esencial*, o la futura *Nueva Gramática de la Lengua Española,* incorporan ejemplos de uso real de la lengua española extraídos del Banco de Datos, principalmente del *CREA*.

Un corpus es un conjunto de textos seleccionados y ordenados de acuerdo con una serie de criterios lingüísticos explícitos, de modo que pueda utilizarse como muestra de una lengua. Hace ahora unos cuarenta años comenzó el desarrollo de la lingüística de corpus, en oposición a la aproximación racionalista basada en la introspección. En los años 60, los primeros corpus, los corpus de primera generación, tenían un tamaño aproximado de un millón de palabras. Los textos que lo formaban se introducían manualmente. Hacia los ochenta se desarrollaron corpus de segunda generación: el COBUILD o el Longman-Lancaster ya tenían muchos más millones de palabras y además ya se introducían los textos a través de escáneres y con programas de Reconocimiento Óptico de Caracteres; en la actualidad se construyen y explotan corpus como el Bank of English o el British Nacional Corpus, formados por cientos de millones de palabras.

Un corpus puede entenderse como una biblioteca organizada, ordenada por materias, por zonas geográficas, por fecha, en cuyos textos podemos encontrar la forma buscada perfectamente documentada: es la ficha de trabajo, la papeleta léxica, en la que a la palabra acompaña, por completo, su referencia bibliográfica. Los textos que forman parte de esa *biblioteca* se seleccionan siguiendo una serie de criterios, mediante la distribución equilibrada entre España y América; entre lengua escrita y lengua oral, entre libros y prensa, novelas y libros de ciencia… Todo ello permite diseñar un programa informático capaz de realizar búsquedas seleccionando la información.

En la *Historia de la Real Academia Española,* don Alonso Zamora Vicente se refiere a los corpus y sus aplicaciones con estas palabras:

> El banco de datos de la Academia es, hoy por hoy, la única herramienta lingüística de gran magnitud existente para nuestra lengua. No cabe duda, por tanto, de que habrá de ser el punto de partida forzoso para investigaciones de diverso tipo, aquellas estrictamente léxicas, pero también de campos tan dispares como el de la publicidad, la terminología, la sociolingüística, etc., así como para la elaboración de una enorme cantidad de productos derivados: gramáticas, diccionarios de todo tipo, métodos y sistemas de didáctica del español, desarrollos informáticos de traducción automática, diccionarios electrónicos, tesauros y correctores ortográficos integrados en procesadores de textos, etc.[1]

En la actualidad, *CORDE* cuenta con unos 300 millones de formas y el *CREA* consta de unos 155 millones, de manera que los investigadores tienen a su disposición alrededor de 450 millones de formas de todos los períodos del español, lo que constituye, sin duda, el recurso más importante del que se haya podido disponer jamás para el estudio de esta lengua. Se encuentra disponible a través de la página electrónica de la Real Academia Española: http://www.rae.es.

Todos los textos que integran el Banco de Datos van acompañados de una serie de etiquetas declarativas de algunos elementos estructurales de los documentos —codificación— y etiquetas analíticas de aspectos lingüísticos.

LA NÓMINA DE AUTORES Y OBRAS DEL BANCO DE DATOS

Los textos están perfectamente documentados en la cabecera electrónica que acompaña a cada uno de ellos: es la ficha bibliográfica del texto en papel (autor, título, lugar de publicación, editorial…) pero también la ficha electrónica: número de palabras, clasificación geográfica, temática. Gracias a estas referencias y a través de una aplicación específica, la *Nómina de autores y obras*, obtenemos los datos reales sobre los que se realiza cada consulta a *CORDE* o a *CREA*: números de textos totales, distribución

1 Alonso Zamora Vicente, *Historia de la Real Academia Española*, Madrid: Espasa-Calpe, 1999.

geográfica, cronológica o temática. En la aplicación se muestra también la referencia bibliográfica completa, el criterio que ha servido para su clasificación cronológica en el Banco de Datos, la clasificación temática y el número de palabras de cada uno de los textos.

Además de mostrar las estadísticas generales, la nómina permite combinar distintos criterios de selección para obtener los datos estadísticos de la consulta realizada: número de palabras de España, de América; en los distintos temas, épocas, etc.

Las consultas a la nómina pueden realizarse en tres ámbitos:

1. Consulta al Banco de Datos: obtenemos datos correspondientes a aquellos criterios de clasificación de textos compartidos por *CREA* y *CORDE*: geográfico y medio (por ejemplo, seleccionar toda la prensa argentina, o únicamente libros chilenos), cronológico (solo registros del siglo XX, desde 1935 a 1986, por ejemplo). También podremos obtener datos correspondientes a autores representados en los dos corpus (Vargas Llosa), etc.

2. Consulta en *CREA*: permite obtener datos relativos a los textos desde 1975 hasta 2004. La tipología de los textos que componen *CREA* y *CORDE* es diferente, tanto temática como porcentualmente. En *CREA* pueden obtenerse datos estadísticos, por ejemplo, distribuidos al 50% entre España y América referentes a textos periodísticos clasificados en política, o los registros procedentes de correos electrónicos, o, incluso, de *weblogs,* cuadernos de bitácora de Internet que representan el material más reciente del que da cuenta *CREA*.

3. Consulta en *CORDE*: para operar únicamente con los datos de *CORDE,* es decir, textos hasta 1975, así como datos correspondientes a aquellos criterios de selección de textos que no comparte con *CREA*: geográficamente, por ejemplo, textos producidos en Filipinas o aquellos escritos en judeoespañol. Además, en *CORDE* se han regularizado las grafías en los títulos de los textos, lo cual facilita la búsqueda en la aplicación, pero se mantiene el título original de la obra en el resultado de la consulta. Es decir, se buscará, *Crónica…* o *Teatro…*, pero se obtendrán los títulos *Chronica…*, *Theatro…*, etc.

CREA: CORPUS DE REFERENCIA DEL ESPAÑOL ACTUAL

El *CREA* recoge las versiones electrónicas de textos producidos entre 1975 y 2004; es, por tanto, un corpus electrónico de carácter sincrónico que sirve para mostrar ejemplos del uso real de la lengua española de los últimos años, tanto en España como en América. Está formado por una amplia variedad de textos escritos y orales con gran diversidad temática. En *CREA* «conviven» un temario de oposiciones a enfermería español, una obra sobre astrología, procedente de Uruguay, un manual de Informática mexicano, la transcripción de la retransmisión de un partido de fútbol argentino o la última novela de Gabriel García Márquez.

Características generales de CREA

Puesto que *CREA* pretende ser representativo del estado actual de la lengua, los materiales que lo integran han sido seleccionados de acuerdo con los siguientes parámetros:

• *Medio*: el 90% corresponde a la lengua escrita y el 10% a la lengua oral. De ese 90%, un 49% son libros, otro 49% es prensa y el 2% restante recoge los textos denominados «miscelánea»: folletos, prospectos, correos electrónicos, blogs…

• *Cronológico*: en períodos de cinco años: 1975–1979; 1980–1984; 1985–1989; 1990–1994; 1995–1999; 2000–2004. *CREA* siempre abarcará los últimos 25 años de la lengua, de manera que, al finalizar la fase 2000–2004, los textos correspondientes a los primeros cinco años pasarán a formar parte del *CORDE*. Los textos de *CREA* contienen un código específico para su correcta inclusión en el parámetro temático correspondiente en *CORDE*.

• *Geográfico*: los materiales se reparten al 50% entre España y América. A su vez, el 50% americano se distribuye en las zonas lingüísticas tradicionales: caribeña, mexicana, central, andina, chilena, y rioplatense.

• *Temático*: cada uno de los tres grandes grupos de materiales («libros y prensa», «miscelánea» y «oral») se clasifica de modo independiente.

Los textos de «libros y prensa», en dos grandes grupos:

I. Ficción: novela, relatos y teatro.

II. No ficción, con seis hipercampos —«Ciencias y tecnología», «Ciencias sociales, creencias y pensamiento», «Política, economía, comercio y finanzas», «Artes», «Ocio y vida cotidiana» y «Salud»— que distribuyen hasta 20 áreas temáticas.

Los textos de «miscelánea» se clasifican en impresa/no impresa y las transcripciones orales en dos grandes géneros con sus correspondientes subgéneros: radiofónico o televisivo es el género I, que incluye noticias, reportajes, entrevistas, debates, tertulias, documentales, retransmisiones deportivas, magacines, revistas deportivas, variedades, sorteos y concursos. En el segundo género, otras grabaciones, se introducen transcripciones de discursos, clases, mesas redondas, etc.

De los 155 millones de formas que integran el *CREA*, actualmente unos diez millones pertenecen a la parte oral. La incorporación de estos diez millones de formas se distribuye en dos grupos completamente diferentes:

a) Textos producidos entre 1975–1999: más de ocho millones y medio de registros procedentes, en su mayor parte, de transcripciones de radio y televisión. Pueden consultarse en la página electrónica de la Real Academia Española, a través de la aplicación de concordancias, la misma que *CREA*.

b) Textos producidos entre 2000–2004: a través de un programa de transcripción, codificación y alineamiento de texto/sonido, se incorporaron a la parte oral del *CREA* más de un millón de palabras transcritas de las que es posible recuperar, además, el fragmento sonoro. Actualmente se trabaja en una aplicación que permita la explotación del corpus oral alineado.

Usos y aplicaciones del CREA

Los corpus significan, ya lo hemos dicho, un recurso de primer orden para la investigación lingüística y lexicográfica; pero no solo para las tareas que se llevan a cabo en la Real Academia Española: la consulta al corpus está disponible, a través de la página electrónica de la RAE, para quienes quieran encontrar ejemplos y testimonios de uso de palabras o expresiones, distribución de las mismas o presencia/ausencia en España y América.

La búsqueda puede ser simple: en la línea de consulta escribimos, por ejemplo *rap*, sin ninguna restricción por tema, lugar geográfico o medio. El sistema nos devuelve un total de 241 casos en 132 documentos. La estadística nos permite observar que se documenta por primera vez en 1981, con un solo caso; en 1995 aparecen 44 casos y 19 ya en 2004. Cabe preguntarse sobre qué número de registros hemos realizado nuestra consulta. Acudimos, para ello, a la nómina y vemos las estadísticas generales. Queremos, además, saber cuántas palabras hay en el corpus correspondientes al año 1995. En la aplicación de la nómina escribimos «1995» en el criterio cronológico, es decir, restringimos el número de registros por año. La estadística nos mostrará, también, el medio del que proceden los registros. Con todo ello podremos determinar la representatividad de *rap* y, después, en la aplicación de concordancias, seleccionar ejemplos de uso.

La aplicación de nómina nos indica que se recogen 10.016.195 registros del año 1995 en el *CREA*; que unos dos millones proceden de América y el resto, de España, y que provienen casi al 50% de textos de libros y prensa.

Por su parte, la aplicación de concordancias nos muestra ejemplos de uso; ahora podemos restringir la búsqueda a libros de España de 1995, pero podríamos seleccionar prensa argentina o libros de música de Perú: eso lo haremos en los criterios de selección. Recuperamos y seleccionamos, por ejemplo, esta concordancia que documenta el uso de *rap* en libros procedentes de España y producidos en 1995:

> Así que uno espera que los nuevos rebeldes hagan algo, muevan el cotarro, les den marcha a los de arriba, como tú hiciste en tu momento. Y sí, lo hacen, porque el *rap* no deja de ser una música de combate, el grito desesperado del negro que se muere en el gueto, pero..., entonces va la industria y mete al rap en la discoteca, y salen hasta raperos blancos descafeinados. (Jordi Sierra i Fabra, *El regreso de Johnny Pickup*, Madrid: Espasa-Calpe, 1995).

Los corpus también pueden servir como banco de pruebas de hipótesis lingüísticas; pueden realizarse búsquedas complejas y utilizar para ello operadores lógicos: por ejemplo, la consulta «computadora o ordenador» nos devolverá todos los documentos que contengan computadora u ordenador y comprobaremos el uso mayoritario de «computadora» en América y «ordenador» en España.

CORDE: CORPUS DIACRÓNICO DEL ESPAÑOL

Uno de los problemas fundamentales con el que se ha encontrado la Real Academia Española en los sucesivos intentos que ha realizado para la elaboración de un diccionario histórico han sido los materiales que le han servido de base para la redacción. Los que se utilizaban hasta ahora, 13 millones de fichas aproximadamente recopiladas a lo largo de tres siglos, eran muy importantes pero imperfectos: en ocasiones, escasos

para la documentación de determinadas formas, en otras, defectuosos (malas ediciones, fichas incompletas) por lo que este material tenía que someterse a distintos tipos de control.

Con el fin de solucionar definitivamente este problema, la Academia ha construido el *Corpus diacrónico del español* (*CORDE*), que cuenta en este momento con más de trescientos millones de registros léxicos desde los orígenes del español hasta 1974, complementados con los más de ciento cincuenta millones que, procedentes de los últimos treinta años, integran el *CREA*. Amparándose en estas bases lexicográficas la Real Academia podrá afrontar definitivamente la puesta en marcha del *Nuevo diccionario histórico de la lengua española*.

Características generales de CORDE

El *Corpus diacrónico del español* pretende ser un corpus representativo de la lengua española a lo largo de la historia. Contiene diferentes variedades de textos distribuidos cuantitativamente de forma proporcional para cada grupo establecido, en un intento de recoger testimonios de todas la épocas y lugares en que se habló español, desde los inicios del idioma hasta nuestros días.

El catálogo de obras que lo componen comprende 5500 títulos, de los que 1000 son de autor anónimo y el resto corresponde a 4500 autores conocidos. Todo ello supone, como ya hemos dicho, un total de más de 300 millones de registros.

Las obras se han seleccionado siguiendo unos criterios muy concretos. En primer lugar se introducen los textos en su integridad, incluyendo prólogos, aprobaciones, tasas... Se han elegido aquellos considerados más representativos debido a su difusión, a su influencia en obras posteriores o al uso que de ellos se ha hecho como apoyo de autoridad en otras obras. Se prefieren las ediciones críticas o, en su defecto, buenas ediciones anotadas, buenas transcripciones de un testimonio, aunque en determinadas ocasiones no ha quedado más remedio que seleccionar la única edición existente. Otro de los aspectos en el que se ha puesto especial cuidado ha sido la dimensión lingüística del texto, la riqueza de vocabulario y su carácter divulgativo. Están, pues, en el *CORDE*, las grandes obras de la lengua española junto a otras que no se han vuelto a editar desde su primera edición y que ahora se han digitalizado para el corpus.

Es importante advertir que el hecho de que la documentación del corpus sea escrita no significa que sea exclusiva ni fundamentalmente literaria. Si se trata de recoger la lengua general, hay que pensar que la literatura no es más que una parcela y no necesariamente la más importante. Por tanto el *CORDE* ha incluido la literatura como uno más entre los apartados de la lengua escrita, pero sin desdeñar los documentos o los textos científicos.

Sin perder de vista el aspecto temático, el *CORDE* tiene una estructuración basada fundamentalmente en criterios formales.

En primer lugar se da entrada al verso. Aunque conscientes de que el lenguaje poético es poco significativo para los estudios lingüísticos y por tanto tienen muy poca

presencia en los corpus sincrónicos, el período cronológico que abarca este corpus ha obligado a su inclusión, ya que sin él se perdería gran parte de la literatura medieval y clásica. El peso de los textos en verso se puede ver en la clasificación genérico-temática del corpus:

- Literarios (44%)
  - Verso 10%
  - Prosa narrativa 27%
  - Prosa dramática 7%

- No literarios (56%)
  - Didáctica 10%
  - Ciencia y Técnica 14%
  - Religión 6%
  - Sociedad 8%
  - Historia 9%
  - Jurídica 6%
  - Prensa 3%

La distribución de los textos desde el punto de vista geográfico concede un 74% al español de España frente a un 25% para el de América. Se justifica esta desproporción en cuestiones históricas evidentes. El 1% restante se asigna a textos judeoespañoles. Para la distribución de los textos americanos se han establecido una serie de zonas basadas fundamentalmente en los estudios de Henríquez Ureña. Así, se consideran los siguientes grupos:

- América 1: México, Guatemala, Honduras, El Salvador.
- América 2: Nicaragua, Costa Rica.
- América 3: Cuba, Puerto Rico, Panamá, República Dominicana, Venezuela.
- América 4: Colombia, Ecuador, Perú y Bolivia.
- América 5: Chile.
- América 6: Argentina, Uruguay, Paraguay.

Desde el punto de vista cronológico el *CORDE* está dividido en grandes períodos temporales. Estos grandes grupos cronológicos se subdividen a su vez en tramos menores establecidos según criterios histórico-lingüísticos que permiten obtener, de forma pormenorizada, una mejor visión histórica de la constitución y desarrollo de la lengua española. Aunque el corpus pretende ser representativo de todas las épocas no quiere decir que todos los períodos estén constituidos por la misma cantidad de textos. Se observará que el porcentaje aumenta progresivamente según nos acercamos a la época contemporánea.

1. EDAD MEDIA (16,5%)
   a. Orígenes–1250
   b. 1251–1491

2. SIGLOS DE ORO (30,5%)
   a. 1492–1598
   b. 1599–1712

3. ÉPOCA CONTEMPORÁNEA (53%)
   a. 1713–1812
   b. 1813–1898
   c. 1899–1939
   d. 1940–1974

Todos estos elementos clasificatorios quedan recogidos en la cabecera bibliográfica que acompaña a cada texto y que sirve además para el control y recuperación de los datos, haciendo posible que las consultas vayan referidas a la totalidad de los textos o bien únicamente a aquellos que poseen unas determinadas características geográficas, temporales, temáticas, etc.

Usos y aplicaciones del CORDE

Una consulta básica al *CORDE* nos proporciona información sobre el nacimiento, vigencia y desaparición de una palabra, además de que nos permite deducir su significado por los contextos en los que aparece. Por ejemplo:

La consulta sobre la palabra *bahareros*, sin ningún tipo de restricción (no hemos limitado zona geográfica, cronología, tema, autores…) arroja la siguiente información:

• 1330–1343 Ruiz, Juan (Arcipreste de Hita), *Libro de buen amor*, ed. Alberto Blecua, Cátedra (Madrid), 1992, página 315:

> «dexa todos aquéstos, toma de nós serviçio.»
> Las monjas le dixieron: «Señor, non avrias viçio:
> son pobres *bahareros* de mucho mal bolliçio;
> señor, vete connusco, prueva nuestro celiçio.»

• *a* 1492 Anónimo, *Cancionero de Pero Guillén*, ed. Brian Dutton, Universidad de Salamanca (Salamanca), 1990, fol. 617v:

> Sabiendo quanto temida
> mi pluma tiene'l ropero,
> c'os mató y os dio herida
> con que vos morís en vida,
> don cobarde *baharero*.

Por tanto, podríamos concluir que *baharero* parece significar «fanfarrón, embustero»; es una palabra que nace en el siglo XIV y llega hasta finales del XV, época en la que desaparece para quedar tan solo en algunos diccionarios u obras de carácter lexicográfico.

En fin, *CREA* y *CORDE* conforman el mayor banco de registros lingüísticos del español reunidos y disponibles para todos; el departamento de Banco de Datos cuenta con una dirección de correo electrónico, dbd@rae.es, desde la que, con sumo gusto, atendemos dudas, cuestiones y sugerencias relacionadas con la explotación de los corpus de la Real Academia Española.

MATERIALES