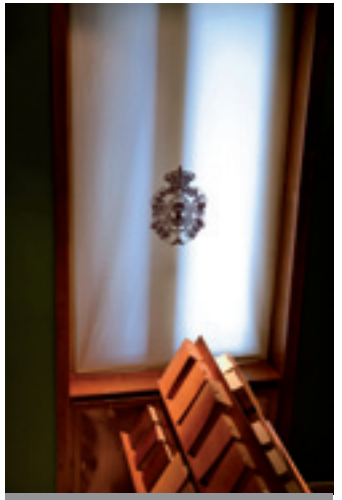
REAL ACADEMIA ESPAÑOLA

El Corpus del Español del Siglo xxi (CORPES)

Primera fase (2001-2012)

Antecedentes: el CREA y el CORDE

La decisión de poner en marcha la confección del Corpus de Referencia del Español Actual (CREA), adoptada por la Real Academia Española en 1995 y seguida pocos meses después por el proyecto de construcción del Corpus Diacrónico del Español (CORDE), ha producido una modificación radical en los medios que tienen a su disposición los equipos técnicos de la Academia, así como de cuantas integran la Asociación de Academias de la Lengua Española y, en general, de quienes se dedican a la investigación del español o a la producción de recursos y materiales para esta lengua. En efecto, las obras publicadas por las academias desde ese momento (la vigésima segunda edición del *DRAE*, el *Diccionario panhispánico de dudas*, el *Diccionario del estudiante*, el *Diccionario esencial*, la *Nueva gramática de la lengua española* en sus tres versiones y la *Ortografía*), así como las que se encuentran actualmente en fase de redacción (la próxima edición del *DRAE*), se han beneficiado de los datos contenidos en el CORDE y, sobre todo, en el CREA.



El CORPES XXI facilitará la elaboración de los diccionarios.

En este conjunto de textos que reflejan la historia completa del español, procedentes de todos los países de habla hispana y de los más diversos tipos y géneros, se encuentran los materiales relevantes que los lexicógrafos y los gramáticos necesitan para llevar a cabo su trabajo. La unión de ambos corpus (más de 260 millones de formas desde los orígenes del idioma hasta 1974 en el CORDE y algo más de 155 millones de formas desde 1975 hasta 2004 en el CREA) proporciona a todos los investigadores o simples interesados en nuestra lengua un recurso en el que poder documentar con comodidad, rapidez y seguridad la mayor o menor frecuencia con que se utiliza una palabra, su distribución por países, años, tipos de texto y áreas temáticas, las formas que figuran habitualmente en su contexto inmediato, etc. En definitiva, estos dos corpus contienen cuanto se necesita para trabajar sobre bases sólidas, tanto en la línea estrictamente científica como en la que fundamenta la toma de decisiones normativas que la Asociación de Academias tiene a su cargo en todo el mundo hispánico.

El Corpus del Español del Siglo XXI

El diseño del CREA, científicamente adecuado y muy ambicioso cuando se formuló, se ha quedado corto para las necesidades actuales. En su concepción original, el período de cinco años en que se articula, contenía, en los tramos más recientes, 37,5 millones de formas, esto es, 7,5 millones de formas para cada año, distribuidas por tipos de textos, soportes, áreas temáticas, etc. y repartidas al 50 % entre España y América. Estas cifras resultan ahora claramente insuficientes para basar en ellas la enorme cantidad de decisiones que las academias han de tomar para llevar a cabo la próxima edición del *DRAE*, prevista para 2014, y, por supuesto, para las necesidades de la investigación lingüística.

De otra parte, la versión pública no incorpora anotación morfosintáctica ni lematización, la tipología textual se ha quedado desfasada como consecuencia de los cambios que han tenido lugar en los últimos años y, finalmente, la aplicación de consulta, concebida y

Sede de la RAE

El 70 % de las formas seleccionadas proceden de América.

desarrollada hace quince años, está necesitada de adaptación a las posibilidades actuales de la informática.

Conscientes de todo ello, y también de la importancia decisiva que tienen estos recursos, en el congreso celebrado en Medellín en marzo de 2007, las academias de la lengua acordaron encomendar a la Real Academia Española la construcción del Corpus del Español del Siglo XXI (CORPES). La RAE ha estado trabajando desde entonces para cumplir el encargo

- con el asesoramiento y la colaboración de las academias de la lengua española;
- con el patrocinio de Banco Santander;
- con la colaboración de grupos editoriales y autores;
- y con la participación de equipos de codificación pertenecientes a diferentes instituciones españolas y americanas.

Las características del CORPES son las siguientes:

- **Tamaño**: 25 millones de formas por año (por tanto, 300 millones en su primera fase, de 2001 a 2012).
- **Distribución**: 70 % América y 30 % España.
- **Codificación** y tipología textual muy detalladas.
- **Anotación** morfosintáctica y lematización.
- **Aplicación de consulta** muy flexible y apta para lograr una auténtica recuperación selectiva de la información.

Cada uno de los textos integrados en el CORPES (algo más de 200 000 en la versión actual, la 0.5) ha sido codificado de forma que incorpora todos los datos bibliográficos pertinentes (autor, título, editorial, fecha, etc.) y aquellos que lo sitúan en cada uno de los subconjuntos en que puede ser analizado el corpus: zona, país, procedencia (libro, prensa, internet, oral), área temática (actualidad, economía, deportes, etc.) y tipo de texto (académico, noticia, biografía, etc.). El sistema de codificación, en XML y basado en la *Text Encoding Initiative*, ha sido desarrollado íntegramente en la RAE. Cada uno de estos parámetros de selección puede ser combinado con todos los demás, de modo que es posible obtener los casos de una palabra o expresión procedentes de, por ejemplo, noticias de prensa publicadas en Panamá en 2010 y que traten de economía. La aplicación de consulta permite siempre la opción de obtener directamente los ejemplos o bien trabajar primero con los datos estadísticos generales y luego ir seleccionando los subconjuntos que interesan. En todos los casos, la aplicación facilita la frecuencia general y la normalizada (en número de casos por millón de formas).

El CORPES XXI es un proyecto panhispánico.

Por otro lado, todos los textos han sido sometidos a una batería de programas informáticos que, mediante técnicas de lingüística computacional, añaden a cada forma la caracterización gramatical que le corresponde y el lema al que se adscriben. Por tanto, una forma como *llegaremos* lleva la indicación de que se trata de la primera persona del plural del futuro de indicativo del verbo *llegar*; a *mirándola* se asocia la información de que debe ser analizada en el gerundio del verbo *mirar* más el pronombre personal átono femenino de tercera persona del singular y acusativo; *a punto de* es una locución; *Canal de Panamá* un nombre propio, etc. Ello significa que la aplicación de consulta permitirá hacer recuperaciones basadas no solo en aspectos léxicos, sino también en factores exclusivamente gramaticales (un sustantivo cualquiera seguido de dos adjetivos cualesquiera, todos los casos de verbos en primera persona de plural del futuro de indicativo, etc.).

Por último, el CORPES permite obtener las coapariciones de una palabra, esto es, el conjunto de términos que aparecen en su contexto inmediato bien en la totalidad del corpus, bien en alguno de los subconjuntos delimitados libremente por quien hace la consulta.

La selección de textos, el diseño de la codificación, la coordinación del trabajo de los equipos colaboradores y la revisión de los resultados han sido realizados por un equipo central radicado en la RAE. Tanto la anotación y lematización como la aplicación de consulta han sido desarrolladas en el Departamento de Tecnología de la RAE. La codificación de los textos ha sido incorporada mayoritariamente por equipos colaboradores radicados en distintas instituciones americanas y españolas.

Situación actual del CORPES y perspectivas de futuro

A quince meses de la finalización de la primera fase del proyecto, que recoge textos producidos entre 2001 y 2012, la versión provisional 0.5, que se presenta en el VI Congreso Internacional de la Lengua Española, consta de algo más de 200 000 textos, que contienen unos 165 millones de formas ortográficas. Lo mismo que sucedió ya con el CREA y el CORDE, está previsto que el CORPES pueda ser consultado en versiones provisionales. En concreto, la planificación actual prevé la disponibilidad pública del CORPES en la página electrónica de la RAE antes de que finalice el año 2013.

Dado que el CORPES ha sido concebido como un corpus semiabierto, una vez terminada la primera fase (prevista para diciembre de 2014), se irá incrementando a un ritmo de 25 millones de formas por año. Con ello, todas las personas interesadas en la lengua española dispondrán de un corpus confiable en el que podrán encontrar respuesta a sus preguntas acerca del español del siglo XXI, basar sus investigaciones o bien obtener los datos necesarios para construir las herramientas y recursos lingüísticos que nuestra lengua necesita para desenvolverse con seguridad en la sociedad del conocimiento.

ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA

Con el patrocinio de

C/ Felipe IV, 4
28014 (28071) Madrid

34 91 420 14 78

rae.es | fprorae.es | asale.org
Twitter.com/raeinforma
YouTube/raeinforma
Facebook.com/RAE