REAL ACADEMIA ESPAÑOLA

Santander 150 AÑOS

El Corpus del Español del Siglo XXI

1. Antecedentes: el CREA y el CORDE

Con la perspectiva que proporcionan los más de diez años transcurridos desde su gestación, es hoy evidente que la puesta en marcha del CREA (Corpus de Referencia del Español Actual) en 1995 y, pocos meses después, del CORDE (Corpus Diacrónico del Español) supuso una modificación radical en el sistema de trabajo de los equipos técnicos de la Real Academia Española, de todas las que integran la Asociación de Academias de la Lengua Española y, en general, de cuantos se dedican a la investigación del español. En efecto, todas las obras publicadas por las Academias desde ese momento (la vigésima segunda edición del *DRAE*, el *Diccionario panhispánico de dudas*, el *Diccionario del estudiante* y, sobre todo, el *Diccionario esencial*), así como las que se encuentran actualmente en fase de preparción (la próxima edición del *DRAE*, la *Nueva gramática de la lengua española* y el *Diccionario académico de americanismos*), se han beneficiado de los datos contenidos en el CORDE y, sobre todo, el CREA. En algunos casos (el *DPD* o el *Diccionario esencial*), el contenido de las entradas se basa precisamente en la información contenida en los corpus.

En este amplísimo conjunto de textos de todas las épocas del español, procedentes de todos los países de habla hispana y de los más diversos tipos, se encuentran los materiales relevantes que los lexicógrafos y los gramáticos necesitan para llevar a cabo su trabajo. La unión de ambos corpus (cerca de 300 millones de formas desde los orígenes del idioma hasta 1974 en el CORDE y algo más de 150 millones de formas desde 1975 hasta la actualidad en el CREA) proporciona a todos los investigadores o, en general, a los interesados en nuestra lengua el recurso para 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, áreas temáticas,

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 de la Asociación de Academias para todo el mundo hispánico.

2. El Corpus del Español del Siglo XXI

Sin embargo, 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 léxicas, 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 insuficientes para fundamentar en ellas la enorme cantidad de decisiones que las Academias han de tomar para llevar a cabo la actualización permanente del *DRAE*. Consciente de ello, la Real Academia Española propuso en el Congreso de Academias celebrado en Medellín el pasado mes de marzo, la preparación de un nuevo corpus del español, elaborado con 25 millones de formas léxicas para cada uno de los años comprendidos entre 2000 y 2011. El Corpus del Español del Siglo XXI constará pues, en su primera fase, de 300 millones de formas, cifra que garantiza que las nuevas ediciones del *DRAE* tendrán el fundamento empírico exigible. La propuesta fue aprobada por unanimidad por el Pleno del Congreso y todas las Academias se comprometieron prestar la máxima colaboración.

De otra parte, la distribución de la procedencia de los textos establecida para el CREA (50% de España y 50% de América), única viable en el momento en que se gestó, resulta inadecuada para los nuevos objetivos. Teniendo en cuenta los muy diversos parámetros que es necesario tomar en cuenta, la repartición propuesta es el 30% para España y el 70% para América.

3. Realización del proyecto

Es claro que el diseño, control, adquisición y codificación de textos y la construcción de las herramientas informáticas que permitan la consulta de un corpus como el que aquí se describe desde cualquier lugar del mundo y con cualquier tipo de computadora requiere una inversión considerable y el trabajo de un grupo de personas muy amplio. La RAE tiene experiencia en la gestión de proyectos de este tipo, derivada de la ampliación del CORDE de 75 a 300 millones de formas en poco más de dos años.

El sistema utilizado, que es, en sus líneas básicas, el mismo que se propone para la construcción del Corpus del Español del Siglo XXI consiste en la división del trabajo entre un equipo central, situado en la RAE, y un número relativamente amplio de equipos colaboradores, vinculados a diversas universidades de España y América. El equipo central, de tamaño relativamente reducido (entre 5 y 8 personas), constituido por personal de la Real Academia Española, se ocupará del diseño, selección de materiales, establecimiento del sistema de codificación, preparación técnica del personal de los equipos colaboradores, control del trabajo realizado por ellos, inserción de los textos codificados en el sistema y mantenimiento de las aplicaciones informáticas que permiten las consultas.

Los equipos colaboradores, radicados en diferentes universidades españolas y americanas, dirigidos por un investigador perteneciente a ellas, podrán especializarse en la codificación de textos de una cierta zona lingüística o de unas determinadas características (prensa, textos orales, etc.). Su vinculación al proyecto se establecerá mediante convenios con la RAE, que se comprometerá a subvencionar a cada equipo en función de la carga de trabajo prevista para cada uno de ellos y, por supuesto, a la formación inicial del personal que lo componga.

El sistema de trabajo, probado ya en la última fase del CORDE, permite, tras el período de puesta en marcha y preparación del personal, un ritmo rápido y constante de crecimiento de los materiales. Presenta, además, la ventaja adicional de que hace posible mejorar la preparación de un buen número de personas en un grupo amplio de

universidades, difundir una tecnología novedosa de acceso a los materiales lingüísticos y homogeneizar los procedimientos de codificación de textos.

Por todo ello, el Corpus del Español del Siglo XXI constituye un proyecto que conjuga a la perfección los objetivos comunes que tienen la Real Academia Española, la Asociación de Academias de la Lengua española y el Santander: la contribución al mejor y más amplio conocimiento del español actual en toda su extensión y complejidad, el empleo de tecnologías punteras en la selección, integración y explotación de los materiales, la voluntad de poner los materiales resultantes a disposición de todos los interesados y el planteamiento del trabajo mediante una estructura de funcionamiento que requiere la colaboración de un número amplio de universidades y centros de investigación de todo el mundo hispánico.

El Santander participa en la iniciativa de la mano de su División Global Santander Universidades, cuyas actividades vertebran la acción social del banco y le permiten mantener una alianza estable con el ámbito académico. A través de su División Global Santander Universidades, el Santander ha destinado 400 millones de euros entre 1996 y 2006 al patrocinio de proyectos académicos, de investigación y tecnológicos en apoyo a la educación superior en los países en que está presente, y ayuda a la difusión y la enseñanza del español mediante su colaboración en iniciativas como Universia –la mayor red de universidades del mundo-, la Biblioteca Virtual Miguel de Cervantes –la mayor colección en internet de las letras hispánicas-, y la Fundación Campus Comillas, un proyecto emblemático del Gobierno de Cantabria cuya línea de trabajo encaja perfectamente en la orientación del proyecto de elaboración del Corpus Español del Siglo XXI.