El CORPES XXI:
la herramienta de consulta más actual de la RAE y la ASALE para traducir, editar y corregir textos de manera más productiva y congruente

En este artículo, la Comisión de Idioma Español introduce el CORPES XXI (Corpus del Español del Siglo XXI), que constituye la herramienta de consulta más actualizada de la Real Academia Española (RAE) y la Asociación de Academias de la Lengua Española (ASALE): no solo cuenta ya con una versión 1.3, sino que lleva incluidas más de 400 millones de formas ortográficas.
A continuación, compartimos la entrevista a la doctora Mercedes Sánchez Sánchez, responsable técnica del corpus.

| Por la traductora pública Graciela B. Forte, integrante de la Comisión de Idioma Español

Mercedes, en primer lugar, muchas gracias por recibirnos y por aceptar esta propuesta de la Comisión de Idioma Español del CTPCBA.

Muchas gracias a vosotros por vuestro interés en el Corpus del Español del Siglo XXI.

Antes que nada, nos gustaría saber cómo fueron tus comienzos en el ámbito académico, es decir, qué fue lo que te llevó a estudiar Filosofía y Letras y, posteriormente, a realizar un doctorado con especialización en Filología Hispánica.

Bueno, siempre tuve interés por la lengua, la historia y la literatura españolas, no de manera aislada, como se estudia normalmente, sino relacionándolas entre sí, de modo que descubrí que mi verdadera vocación era la filología, entendida en el sentido que recoge el *Diccionario*: el estudio de la cultura a través de su lengua y su literatura, en especial mediante los textos escritos.

Estudié Filología Hispánica en la Universidad Autónoma de Madrid. Durante la carrera tuve como profesor a Pablo Jauralde, especialista en el Siglo de Oro y, muy particularmente, en Francisco de Quevedo. Promovía actividades encaminadas a la formación integral de sus alumnos y nos brindaba la ocasión de aprender a organizar un congreso, poner en marcha una revista de investigación o catalogar poesía de los siglos XVI y XVII conservada en manuscritos de la Biblioteca Nacional. En ese contexto surgió la posibilidad de realizar, bajo su dirección, una tesis doctoral sobre un manuscrito con cartas autógrafas de Quevedo. La edición y anotación de estas cartas reunía exactamente todo lo que yo deseaba estudiar: literatura, lengua e historia. Esa investigación se convirtió en mi tesis doctoral y confirmó mi decisión de dedicar mi vida académica y profesional a la filología.

REAL ACADEMIA ESPAÑOLA

REAL ACADEMIA ESPAÑOLA

¿Cómo se inicia y desarrolla tu carrera dentro de la RAE?

Mientras avanzaba en mi tesis doctoral y en otros proyectos, en 1994 recibí una llamada del Instituto Nacional de Empleo: *¿Le interesa un trabajo en la Real Academia Española?* Puedes imaginar mi sorpresa. Tras una entrevista en la propia sede, me informaron de que la colaboración —enmarcada en un convenio entre la RAE y el INEM— tendría una duración inicial de tres meses. El objetivo era comenzar la informatización de los ficheros léxicos de la institución.

En las últimas semanas de aquel periodo pidieron voluntarios para aprender a utilizar un escáner con el fin de «informatizar textos», como se decía entonces. Naturalmente me ofrecí, junto con algunos compañeros. Así participé en los primeros pasos de un proyecto que resultaría decisivo para el estudio y la investigación de la lengua española: la creación de un corpus. Los textos seleccionados se escaneaban, se procesaban con programas de OCR, se corregían y se enriquecían con marcas de codificación. Así nació el CREA (Corpus de Referencia del Español Actual), formado por textos producidos entre 1975 y el año 2000. Poco después surgió el CORDE (Corpus Diacrónico del Español), que abarcaba desde los orígenes de la lengua hasta 1974.

Tuve la fortuna de presenciar el inicio de la construcción de los corpus de la RAE y de ser consciente, ya entonces, de la trascendencia que aquello suponía en un momento de profunda transformación en la forma de trabajar de la institución. Poco después de aquellos tres meses, me incorporé de manera definitiva a la construcción del CREA y del CORDE.

Años después, siempre con el académico Guillermo Rojo al frente de los corpus académicos, se puso en marcha el CORPES XXI, que supuso un cambio fundamental al integrarse como proyecto de la Asociación de Academias de la Lengua Española.

He dado también clases en la Escuela de Lexicografía Hispánica, en la que fui jefa de estudios, y formé parte del Departamento de Comunicación. Además, construimos la parte correspondiente a los corpus en Enclave de Ciencia y colaboramos, desde la RAE, en la creación de CAPITEL (Corpus Anotado del Plan de Impulso de las Tecnologías del Lenguaje).

En definitiva, salvo mis etapas en la Escuela de Lexicografía y en Comunicación, mi carrera en la RAE se ha desarrollado fundamentalmente en torno al diseño, la construcción y el desarrollo de los corpus académicos.

¿Podrías contarnos cómo surgió el proyecto del CORPES XXI, dirigido por Guillermo Rojo, y cuáles son sus características básicas?

El académico Guillermo Rojo asumió la responsabilidad de poner en marcha el banco de datos de la RAE, en el que se enmarca el CORPES. Tras la experiencia con los corpus CREA y CORDE, que marcaron un hito en la investigación del español, pronto se vio la necesidad de replantear aquel diseño inicial. La expansión de Internet, el auge de los medios digitales y los cambios tecnológicos hicieron evidente que había que dar un paso más.

En 2007, la RAE presentó en el XIII Congreso de la ASALE, celebrado en Medellín, el proyecto de un corpus más ambicioso: el CORPES (Corpus del Español del Siglo XXI). El pleno de la Asociación de Academias aprobó que fuera la RAE quien se encargara de su diseño y desarrollo.

El CORPES XXI es, en cierto modo, la evolución natural del CREA. Aunque este último supuso un diseño pionero, pronto resultó insuficiente para atender las necesidades de las Academias en la elaboración del *Diccionario*, la *Gramática* y la *Ortografía*. A partir de los años 2000, con la llegada de las ediciones electrónicas, los libros digitales, la prensa en línea y las redes sociales, ya no era necesario escanear textos como al principio; el escenario era otro y el corpus debía adaptarse.

El diseño del CORPES responde a esos desafíos y estas son sus características básicas: para cada año del siglo XXI se incorporan 25 millones de formas procedentes de todos los países hispanohablantes, incluidos Guinea Ecuatorial y Filipinas. El 70 % de los textos procede de América y el 30 % de España. Una parte de los textos son transcripciones de la lengua oral, y algunos tienen la posibilidad de recuperar el audio o el vídeo del que proceden.

La tipología de los materiales se organiza en dos grandes bloques: ficción —novelas, relatos, guiones cinematográficos, teatro— y no ficción, donde se agrupan textos de prensa y ensayos clasificados en distintas áreas temáticas (actualidad, ciencia y tecnología, salud, arte, política,

entre otras). Participar en la construcción de este corpus fue, para mí, no solo un reto apasionante, sino también la confirmación de que mi vida profesional estaba unida de forma definitiva a la filología y a la creación de corpus lingüísticos.

¿En qué consiste tu trabajo como responsable técnica del CORPES XXI?

Mi labor principal consiste en coordinar todo el proceso de construcción del corpus y al equipo de personas que lo hace posible, siempre siguiendo las indicaciones del académico director del proyecto, Guillermo Rojo. Esto abarca desde la selección de los textos y la definición del sistema de codificación hasta la publicación de cada versión en la aplicación accesible desde la web de la RAE.

El trabajo se organiza en dos niveles. Por un lado, existe un equipo central en la propia Academia, encargado de garantizar que se cumpla el diseño general del corpus. Por otro, colaboran ocho equipos asociados a universidades e instituciones de distintos países —entre ellas, la Academia Argentina de Letras y la Academia Peruana de la Lengua—. Estos equipos reciben de la RAE la tarea de codificar una cantidad determinada de palabras, siguiendo criterios que aseguran un equilibrio geográfico (todos los países hispanohablantes están representados), temático (prensa, literatura, textos científicos, etc.) y cronológico (textos de distintos años del siglo XXI).

Una vez que esos materiales llegan a la RAE, pasan por un proceso de revisión, control de calidad e integración en el conjunto del CORPES. Después se aplican programas de anotación automática, que añaden información lingüística —como la categoría gramatical de cada palabra—, y finalmente se incorporan a la aplicación de consulta en línea. De este modo, los usuarios no solo acceden a las formas ortográficas, sino que es posible también realizar consultas mucho más exquisitas, basadas en la información gramatical.

¿Cuáles son los criterios de selección y codificación de los textos que se incluyen?

La selección de los textos debe ajustarse siempre al diseño del corpus y a los metadatos con los que van a codificarse, de modo que después sea posible recuperar la información de forma selectiva. Por ello, en la fase de codificación prestamos una atención especialmente rigurosa a la comprobación de esos metadatos, en particular, en lo relativo a la adscripción geográfica y cronológica de cada texto.

Esa precisión marca una diferencia fundamental con respecto a otras herramientas que también pueden utilizarse como corpus. En el CORPES, por ejemplo, un artículo firmado por Leila Guerriero y publicado en el diario *El País* de España no se clasifica como español, sino como argentino, porque lo determinante es la procedencia del autor y no únicamente el lugar de publicación.

Este tipo de decisiones resulta crucial para los investigadores: una adscripción correcta permite estudiar con fiabilidad cómo se usa el español en cada país y en cada momento histórico, sin distorsiones provocadas por la circulación internacional de autores o medios. De este modo, el CORPES ofrece una representación equilibrada y fiel de la diversidad del español, lo que constituye uno de sus principales valores frente a otras bases de datos textuales.

Por último, nos gustaría pedirte una reflexión sobre cómo ves el español en la actualidad, no solo teniendo en cuenta tu rica formación académica, sino también el hecho de que formás parte de una institución clave en la constitución y evolución de nuestra lengua.

El español de hoy es una lengua viva, diversa y en continua transformación. Como filóloga me interesa esa riqueza cultural que la lengua refleja, y como responsable técnica del CORPES tengo la oportunidad de observarla directamente en los textos que lo integran: voces de distintos países, registros muy variados y usos que cambian al ritmo de la sociedad.

Creo que el gran reto es asumir esa diversidad no como un problema, sino como una fortaleza. El español no es único ni uniforme: es plural, y en esa pluralidad reside precisamente su riqueza y su vitalidad.

Desde la RAE y la ASALE, esa visión se refleja en proyectos como el CORPES, que busca ofrecer una muestra fiel y equilibrada del español en todas sus variedades. Es una forma de poner al servicio de investigadores, docentes y profesionales una herramienta que documenta cómo evoluciona nuestra lengua y que ayuda a comprender mejor su complejidad y su riqueza compartida.

REAL ACADEMIA ESPAÑOLA

El funcionamiento del CORPES

Ahora que pudimos conocer un poco más a la responsable técnica del CORPES XXI, pasemos a la utilización práctica de esta poderosa herramienta que nos permite, tal como Mercedes Sánchez Sánchez lo expuso en su presentación durante el 5.° Encuentro Nacional de Correctores celebrado por UniCo a fines de 2024, «conocer el significado y características de palabras, expresiones y construcciones a partir de los *usos reales registrados*, y escoger, por tanto, la más adecuada para cada ocasión, ya sea que estemos editando, traduciendo o corrigiendo un texto».

Lo primero que encontramos en la interfaz son tres grandes bloques:

— Búsqueda

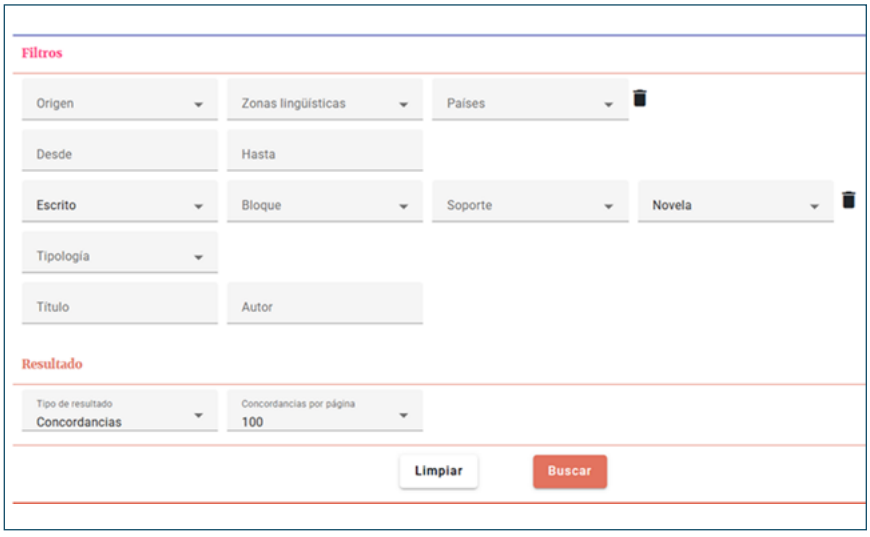
— Filtros

— Resultado

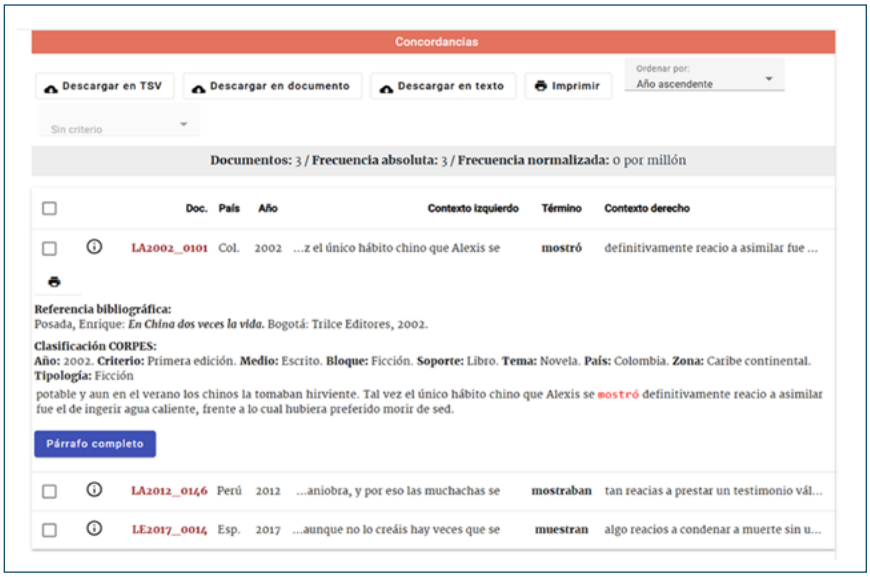
Dentro de cada uno de ellos, encontramos, a su vez, una serie de subcorpus que permiten acotar cada una de nuestras búsquedas. En la pestaña de «Búsqueda» tendremos la posibilidad de buscar «Palabras ortográficas» o «Elementos gramaticales». Dentro de la pestaña «Filtros» podremos especificar la distribución geográfica, cronológica, temática o tipológica de la palabra o el elemento gramatical que deseamos encontrar. Finalmente, en la pestaña «Resultado», podremos elegir las opciones de «Estadísticas», «Concordancias», «Documentos», «Coapariciones» e «Inventarios». En cada una de las búsquedas, además, el CORPES nos brindará información sobre la *frecuencia absoluta* (número total de apariciones de la palabra ortográfica o el elemento gramatical), el *número de documentos* (en los que aparecen registrados) y la *frecuencia normalizada* (número de apariciones por cada millón de palabras).

Para poner en práctica la herramienta, decidimos buscar la expresión *mostrarse reacio a*, precisamente porque no es de uso frecuente. Buscamos entonces los siguientes lemas: *mostrar* + *se* + *reacio* + *a*. Como vemos a continuación, al elegir la opción «Elementos gramaticales», aparece una línea donde escribimos el lema principal en su forma canónica, es decir, el verbo *mostrar* en infinitivo. Al hacer clic en el botón (+) que está a la derecha del casillero «Categoría gramatical», aparecen los *criterios de distancia*: es decir, podemos seguir completando los casilleros con el resto de los lemas de la expresión:

En el caso del lema *se*, elegimos un «Intervalo 1» por la «Izquierda» y por la «Derecha», y, en los restantes elementos, «Distancia 2» y «Distancia 3» por la «Derecha», ya que esa es la proximidad a la que se encuentran respecto del lema principal (expresada en número de palabras). Luego definimos los «Filtros» y el criterio del «Resultado».



De esta forma, observamos que el CORPES nos devuelve tres ejemplos de uso de la expresión *mostrarse reacio a* y que nos permite, además, obtener su contexto de producción.



Este artículo intenta ser una suerte de introducción a una herramienta sumamente rica que puede ser útil para todo tipo de profesional de la lengua. En la Comisión de Idioma Español, estamos seguros de que el CORPES XXI vino para quedarse y para ayudarnos, sobre todo, a lograr una mayor congruencia y coherencia en nuestros trabajos.■

Para ver la entrevista completa y más detalles sobre el CORPES XXI, los invitamos a visitar la página de la Comisión en el sitio web del Colegio: https://www.traductores.org.ar/matriculados/comisiones/comision-de-idioma-espanol/.

Por otra parte, adelantamos que la doctora Mercedes Sánchez Sánchez participará de manera virtual en la Jornada por el Día de la Corrección que celebraremos el próximo 27 de octubre.