REAL ACADEMIA ESPAÑOLA

NOTA DE PRENSA

El CORPES XXI de la RAE supera los 312 millones de formas ortográficas en su nueva actualización

- La versión 0.92 del Corpus del Español del Siglo XXI (CORPES XXI) incorpora alrededor de 33 millones de nuevas formas ortográficas a una recopilación que ya suma más de 300 000 documentos.
- El proyecto, dirigido por el académico Guillermo Rojo, constituye una base de datos fundamental para el estudio de la lengua española en la actualidad.
- Su consulta ya está disponible en línea en **este enlace**.

4 de junio de 2020

Ya está disponible para su consulta en línea la nueva actualización del Corpus del Español del Siglo XXI (CORPES XXI) de la Real Academia Española en colaboración con la Asociación de Academias de la Lengua Española (ASALE). La versión 0.92 de esta herramienta lingüística reúne **más de 300 000 documentos, que suman en torno a 312 millones de formas ortográficas, procedentes tanto de textos escritos como de transcripciones orales.** Con respecto a la versión anterior, esta actualización supone un incremento de alrededor de **33 millones de nuevas formas** incorporadas a esta excepcional base de datos del español dirigida por el académico de la RAE Guillermo Rojo.

Desde su lanzamiento en 2013, el CORPES XXI ha ampliado sus contenidos y mejorado su herramienta para lograr el propósito básico de este corpus de referencia: ser un **fiel retrato del español de nuestros días.** Para ello contiene textos de todos los tipos (novelas, obras de teatro, guiones de cine, noticias de prensa, ensayos, transcripciones de noticiarios radiofónicos o televisivos, transcripciones de conversaciones, discursos, etc.) y también de todos los países que constituyen el mundo hispánico.

EL CORPES XXI EN CIFRAS

Por lo que respecta al bloque de **ficción** (novelas, guiones de cine, relatos, obras de teatro), las formas del CORPES XXI sobrepasan los 88 millones, mientras que las contenidas en textos de libros de **no ficción** y en publicaciones periódicas (ciencias sociales, salud, política, artes, tecnología…) se acercan a los 219 millones.

Los textos procedentes de **libros** suponen casi 155 millones de formas; las **publicaciones periódicas** están representadas con unos 145 millones. Seis millones y medio más provienen de blogs, entrevistas digitales y miscelánea.

De los alrededor de 33 millones de formas incorporadas a la herramienta en la versión 0.92 del CORPES XXI, casi cinco millones proceden de **textos orales** (programas de radio y televisión, entrevistas en medios de comunicación, Youtube, etc.). Algunos archivos ofrecen el sonido alineado correspondiente a la transcripción; en otros es posible la descarga del archivo de audio, además de la visualización del vídeo de acuerdo con la procedencia del texto fuente.

En cuanto a la **distribución temporal**, aumenta el número de textos producidos entre 2016-2020, con algo más de 24 millones de formas. Por lustros, el mayor peso recae en el segmento 2006-2010, con unos 107 millones de formas; más de 97 millones corresponden a formas producidas entre 2001 y 2005, y el periodo de 2011 a 2015 alcanza casi 79 millones de formas.

Respecto a la **procedencia** de los documentos del CORPES XXI, el equilibro previsto entre España y América (30 %-70 %) se mantiene: las formas producidas en textos clasificados como pertenecientes a España suponen algo más del 30 % y los de América superan los 204 millones de formas.

HERRAMIENTA FUNDAMENTAL DE LA LINGÜÍSTICA

El Corpus del Español del Siglo XXI (CORPES XXI) es, al igual que CREA, un corpus de referencia. En lingüística, se llama *corpus* a un conjunto lo más extenso y ordenado posible de textos. Los corpus son empleados habitualmente para conocer el contexto y las propiedades de las palabras, expresiones y construcciones a partir de los usos reales registrados. Dado el tamaño que poseen, los corpus tienen que estar en formato electrónico.

Un corpus general (llamado *de referencia*) tiene como propósito básico el de servir para obtener las características globales que presenta una lengua en un momento determinado de su historia. En el caso del español actual, el corpus debe contener textos de todos los tipos y también de todos los países que constituyen el mundo hispánico.

Prensa Real Academia Española
comunicacion@rae.es
91 420 14 78 297