REAL ACADEMIA ESPAÑOLA

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

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

REAL ACADEMIA ESPAÑOLA

La Real Academia Española (RAE) define a un corpus como un «conjunto de (fragmentos de) textos, orales o escritos, producidos en condiciones naturales, conjuntamente representativos de una lengua o una variedad lingüística, en su totalidad o en alguno(s) de sus componentes, que se almacenan en formato electrónico y se codifican y anotan con la intención de que puedan ser analizados científicamente». El objetivo de un corpus es ofrecer una muestra representativa del uso de la lengua: dar a conocer el significado y las características de las palabras, expresiones y construcciones a partir de un uso concreto y registrado. En 2007 la Asociación de Academias de la Lengua Española (ASALE) encomendó a la RAE la construcción del CORPES XXI (Corpus del Español del Siglo XXI), el cual se presentó por primera vez en 2013 y se ha ido actualizando y transformando con el tiempo. Cuando conocemos la finalidad, la historia y las características de este recurso, nos damos cuenta de su riqueza y valoramos aún más la importancia de tenerlo entre nuestras fuentes de autoridad y de referencia.

En tiempos de la inteligencia artificial (y ante una mar de conocimiento que sigue expandiéndose en tiempo real), es fundamental contar con herramientas confiables que no solo aporten efectividad a nuestro trabajo, sino que destaquen las ventajas de consultar un corpus nacido de textos orales y escritos producidos, seleccionados y codificados por seres humanos (aunque esto último pueda sonar apocalíptico).

Fue así como durante el 5.º Encuentro Nacional de Correctores celebrado por UniCo a fines de 2024 descubrimos a Mercedes Sánchez Sánchez, doctora en Filosofía y Letras (Filología Hispánica) por la Universidad Autónoma de Madrid, quien a fines de 1994 se incorporó al proyecto de digitalización de los ficheros de papel de la RAE y dio comienzo así a una vasta carrera que la llevaría a formar parte del desarrollo de los distintos corpus de la institución, entre ellos, el CREA (Corpus de Referencia del Español Actual), el CORDE (Corpus Diacrónico del Español) y el CORPES XXI. También fue una de las redactoras del *Libro de estilo de la lengua española según la norma panhispánica.*

De allí surgió nuestro interés por entrevistar a Mercedes, quien es la responsable técnica del CORPES XXI. Este corpus constituye la herramienta de consulta más actualizada de la RAE y la ASALE: no solo cuenta ya con una versión 1.3, sino que lleva incluidas más de 400 millones de formas ortográficas. Pero dejemos que sea Mercedes quien nos dé los detalles...

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.

¿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

REAL ACADEMIA ESPAÑOLA

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.

¿Qué sucede cuando se encuentran con textos ambiguos o contradictorios sobre determinado tema?

La clave a la hora de valorar la inclusión de un texto en el CORPES es el diseño al que nos debemos: somos nosotros quienes seleccionamos los materiales. Si durante el proceso de lectura detectamos que un texto no se ajusta a los criterios establecidos, simplemente se descarta.

Al tratarse de una herramienta en continuo desarrollo, el control de calidad resulta esencial y se aplica de manera constante. Por eso, ante cualquier situación que impida la incorporación o el mantenimiento correcto de un texto en el corpus, se prefiere prescindir de él antes que comprometer la fiabilidad del CORPES.

Ante el avance de la inteligencia artificial, ¿cómo dirías que este hecho afecta la producción y la actualización del CORPES?

Podemos decir que la inteligencia artificial influye en el CORPES de dos maneras. Por un lado, es una aliada: nos ayuda a agilizar tareas como la anotación automática o la detección de errores, lo que hace el trabajo más eficiente. Por otro, supone un desafío, porque nos obliga a extremar el cuidado en la selección de textos para garantizar que reflejen el español real de los hablantes y no producciones generadas por máquinas.

El reto, en definitiva, está en mantener ese equilibrio: aprovechar lo que la inteligencia artificial aporta sin renunciar a la fiabilidad que distingue al CORPES.

Si comparamos el CORPES con herramientas como Google Ngram, ¿cuáles dirías que son las ventajas y desventajas de utilizar la herramienta que propone la RAE?

La diferencia fundamental entre el CORPES y Google Ngram está en el diseño y en el control de calidad. Google Ngram se alimenta de millones de libros digitalizados sin una revisión exhaustiva: es útil para observar tendencias generales, pero resulta propenso a errores de metadatos como la datación y, además, se limita al ámbito de los libros, sin incluir géneros como la prensa o los materiales orales, a no ser que formen parte de lo que consideran «libros».

El CORPES, en cambio, se construye con criterios rigurosos: cada texto se selecciona y codifica con precisión geográfica, cronológica y temática. Esto asegura que los resultados sean representativos y comparables.

Mientras Ngram contiene textos en español hasta 2022, el CORPES incorpora ya materiales de 2025. Su desventaja frente a Ngram es que crece más lentamente, porque ese rigor exige más tiempo y recursos; pero su principal virtud es la fiabilidad, la variedad de géneros y la utilidad para la investigación lingüística.

Teniendo en cuenta que la RAE ha sido siempre la gran «reguladora» de la lengua española, ¿tienen referencias sobre la repercusión de esta herramienta de consulta en las distintas comunidades hispanohablantes?

Conviene recordar que, desde la adopción de la política panhispánica, el trabajo académico sobre el español se realiza de manera conjunta en el marco de la ASALE. Todas las obras normativas se publican por consenso entre las Academias: el *Diccionario* ya no es el *DRAE*, sino el *DLE*; lo mismo ocurre con la *Nueva gramática* y con la *Ortografía*.

En este contexto, el CORPES es también un proyecto panhispánico. Ha inspirado la creación de otros corpus, se utiliza con frecuencia en investigaciones y despierta un gran interés entre profesionales de la lengua, como traductores y correctores. Su mayor repercusión está en ofrecer a investigadores y especialistas un instrumento fiable y actualizado para conocer mejor la realidad del español en todas sus variedades.

En una entrevista del año 1983, Julio Cortázar afirmó: «Dentro de unos años, ciertas palabras, que ahora todo el mundo usa, como "chantapufi", van a desaparecer». El mundo ha cambiado mucho desde entonces, por eso nos gustaría preguntarte esto: ante la vorágine de información y nuevas tecnologías a las que todos estamos expuestos, ¿es posible mantener actualizada una herramienta como el CORPES?

Las palabras no desaparecen, simplemente dejan de usarse. Si quedan registradas en el CORPES, siempre permanece un testimonio de ellas en su contexto, lo que permite rastrear su historia y su evolución. El corpus, así, no solo refleja el presente del español, sino que también conserva la memoria de su cambio.

Mantenerlo actualizado es, por supuesto, un reto en un mundo de sobreabundancia informativa y cambios tecnológicos constantes. La clave, una vez más, está en el diseño: el CORPES no pretende abarcarlo todo, sino ofrecer una muestra representativa y fiable del español actual, gracias a criterios de selección, sistemas de revisión y el apoyo de herramientas tecnológicas que agilizan la incorporación de materiales.

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 ella lo expuso en su presentación durante el 5.º 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».

Partamos entonces del aspecto general para ir de a poco a lo particular: 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).

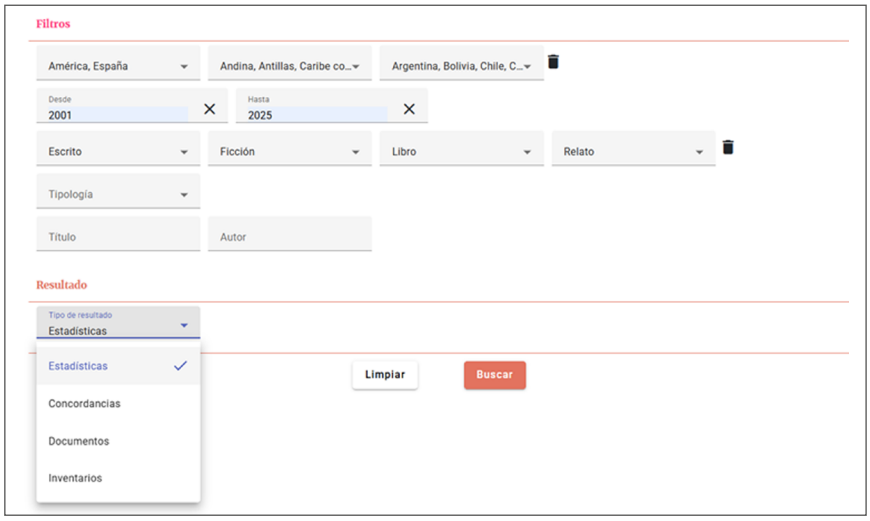
No hay duda de que para poder utilizar una herramienta como el CORPES es necesario practicar con ella. Por ese motivo, y dado que, como dijo alguna vez Jorge Luis Borges, el tres es un número mágico, veremos a continuación tres ejemplos de uso práctico según distintos criterios de búsqueda y recuperación de la información.

1) Búsqueda por palabras ortográficas

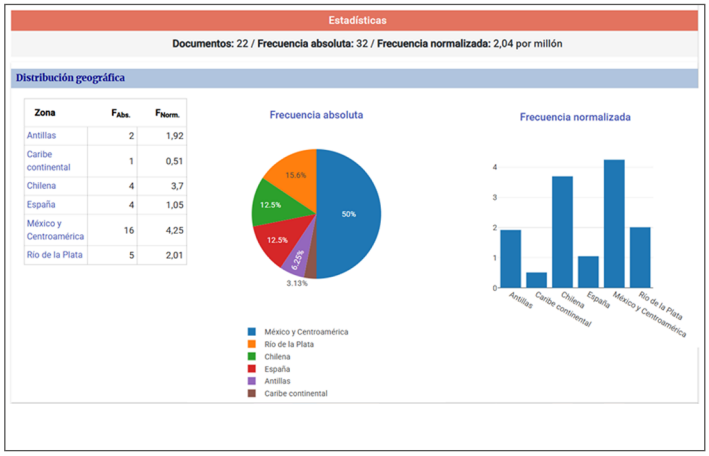
Para hacer una búsqueda por palabras ortográficas, elegimos esa opción en el casillero de «Búsqueda» y escribimos el término que deseamos consultar:

Así, si deseamos averiguar cuál ha sido la distribución geográfica, cronológica, temática y tipológica, por ejemplo, de la palabra *rayuela* desde el año 2001 hasta el año 2025 inclusive, tendremos la posibilidad de seleccionar distintos «Filtros» para acotar nuestra búsqueda: medio (escrito u oral), bloque (ficción o no ficción), soporte (libro, en este caso) y área temática (relato, para el ejemplo elegido).

Una vez que elegimos los filtros que deseamos aplicar, vamos a la pestaña «Resultado», la cual nos permitirá indicarle a la herramienta cuál es nuestro interés específico. En este caso, decidimos elegir la opción «Estadísticas», ya que nuestro interés principal es obtener tablas y gráficos de frecuencia de la palabra *rayuela* según su distribución geográfica, cronológica, temática y tipológica.

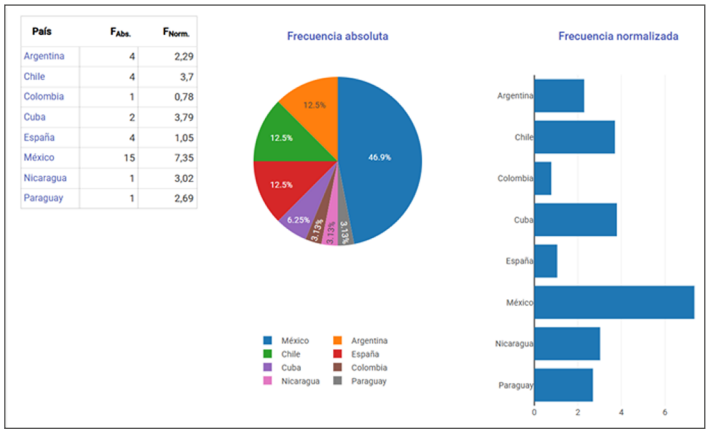


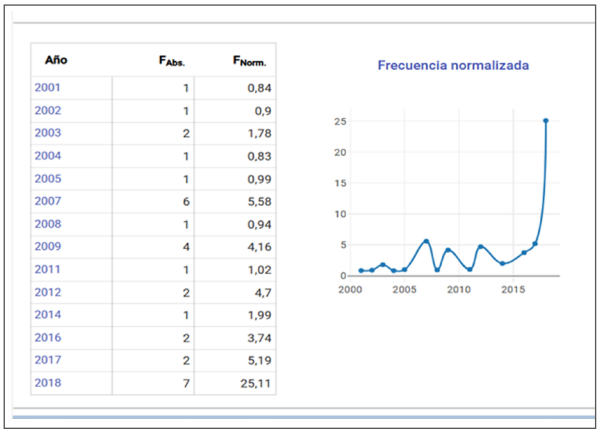
Al hacer clic en el botón «Buscar», se obtiene la información deseada:



De esta forma, es posible ver que la palabra *rayuela* se viene utilizando con mucha más frecuencia en México y Centroamérica que, por ejemplo, en la Argentina, como muchos de nosotros hubiéramos podido pensar (Cortázar, siempre Cortázar...).

Asimismo, el uso del término no deja de crecer desde el año 2015, en textos escritos de ficción cuya área temática es el relato, tal como lo establecimos en nuestra búsqueda.





REAL ACADEMIA ESPAÑOLA

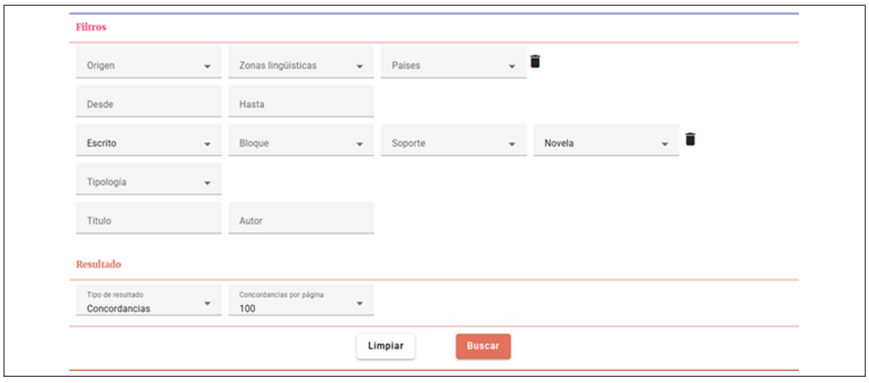
2) Búsqueda por elementos gramaticales

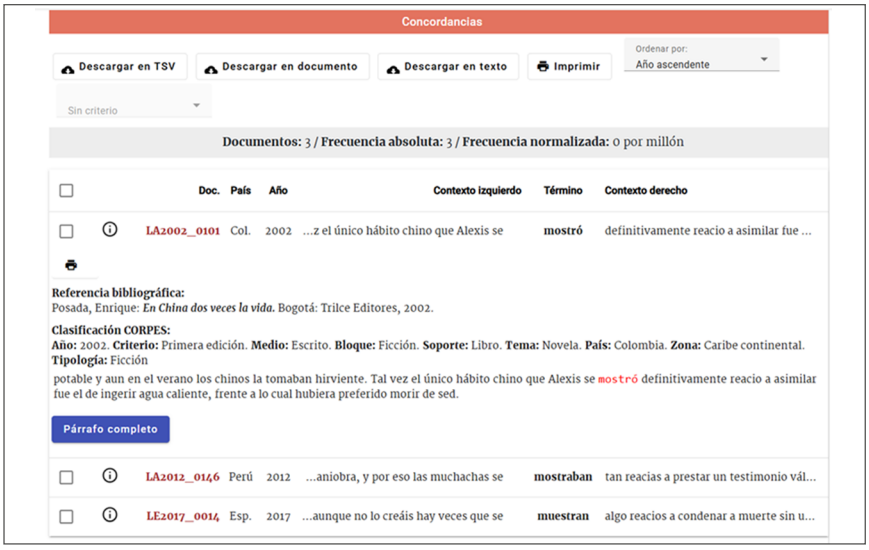
La segunda opción de búsqueda que nos ofrece el CORPES XXI es por elementos gramaticales.

Para ponerla en práctica, 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» (en nuestro ejemplo seleccionamos la opción «Concordancias», ya que estamos interesados en encontrar ejemplos de uso de la expresión).

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.



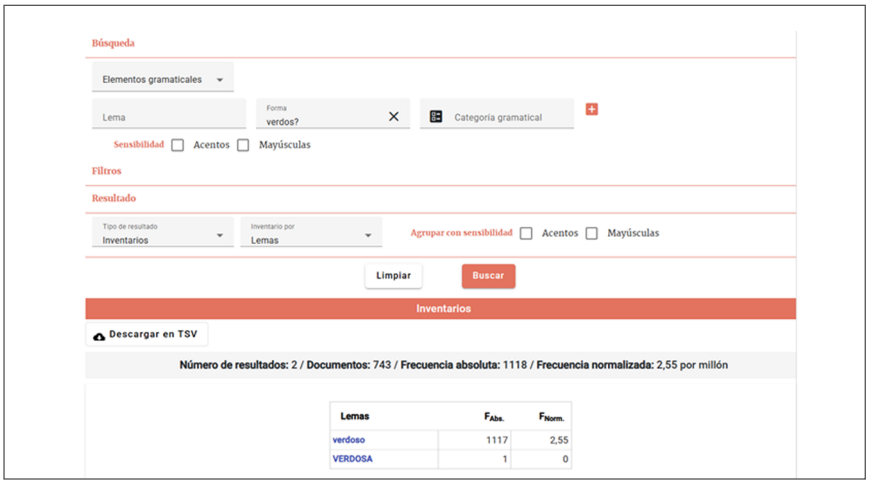


3) Combinación de criterios de búsqueda y comodines

Una de las cuestiones más novedosas del CORPES XXI es la posibilidad de combinar distintos criterios de búsqueda. Por ejemplo, podríamos hacer una búsqueda destinada a averiguar con qué frecuencia un sustantivo se escribe seguido de uno o más adjetivos (en cuyo caso podríamos dejar vacíos el casillero «Lema» y el casillero «Forma», y concentrarnos en la «Categoría gramatical»), o pedirle a la herramienta que nos confeccione un listado de adverbios terminados en *-mente.*

En el último caso, puede servirnos mucho hacer uso de los *comodines*. Estos son signos que nos ayudan a hacer búsquedas personalizadas, tal como sucede con muchos de los navegadores que utilizamos a diario. Por ejemplo, si hacemos una búsqueda de elementos gramaticales y en el casillero «Lema» escribimos **mente*, en el casillero «Categoría gramatical» elegimos «Adverbio» y en el criterio de «Resultado» elegimos la opción «Inventarios», el CORPES nos brindará un listado de adverbios terminados en *-mente*, ya que, tal como se detalla en su *Guía básica de consulta*, el asterisco (*) «sustituye a cualquier número de caracteres (incluyendo ninguno) a partir de la posición en que aparece».

El comodín representado por el signo de interrogación (?), en cambio, reemplaza a un solo carácter, lo cual puede ser de gran utilidad a la hora de hacer una búsqueda para obtener la frecuencia de aparición de una forma masculina o femenina:



Conclusiones

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. Si mientras estamos traduciendo un texto jurídico, por ejemplo, dudamos del uso real de un término, ¿dónde conviene realizar una búsqueda?, ¿en textos escritos u orales? Y, a la hora de tener que traducir, editar o corregir un video con lenguaje adolescente, ¿no conviene conocer la distribución cronológica de aquellos términos o jergas que nos resultan desconocidos?

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 aquellas personas que estén interesadas en seguir probando la herramienta o comenzar a trabajar con ella, dejamos a continuación los enlaces del CORPES XXI y de su *Guía básica de consulta*:

CORPES

Guía básica de consulta

Para contarnos qué les pareció este artículo, hacernos comentarios o sugerencias, pueden escribirnos a espanol@traductores.org.ar.