

Usos y aplicaciones de corpus lingüísticos en la enseñanza de ELE

Rosalía Calle Bocanegra

Departamento de Lenguas Romances, Universidad Palacký de Olomouc

rosalia.callebocanegra@upol.cz

Resumen

Los corpus lingüísticos son fuentes que aportan una gran cantidad de *input* auténtico que pueden servir para resolver dudas lingüísticas, tanto de profesores como estudiantes, o diseñar materiales para clase de ELE, pero al mismo tiempo pueden ser complejos de usar y presentar interfaces áridas. En este artículo comentaremos los diferentes tipos y características de los corpus lingüísticos del español, las ventajas y desventajas de su uso en clase de ELE y algunos ejemplos concretos de cómo usar los corpus en clase de ELE.

Palabras clave: corpus lingüísticos, ELE, CORPES XXI, colocaciones

Abstract

Linguistic corpora are huge sources of authentic input that can shed light on to linguistic doubts both from teachers and students and they can be used for designing materials for teaching Spanish as a foreign language, but at the same time it can be difficult to use them and they can present arid interfaces. This paper will discuss the different types of corpora and their characteristics, the advantages and disadvantages of using them for teaching

Spanish to foreigners and some concrete examples of how they could be used in the classroom.

Keywords: linguistic corpora, teaching Spanish to foreigners, CORPES XXI, collocations

1. Introducción

En primer lugar, debemos responder a la pregunta: ¿qué es un corpus lingüístico? Un corpus lingüístico es un conjunto de textos —orales y/o escritos— normalmente digitalizados que recogen millones de palabras, cuyo objetivo es representar a la lengua y ofrecer un modelo de la misma. Se suelen encontrar en soporte electrónico por su gran tamaño y se usan como fuente de conocimiento del significado de las palabras y expresiones, ya que se trata de datos reales de la lengua tomados de novelas, periódicos, radio, televisión y un largo etcétera.

Uno de los corpus más usados en todo el mundo es precisamente Google, que se caracteriza por la gran cantidad de datos que lo conforman, la rapidez a la hora de realizar las búsquedas y la constante actualización de los datos presentes en el mismo. Sin embargo, algunas de las desventajas que se desprenden de su uso son la imposibilidad de hacer búsquedas refinadas —por registro o tema, por ejemplo—, la no selección que hace que podamos encontrarnos ante textos con faltas de ortografía, por ejemplo, y que no está lematizado ni etiquetado. Un corpus lematizado es aquel que, cuando se realiza la búsqueda de una palabra en el campo “lema”, arroja resultados en todas las formas flexivas. En la Figura 1 podemos apreciar los resultados de la búsqueda del verbo “caber” insertado en el campo de “lema”:

The screenshot shows the search interface of CORPES XXI. At the top, there are navigation tabs: **Concordancias**, **Coapariciones**, **Configuración**, **Ayuda**, **Estadística**, **Modo de cita**, **Sugerencias**, and **Preg**. Below these, there is a search bar with the following fields: **Lema** (containing 'caber'), **Forma** (empty), **Clase de palabra** (dropdown menu), and a checkbox for **Grafía original**. There are also buttons for **+**, **-**, **Concordancia**, **Estadística**, and **Nueva consulta**.

REF. (Clasificación, país)	CONCORDANCIA
<input type="checkbox"/> 41 2001 Arg.	lucha sangrienta y llena de odio, que aún no han tenido la grandeza de superar, cabe
<input type="checkbox"/> 42 2001 Arg.	Sólo nos cabe
<input type="checkbox"/> 43 2001 Arg.	los artistas. Ni que se haya operado una especie de negación sobre el lugar que le cabe
<input type="checkbox"/> 44 2001 Arg.	No tuvo, ni por asomo, la misma suerte otro artista que no cupo
<input type="checkbox"/> 45 2001 Arg.	George Yúdice. Esa necesidad de estéticas y concepciones muy nacionales pero que quepan
<input type="checkbox"/> 46 2001 Arg.	corresponden a la Ciudad de Buenos Aires y la provincia de Chubut respectivamente. Cabe
<input type="checkbox"/> 47 2001 Arg.	(ya sea en forma de leche fluida, fórmulas o semisólidos), y en este sentido nos cabe
<input type="checkbox"/> 48 2001 Arg.	minutos de diferencia entre la hora del nacimiento de Mirta y la de mi hermana, cabía
<input type="checkbox"/> 49 2001 Arg.	lo viera, tomó una de las tantas botellas de vino y la guardó en su mochila. Aún cabía
<input type="checkbox"/> 50 2001 Arg.	nos encontramos con que las encuestas dan un 83 por ciento a favor del Presidente. Cabría
<input type="checkbox"/> 51 2001 Arg.	La mano, el cigarrillo, toda esta torpe escenografía que no cabe
<input type="checkbox"/> 52 2001 Arg.	ese estereotipo, que, no descartaba, en determinadas condiciones, también pudiese caber

Figura 1. Captura de pantalla de la búsqueda de “caber” en CORPES XXI

Si realizáramos la búsqueda del verbo “caber” en el campo de “forma”, el corpus solo nos ofrecería los resultados de ese infinitivo en concreto. Un corpus etiquetado es aquel en el que las palabras que lo conforman están previamente clasificadas según su categoría gramatical. Si está etiquetado, podemos buscar en un corpus la forma “presente” como un sustantivo (como en “vive el presente”) o como un adjetivo (“el presente libro se dirige a un público no experto”) o verbo (“aprovecharemos la oportunidad que se nos presente”). En la Figura 2 aportamos un ejemplo de la búsqueda de “presente” como sustantivo:

The screenshot shows the search interface of the CORPES XXI database. At the top, there is a navigation bar with tabs: **Concordancias**, **Coapariciones**, **Configuración**, **Ayuda**, **Estadística**, **Modo de cita**, **Sugerencias**, and **Preguntas**. Below this, there is a search form with the following fields: **Lema** (empty), **Forma** (set to 'presente'), **Clase de palabra** (set to 'sustantivo'), and a checkbox for **Grafía original**. Below the search form are buttons for **+**, **-**, **Concordancia**, **Estadística**, and **Nueva consulta**.

The search results are displayed in a table with two columns: **REF. (Clasificación, país)** and **CONCORDANCIA**. The results are as follows:

REF. (Clasificación, país)	CONCORDANCIA
<input type="checkbox"/> 1 2001 Méx.	futuro, atemorizada de que ciertos acontecimientos sucedan, y yo prefiero vivir el presente
<input type="checkbox"/> 2 2001 Arg.	habituales. El concepto de este disco era un repaso por toda mi carrera, por el presente
<input type="checkbox"/> 3 2001 Esp.	hacia el continente europeo, y librando las batallas del pasado en vez de las del presente
<input type="checkbox"/> 4 2001 Méx.	-Guardar periódicos es como guardar el presente
<input type="checkbox"/> 5 2001 Méx.	rostro de don Lupe es de un amarillo espectral, como de aparecido- ¿Te devuelvo al presente
<input type="checkbox"/> 6 2001 Méx.	En ese apagón del tiempo hay tiempo para modificar el presente
<input type="checkbox"/> 7 2001 Méx.	hay tiempo para modificar el presente, y el curso de la historia que procede del presente
<input type="checkbox"/> 8 2001 Méx.	éstos me dan una impresión de inmovilidad temporal, de haberme quedado fijo en el presente
<input type="checkbox"/> 9 2001 Méx.	La Zona del Silencio tal vez está en el presente
<input type="checkbox"/> 10 2001 Méx.	-Ésa es mi historia, vuelve al presente
<input type="checkbox"/> 11 2001 Méx.	-Trato de devolverte al presente
<input type="checkbox"/> 12 2001 Méx.	sé qué estoy mirando ni dónde me encuentro, si aquí o allá, si en el ayer o en el presente

Figura 2. Captura de pantalla de la búsqueda de “presente” como sustantivo en CORPES XXI

Hasta ahora hemos visto ejemplos tomados del CORPES XXI, que se trata de un corpus general de la lengua española porque contiene textos, tanto orales como escritos, de todos los tipos y países. Otro tipo de corpus son los corpus diacrónicos como el CORDE, el cual contiene textos de todos los periodos y lugares en los que se habló español desde sus inicios hasta el año 1794. Para los estudiosos de la adquisición de lenguas extranjeras son especialmente provechosos los corpus de aprendices CAES y SPLLOC, ya que a través de ellos los profesionales de ELE (profesores, creadores de materiales, investigadores del campo de la didáctica o adquisición de lenguas extranjeras, etc.) pueden indagar acerca del vocabulario que resulta complicado o del campo gramatical que genera un mayor número de errores entre los aprendices de una lengua extranjera. El corpus de aprendices de español (CAES) recopila textos escritos por 1423 estudiantes de ELE con niveles de competencia que comprenden desde el A1 hasta el C1 del Marco Común Europeo de Referencia para las Lenguas (MCER). Los aprendices tienen como primera lengua (L1) o lengua materna el árabe, ruso, portugués,

francés, chino mandarín o inglés. Sin embargo, *Spanish Learner Language Oral Corpora* (SPLLOC) es una recopilación de textos orales producidos por aprendices con inglés como L1. También especialmente útiles para los traductores son las combinaciones de corpus y diccionarios como Linguee, que es un diccionario multilingüe en el que aparecen las traducciones de textos reales en su contexto.

2. Ventajas y desventajas del uso de corpus en clase de ELE

La principal ventaja del uso de corpus en clase es que suponen una fuente de *input* de calidad y auténtico, lo cual contrasta con los materiales adaptados y con *input* modificado que se suelen encontrar en los manuales de ELE. Es común que un profesor nativo de ELE se encuentre en situaciones en las que recurre a su intuición para decir si algo en español le parece correcto o no simplemente porque le suena bien o mal, pero no sabe el motivo ni puede explicárselo a los estudiantes. Los corpus pueden ayudar, a través de los ejemplos que aportan, a sistematizar el conocimiento intuitivo que un hablante nativo pueda tener de su lengua materna (Pitkowski *et al.* 2009).

Otra de las ventajas que tienen los corpus es la gran cantidad de usos que pueden tener para ELE: se pueden usar para encontrar más ejemplos reales y contextualizados cuando el profesor o los estudiantes no encuentran más, resolver dudas lingüísticas de los aprendices, diseñar materiales didácticos que pueden usarse en clase o fomentar la autonomía del aprendiz animándolo a que él mismo use el corpus para resolver actividades.

Por otro lado, algunas de las desventajas de su uso son que pueden presentar una interfaz árida y poco atractiva (González García 2018) y que el término “corpus” se asocia con expertos por lo que pueden dar la impresión de ser difíciles de usar (Buyse 2017). Sin embargo, nosotros animamos a los profesionales de ELE a usar esta herramienta y vamos a ilustrar algunos ejemplos de uso de los mismos en el siguiente apartado.

3. Ejemplos de uso del CORPES XXI en clase de ELE

En primer lugar, vamos a mencionar algunas de las características del CORPES XXI para luego pasar a los ejemplos de uso. La versión actual del corpus (0.94), publicada en julio de 2021, recoge aproximadamente 350 millones de formas que proceden tanto de textos orales como escritos. De ellas, el 70% provienen de textos americanos (variedad andina, chilena, rioplatense, mexicana y de Centroamérica, estadounidense, filipina, de Guinea Ecuatorial, de las Antillas y del Caribe continental) y el 30% restante de textos de España. Los textos del bloque de ficción provienen de guiones, novelas, relatos, etc. y los textos del bloque de no ficción provienen de libros y prensa sobre temas de actualidad, artes, ciencias, política, salud, etc.

3.1. Encontrar más ejemplos de una expresión

Imaginemos que en clase estamos trabajando una serie de expresiones y una de ellas es “sacar el parecido”. El profesor se queda sin ejemplos y los estudiantes precisan más. Una de las ventajas del CORPES XXI es que permite hacer búsquedas por proximidad, es decir, permite buscar una palabra con otra a una cierta distancia (a la derecha o izquierda) de la primera. La búsqueda de esta expresión se podría hacer de la siguiente forma: escribimos el verbo “sacar” en el campo de “lema” y luego pulsáramos el botón “+ proximidad” y en el campo “forma” escribimos “parecido” que debe de aparecer en dos intervalos a la derecha (dejamos la posición del primer intervalo a la derecha libre para el artículo “el” de “sacar el parecido”). Pulsamos el botón de concordancia y nos aparecen los ejemplos de la Figura 3:



Figura 3. Captura de pantalla de la búsqueda de la expresión “sacar el parecido” en CORPES XXI

Otra información que aporta el corpus en la izquierda es el año y lugar de procedencia de los textos, como podemos ver en la Figura 3.

3.2. Ilustrar la frecuencia de uso de formas en diferentes variedades del español

Además de la búsqueda por proximidad, CORPES XXI también permite comparar las estadísticas de la aparición de formas lingüísticas dependiendo de su distribución por zonas, países, períodos o temas. Además de ofrecer la frecuencia absoluta de la forma buscada, también ofrece la frecuencia normalizada (número de ocurrencias por cada millón de palabras en cada zona, país, período, etc.) y un gráfico circular con los datos. Esta función es especialmente útil para mostrar a nuestros estudiantes dónde se usa más una forma y dónde menos. Por ejemplo, si buscamos la forma “vos” en el campo de forma y pulsamos el botón de estadística nos aparecen datos de la Figura 4:

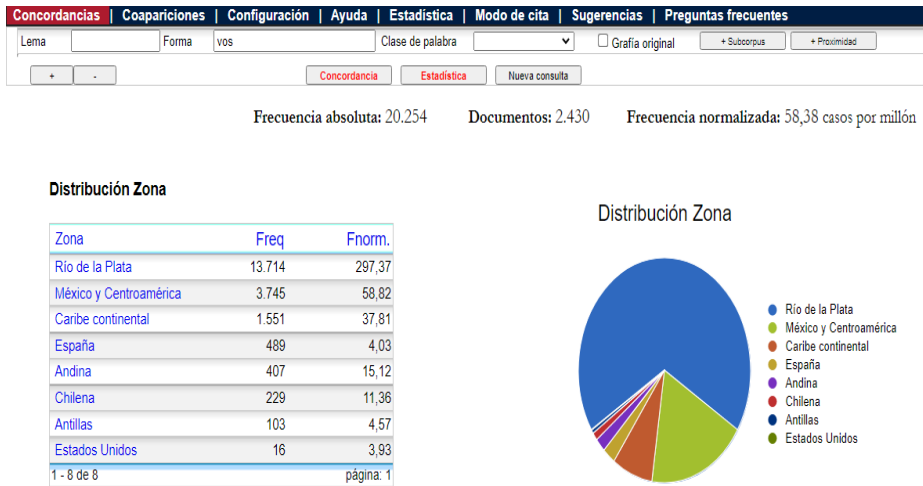


Figura 4. Captura de pantalla de la búsqueda de estadísticas de la forma “vos” en CORPES XXI

Como vemos en la Figura 4, la variedad rioplatense es la que más usa esta forma en comparación con el resto. Esta búsqueda también podría hacerse comparando las estadísticas de la forma del pretérito perfecto “he ido” en comparación con “fui” para mostrar en qué países o zonas se usa más, por ejemplo.

3.3. Buscar colocaciones

Otra de las funciones útiles del corpus es que permite buscar palabras que frecuentemente aparecerán junto a otras, o lo que es lo mismo, las colocaciones. Esto podemos hacerlo con el apartado “coapariciones”. En la Figura 5 aportamos un ejemplo de la búsqueda por coapariciones del lema “meter”.

	Clase	Freq	MI	LL SIMPLE	T-SCORE
berenjenal	sustantivo	66	11,71	431,48	8,12
calzador	sustantivo	20	10,82	116,32	4,47
lío	sustantivo	1013	10,5	5.666,68	31,82
fregado	sustantivo	15	10,48	83,58	3,87
emboque	sustantivo	10	9,92	51,92	3,16
plomazo	sustantivo	20	9,81	102,38	4,47
frentazo	sustantivo	10	9,71	50,51	3,16
cizaña	sustantivo	21	9,67	105,58	4,58
embrollo	sustantivo	61	9,25	290,00	7,81
baza	sustantivo	82	9,14	384,50	9,05
lavavajillas	sustantivo	16	9,06	74,16	4,00
brete	sustantivo	15	8,98	68,81	3,87
congelador	sustantivo	80	8,78	357,01	8,94
pata	sustantivo	833	8,76	3.707,76	28,82
bañera	sustantivo	118	8,7	519,98	10,86
bolsillo	sustantivo	1194	8,69	5.266,87	34,49
follón	sustantivo	22	8,66	96,49	4,69

Figura 5. Captura de pantalla de la búsqueda de coapariciones con el lema “meter” en CORPES XXI

Así, podemos observar que algunas de las palabras que aparecen frecuentemente junto al verbo “meter” son: berenjenal, lío, cizaña, embrollo, baza, pata (de la expresión “meter la pata”), follón, etcétera.

3.4. Diseñar actividades para clase

Vamos a presentar una propuesta en la que hemos usado el corpus para crear una actividad para la clase de ELE que trate la morfología del español (adaptada de González García 2018). Usando el comodín “asterisco” podemos buscar palabras en CORPES XXI que empiecen por cualquier cosa pero que terminen en lo que hay detrás del asterisco. Por ejemplo, si

buscamos en el campo de forma *zón nos aparecerán ocurrencias de todas las palabras terminadas en -zón. En la Figura 6 y 7 se muestran los resultados de *zón como sustantivo y como adjetivo, respectivamente:

121.959 casos en 42.110 documentos.

REF. (Clasificación, país)	CONCORDANCIA	Ordenar por:
1 2001 Méx.	generalmente esto sucede cuando la gente se deja guiar por los prejuicios de la razón . Le sugiero que para que estos pen...	Año ascen
2 2001 Méx.	importa. Y ya lo dejen a un lado y empiecen a escribir. Al final del día verán que la razón es muy limitada y sólo se le ocurrir	
3 2001 Cuba	deben a que yo casi no veo. Se ha estropeado la lente pero veo con la cabeza y el corazón . Son para que no me moleste la cla	
4 2001 Arg.	Joaquín Dicenta, con quien compartí emociones y sensibilidades. Se le rompió el corazón cuando amanecieron sus primeros 5	
5 2001 Arg.	60. ¿Qué hay que hacer para ser un genio? ¿Nacer ya así? Abrazos de corazón .	
6 2001 Arg.	Que la legislación, con respecto al cannabis, es una vergüenza y un atropello a la razón .	
7 2001 Esp.	Un día en el cricket, aunque Lord Tebbitt no tenga razón , es en cualquier caso un experiment	
8 2001 Esp.	llorón, indisciplinado, errático, infantil, guerrero, familiar, explosivo, todo corazón . Sólo tenía de inglés el ser un perdo	
9 2001 Esp.	Nos gustaría enviarte los comentarios de los niños y jóvenes a los que tocas el corazón . ¿Dónde podemos hacerlo? Si lo en	
10 2001 Esp.	dan reglas, pero se recalca una y otra vez la necesidad de hacer las cosas por una razón , que ante todo se tengan en cuenta	
11 2001 Bol.	Don Juan Tenorio pagó su culpa con un par de tijeras clavadas en el centro de su corazón fermentido. El hecho ocurrió en la ca	
12 2001 Bol.	tijeras que guardaba debajo de la almohada y se las clavó una y veinte veces en ese corazón que realmente no merecía seguir la	

Figura 6. Captura de pantalla de la búsqueda de *zón como sustantivo en CORPES XXI

732 casos en 545 documentos.

REF. (Clasificación, país)	CONCORDANCIA	Ordenar por:
1 2001 Méx.	puede confundir a la Harpagus horribilis con el pez de las cuevas. Este último es cabezón , tiene protuberancias en lugar d	Año ascen
2 2001 Arg.	huele las manos y percibe que aún le quedan las huellas tenues del perfume algo dulzón de Sandra.	
3 2001 Arg.	auto me acomodé lo mejor que pude. El chofer, un tipo de pelo corto, rubio y algo cabezón , apenas sí me saludó y parecía	
4 2001 Arg.	todo lo que quisiese. De su marido ejemplar y de sus tres jóvenes gracias y del retozón rubio alegría de la familia. Por q	
5 2001 Ur.	-Es el vino blanco -dice Angélica-. El vino blanco cuando no es bueno es cabezón , y por aquí no he encontrado vir	
6 2001 Chile	con leche recién preparado que lo esperaba en la amplia terraza. Pero otro olor dulzón , como a claveles frescos, predo	
7 2001 Nic.	pintando cartelones: ¡Abajo la Dictadura... ! ¡Basta ya...! ¡Muera el dictador León Panzón ...! ¡Ni un paso atrás... Ni un pas	
8 2001 C.Rica.	mis ojos en el cielo raso color cereza y comencé a buscar en mi memoria un mambo retozón , pero Ernesto me ganó la partid.	
9 2001 C.Rica.	gesto de comodidad que prometía una velada placentera. Su perfume no era brusco ni dulzón ; más bien invitaba a padecerlo l	
10 2001 C.Rica.	casa silbando una melodía muy conocida por aquellos días. El cartero era un tipo cabezón que se empeñaba en destacar l	
11 2001 Par.	que ahora le mojaba la remera, y la chaqueta, y que le daba una especie de sopor dulzón , lo que le vino incomodando des	
12 2001 Guat.	un circo, dejando atrás a toda su familia. Se fue con un circo, me dijeron, así, panzón y viejo, y que lo habían visto via	

Figura 7. Captura de pantalla de la búsqueda de *zón como adjetivo en CORPES XXI

A partir de algunas frases reales escogidas del corpus, se puede diseñar una actividad en la que se le pida al estudiante que escoja en cuáles de las frases –zón es un sufijo y en cuáles no. Un ejemplo de cómo sería esta actividad se

¿Cuándo es –zón sufijo y cuándo no lo es?

- El chófer era un tipo de pelo corto, rubio y algo **cabezón**.
- Nos gustaría mandarte los comentarios de niños y jóvenes a los que tocas el **corazón**.
- Él siempre tiene la **razón**.
- De una mesa se levantó un tipo **panzón**.
- Su perfume no era brusco ni **dulzón**.
- En esa calle hay un **buzón** de correos.

encuentra en la Figura 8:

Figura 8. Actividad creada a partir del corpus inspirada en González García (2018)

Además, se podría trabajar qué matiz aporta el sufijo –zón a algunas palabras como “dulzón” (olor desagradable) o panzón, cabezón y narizón (gran panza, cabeza o nariz).

Bibliografía

Buyse, K. 2017. “Los corpus como herramienta de aprendizaje del léxico”. En F. Herrera y N. Sans (eds.), *Enseñar léxico en el aula de español* (pp. 121-140). Barcelona: Difusión.

González García, V. 2018. Webinario “Corpus para tus clases de ELE”. Difusión. Disponible en: <https://www.difusion.com/tienda/corpus-clase/> [Consultado el 15/9/2021]

Pitkowski, E. F., & Gamarra, J. V. 2009. “El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE”. *Tinkuy: boletín de investigación y debate*, 11, 31–51.

Corpus

CORPES XXI, disponible en: <https://www.rae.es/banco-de-datos/corpes-xxi>

CORDE, disponible en: <https://www.rae.es/banco-de-datos/corde>

CAES, disponible en: <https://galvan.usc.es/caes>

SPLLOC, disponible en: <http://www.splloc.soton.ac.uk/>