

CORPUS DIACRÓNICO DEL ESPAÑOL (CORDE)

José María Echazarreta Arzac

I.E.S. Joaquín Araújo

Fuenlabrada

INTRODUCCIÓN A LOS CORPUS

Los corpus, en general, se definen como el «conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación» (RAE, 2017).

Un corpus lingüístico se basa en un amplio conjunto de textos, desde ensayos a transcripciones radiofónicas, obras de teatro o guiones de cine, donde se almacena un conjunto aún más extenso de términos, y que sirven para recopilar lo que se escribe o se ha escrito en una lengua en un momento en concreto.

Para la investigación de una lengua, es necesario recurrir a los corpus lingüísticos si queremos conseguir ejemplos reales. A partir de ellos podemos analizar el uso de determinadas palabras que están dentro de una oración, conocer la frecuencia de uso de un término en concreto, entre otras muchas opciones.

El corpus se ha convertido en el método de investigación de la lengua más importante gracias a su ingente tamaño, la gran fiabilidad y lo fácil que es acceder a él, gracias a su digitalización.

Los datos recogidos en un corpus pueden ser escritos, muestras orales (generalmente transcritas), o mixtos. Lo más común es el uso de datos escritos por la dificultad que suponía la transcripción y la pérdida de datos que esta técnica conllevaba.

Este tipo de corpus tiene un carácter abierto, es decir, está siendo actualizado continuamente: se introducen nuevos textos y no tiene un límite establecido. Además, en los corpus textuales se amplía el repertorio de las lenguas que disponen de corpus y se elaboran, además de corpus monolingües (como el CREA o el CORGA), corpus de tipo multilingüe, que pueden ser comparables: es el mismo texto dispuesto en distintas lenguas; y alineados. Con ellos se consigue mostrar la traducción en otra lengua de un modo coherente.

Ejemplos de corpus ingleses

Uno de los ejemplos más importantes de corpus textuales con las características antes mencionadas es *The Bank of English*, también conocido como proyecto COBUILD, compuesto por más de quinientos millones de palabras procedentes del inglés moderno tanto de manera escrita como de manera oral del inglés británico, americano, canadiense y australiano. Otro corpus de datos ingleses es el conocido *British National Corpus* (BNC), compuesto por unos cien millones de palabras de origen británico tanto escrito como hablado.

Ejemplos de corpus españoles

El Banco de Datos de la Real Academia Española está constituido por dos grandes corpus: el Corpus de Referencia del Español Actual y el Corpus Diacrónico del Español. Estos dos corpus son complementarios, ya que el CREA agrupa textos desde 1974 hasta 2004, mientras que el CORDE se ocupa de todo lo anterior. La manera de agrupar los datos se basa en clasificar los textos que entran en el periodo que caracteriza al CREA, y todos los demás datos que quedan fuera de este pasan a formar parte de CORDE.

Se utiliza CORDE para textos hasta 1974 y datos correspondientes a aquellos criterios de selección de textos que no comparte con CREA: geográficamente, por ejemplo, textos producidos en Filipinas o textos escritos en judeoespañol. Además, conviene resaltar que en CORDE se han regularizado las grafías en los títulos de textos, cosa que ayuda a que la búsqueda en la aplicación sea más fácil y rápida, pero se mantiene el título original de la obra en el resultado de la consulta. Los corpus que forman el Banco de Datos son la fuente de documentación principal que utiliza el Instituto de Lexicografía de la Real Academia Española para preparar diferentes materiales. El Banco de Datos ha servido para redactar el Diccionario del Estudiante, el Diccionario Panhispánico de Dudas, el Diccionario Esencial, o la Nueva Gramática de la Lengua Española.

En cuanto a los corpus de origen español el Corpus de Referencia del Español Actual (CREA) es el más destacable. Es una herramienta que facilita la búsqueda de palabras españolas actuales utilizadas en medios de comunicación escritos. Este buscador cita las fuentes donde se encuentra el término o términos y los contextualiza en las fuentes encontradas. Posee algo más de 160 millones de formas y tiene una amplia variedad de textos escritos y orales, procedentes de todos los países de habla hispana desde el año 1975 hasta el año 2004. En este corpus se incluyen textos escritos de libros, periódicos y revistas, y transcripciones de documentos sonoros de radio y televisión principalmente.

INTRODUCCIÓN AL CORPUS DIACRÓNICO EL ESPAÑOL (CORDE)

El CORDE es un corpus textual que recoge datos de todas las épocas y lugares donde se habló el español, contando desde los inicios del idioma hasta el año 1974, cuando entra en contacto con el Corpus de Referencia del Español Actual, mencionado antes.

El Corpus Diacrónico del Español es un banco de datos de la lengua española que contiene unos 260 millones de palabras provenientes de textos escritos de diferentes tipos.

El problema principal que afrontó la RAE fue determinar los materiales que le servían de base para la redacción. Los materiales que se utilizaban antes de la creación del CORDE eran unas 13 millones de fichas recopiladas a lo largo de tres siglos, pero a veces defectuosas e imperfectas. Para evitar esto, se creó este corpus, que facilitó la creación del Diccionario histórico de la lengua española.

Es una herramienta muy útil para el análisis o estudio de la lengua española desde sus orígenes y sus modalidades según la época y el lugar en el que se encuentre. De él se puede extraer información para analizar los significados de las palabras, y el uso de su gramática a través del tiempo. Este corpus agrupa textos de géneros muy variados, distribuidos en prosa y verso. Además, dentro de estos grupos, se puede hacer otra clasificación que separa los textos

narrativos de los líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc.

En 1994, la RAE utilizó las nuevas tecnologías de la información para desarrollar un banco de datos que mejorara el estado de sus materiales de trabajo y su acceso, haciendo más rápido y eficaz su uso. La Real Academia Española utiliza el CORDE para documentar palabras, concretar si esas palabras no se usan o son anticuadas, teniendo en cuenta el origen de los términos, la tradición que tienen en su lengua, entre otros aspectos.

El CORDE es el mayor conjunto de registros léxicos de la historia del español.

CARACTERÍSTICAS DEL CORPUS DIACRÓNICO DEL ESPAÑOL (CORDE)

El Corpus Diacrónico del Español presenta una serie de características que hacen que destaque:

- El CORDE quiere ser representativo de la lengua española a lo largo de la historia, recopilando testimonios de todas las épocas y lugares en que se habló el español, desde los inicios de esta lengua hasta nuestros días.
- El catálogo de obras que forman el CORDE está compuesto por 5500 títulos, de los cuales 1000 son anónimos y el resto corresponde a 4500 autores conocidos.
- Las obras que lo componen han sido seleccionadas siguiendo unos criterios: en un principio se introducen los textos íntegros. Se escogen textos representativos por la difusión o la influencia que han tenido a lo largo de la historia.
- Los documentos que aparecen en el corpus son escritos, pero no exclusivamente literarios, sino que incluye también documentos o textos de origen científico.
- La estructura de CORDE está basada sobre todo en criterios formales. En la siguiente tabla se muestra la distribución de textos dentro del corpus:

Literarios 44%	No literarios 56%
Verso 10%	Didáctica 10%
Prosa narrativa 27%	Ciencia y Técnica 14%
Prosa dramática 7%	Religión 6%
	Sociedad 8%
	Historia 9%
	Jurídica 6%
	Prensa 3%

- En cuanto a la clasificación de los textos desde el punto de vista geográfico un 74% procede del español de España, mientras que un 25% procede del español hablado en América. El 1% que resta pertenece a datos procedentes de textos judeoespañoles.
- En cuanto a la clasificación de los textos desde el punto de vista cronológico está dividido en grandes periodos temporales, y subdividido a su vez en tramos menores. Está claro que aunque el corpus quiere ser representativo por partes iguales de cada época, no todos los periodos tienen la misma cantidad de textos, por lo que, por ejemplo en la época contemporánea, el número de materiales aumenta:

1. Edad Media (16,5%)
 - a. Orígenes – 1250
 - b. 1251 – 1491
2. Siglos de Oro (30,5%)
 - a. 1492 – 1598
 - b. 1599 – 1712
3. Época Contemporánea (53%)
 - a. 1713 -1812
 - b. 1813 – 1898
 - c. 1899 – 1939
 - d. 1940 – 1974

USOS Y APLICACIONES DEL CORPUS DIACRÓNICO DEL ESPAÑOL (CORDE)

Al realizar una búsqueda, CORDE nos facilita la información sobre el origen, vigencia y desaparición de una palabra, a la vez que nos ayuda a saber, según el contexto en el que aparezca, el significado de esa palabra.

Al introducir en el buscador de CORDE una palabra, nos aparecen diferentes celdas en las que elegir el filtro que deseemos aplicar a la búsqueda. Entre los criterios de selección aparece el autor, la obra, la época, el medio (donde podemos elegir que la búsqueda se limite a libros, periódicos, revistas, miscelánea, orales...), la geografía (donde aparecen los distintos lugares en los que se habla el español) y el tema (donde encontramos categorías como la lírica, la narrativa, el teatro; y dentro de estas, subgrupos más pequeños), para hacer acotar el margen de búsqueda y llegar así más fácilmente a lo que se desea encontrar.

Más abajo nos sale la opción **Consulta CREA**, herramienta que nos dirige directamente al Corpus de Referencia del Español Actual de la Real Academia Española.

También aparece la opción **Nómina de autores y obras**, herramienta que nos permite buscar una obra en concreto añadiendo su autor y seleccionado si queremos buscarlo en CREA o CORDE, además de darnos la opción de elegir la cronología de la obra, el medio (libros, prensa o miscelánea en CREA), y la geografía.

Una vez dentro del Corpus Diacrónico del Español, nos aparece esta ventana de búsqueda:



Real Academia Española - Corpus Diacrónico del Español (CORDE)

Consulta:	piano		
Criterios de selección:			
Autor:		Obra:	
Cronológico:		Medio:	Geográfico:
		(Todos) Libros Periódicos Revistas Miscelánea Orales	(Todos) Argentina Bolivia Chile Colombia Costa Rica
Tema:	1221.- Relato extenso novela y otras formas similares 1222.- Relato extenso diálogo y miscelánea 1223.- Otras formas 1229.- Otros 13.- Teatro 131.- Breve 132.- Extenso		

Buscar Limpiar

[Consulta CREA](#) [Nómina de autores y obras](#) [Cómo citar el CORPUS](#) [Ayuda.](#)

ISSN 2340-5651 Corpus diacrónico del español

En este caso, la palabra que se ha consultado es *piano*, y dentro de las categorías que puedo elegir para filtrar mi búsqueda he escogido el teatro. Al presionar el botón **Buscar**, el resultado dice que hay 57 casos en 22 documentos de la palabra piano en teatro.



Resultado de la consulta al banco de datos

Cómo citar el CORPUS

Consulta:	<i>piano</i> , en todos los medios, en CORDE, en Teatro
Resultado:	57 casos en 22 documentos.

Ver estadística

Filtros:	Casos
Ratio:	10

☐ Mantener documentos (Solo para filtro sobre casos).

Filtrar

OBTENCIÓN DE EJEMPLOS

Recuperar	Concordancias	Normal.
Clasificación:		
Agrupación:		Marcas:

Nueva consulta: [CREA](#) [CORDE](#) [Nómina de autores y obras](#) [Ayuda.](#)

En la búsqueda se puede comprobar el uso de la palabra seleccionada. Por ejemplo, en el caso de la palabra *piano* en teatro, en España hay 28 casos, lo que se resume en un 49,12% de los datos encontrados con esas características.



Resultados con estadísticas (RAE)

Consulta:	piano, en todos los medios, en CORDE, en Teatro
Resultado:	57 casos en 22 documentos.

Filtros: Casos
Ratio: 10

☐ Mantener documentos (Sólo para filtro sobre casos).

Filtrar

OBTENCIÓN DE EJEMPLOS

Recuperar

Concordancias
Normal.

Clasificación:
Agrupación:
Marcas:

Cómo citar el CORPUS

Estadísticas

Año	%	Casos	País	%	Casos	Tema	%	Casos
1949	23.52	12	ESPAÑA	49.12	28	13.- Prosa dramática	100.00	57
1951	21.56	11	MÉXICO	33.33	19			
1932	13.72	7	CUBA	8.77	5			
1955	13.72	7	ARGENTINA	3.50	2			
1906	5.88	3	COLOMBIA	1.75	1			
1831	3.92	2	EL SALVADOR	1.75	1			
1911	3.92	2	GUATEMALA	1.75	1			
1916	3.92	2						
1833	1.96	1						
Otros	7.84	4						

Por último, otra de las herramientas que facilita este corpus, es la obtención de ejemplos en los que aparece la palabra buscada y todos los datos del ejemplo que quieras seleccionar:

Recuperar Concordancias. Normal. Clasificación: Agrupación: Marcas: CORDE

Cómo citar el CORPUS

Concordancias.

Pantalla: 2 de 3. Siguiente 1 2 3 Ver párrafos

Nº	CONCORDANCIA	AÑO	AUTOR
26	ado yo pagando cuatro años seguidos el maestro de piano; hija mía de mis entrañas, ¿de qué te sirve hab	** 1831	Larra, Mariano José de
27	e Oshidori será una fantasía morisca para canto y piano... Sergio Nada de fantasías moriscas, Adelaida.	** 1932	Jardiel Poncela, Enriq
28	a! ¡Se me ha puesto convidaros a todas! Si no hay piano, se busca. ¡Aquí se responde con cartera! ¡Madr	** 1926 - 1930	Valle-Inclán, Ramón Ma
29	busto sobre la diversa botillería. El Café tiene piano y violín. Las sombras y la música flotan en el	** 1920 - 1924	Valle-Inclán, Ramón Ma
30	ENOS ACTO PRIMERO En un café madrileño, de los de piano y violín. Antes de levantarse el telón se oye l	** 1911	Benavente, Jacinto
31	o un Joven. Durante esta escena se oirá tocar muy piano la polkita de que se habla anteriormente	** 1911	Benavente, Jacinto
32	MERA Eleonora Gaddi y Alejandro. Ella, sentada al p	** 1911	Castellón, José
33	lante. Una pausa. En la escena solitaria óyese un p	** 1911	Carlos Felipe (Carlos
34	tumbre, la fuente.) raimundo ¡Es ella! Estudia su p	** 1911	Carlos Felipe (Carlos
35	ace tarde y Lila se irá a dormir. Ya no se oye el p	** 1911	Carlos Felipe (Carlos
36	ndo hajito con la que tenía al lado. Depué tocó e piano un señó, l'aminitrado de Correo. Depué cantó la	** 1949	Jardiel Poncela, Enriq
37	eral y en el ángulo del tercer término se alza un piano de cola y cuatro atriles musicales. El mobili	** 1940	Álvarez Quintero, Sera
38	ecedoras de rejilla; donde estaba el arcón hay un piano; por doquiera hay plantas y flores; en los arco	** 1906	Álvarez Quintero, Sera
39	mi doncella, con mis flores, con el loro, con el piano, con la perrita y con toda la balumba que conmi	** 1906	Álvarez Quintero, Sera
40	legria! (Hace sonar distraídamente las teclas del piano.) Julio Oye. Consolación ¿Qué quieres? Julio ¿P	** 1906	Neville, Edgar
41	ro Te diré. Julián Ya sé que no podría llevar un piano de un lado para otro; pero también de jóvenes,	** 1952	Gorostiza, Manuel Edua
42	qué me he de hacer yo entretanto, sin libros, sin piano...? DON EDUARDO En efecto, no tienes hoy mucho	** 1833	Ibargüenoitia, Jorge
43	e cobre y otras chácharas, una cómoda antigua, un piano viejo vertical y una mecedora. En la pared, bla	** 1955	Ibargüenoitia, Jorge
44	era chiquita, hicimos el sacrificio de comprar el piano y pagar maestra. Estudió dos años. Así que toca	** 1955	Ibargüenoitia, Jorge
45	espíritu con su música. Bueno, pues nunca abre el piano. Ya se le olvidó lo que sabía, sólo toca cuando	** 1955	Ibargüenoitia, Jorge
46	o como si el otro fuera a la guerra. Se sienta al piano y toca los primeros acordes de la gavota Pavlov	** 1955	Ibargüenoitia, Jorge
47	en. Clotilde va a la terraza a verlos irse. Va al piano y toca la gavota Pavlova, muy impaciente. Anton	** 1955	Ibargüenoitia, Jorge
48	Hilarión tocaba el chelo. Clotilde Pue yo toco el piano. Abre el piano y se sienta en el teclado con es	** 1955	Ibargüenoitia, Jorge
49	el chelo. Clotilde Pue yo toco el piano. Abre el piano y se sienta en el teclado con estruendo. Nicola	** 1955	Ibargüenoitia, Jorge
50	Nunca me dijo una palabra de reproche, pero aquel piano cerrado, aquel sillón vuelto de espaldas a la v	** 1949	Casona, Alejandro

Ir arriba Pantalla: 2 de 3. Siguiente 1 2 3 Ver párrafos

Nueva consulta: CREA CORDE Nómina de autores y obras Ayuda.

BIBLIOGRAFÍA

- <http://articles.unono.net>
- https://es.wikipedia.org/wiki/Corpus_ling%C3%BC%C3%ADstico
- <https://recursosdelengua.wordpress.com>
- DialNet, *“El Banco de Datos de la Real Academia Española: CREA Y CORDE”*, de Mercedes Sánchez Sánchez y Carlos Domínguez Cintas.