# Lingüística de corpus y español como L2/LE: de la investigación al aula

Mar Cruz Piñol



III Jornadas internacionales sobre investigaciones lexicográficas y lexicológicas



### https://sites.google.com/view/corpus-ele-jaen-2019/



#### Referencias

### Corpus y Webs

"Lingüística de corpus y español como L2/LE: de la investigación al aula" (InLéxico 2019)

Mar Cruz Piñol (abril 2019)

https://sites.google.com/site/marcruzpinolub/

## ¿Qué es un corpus?

## corpus<sup>2</sup>

Del ingl. corpus, y este del lat. corpus 'cuerpo'.

1. m. Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación.

Real Academia Española © Todos los derechos reservados

"A corpus is a collection of pieces of text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2004)

## Objetivos de la LC

[Torruella y Llisterri, 1999]

- "La función principal de un corpus, tanto textual como oral, es establecer la relación entre la teoría y los datos: el corpus tiene que mostrar, a pequeña escala, cómo funciona una lengua natural"
- (los corpus) "proporcionan bases mucho más reales para el estudio de las lenguas que los métodos intuitivos tradicionales"
- "la finalidad última [... es] entender mejor cómo funciona el lenguaje humano"

## Beneficios del trabajos con corpus

- Verificación objetiva
- Oraciones en contexto
- Muestras reales
- Comprobación de teorías
- Reglas a partir de la observación (inductivo)
- Información cuantitativa
- Gran volumen de datos

## **Conceptos clave**

- Ocurrencias
- Concordancias
- Frecuencias
- Representatividad
- Lemas / Formas / Variantes
- Anotación

## Ocurrencias

### Concordancias

#### OBTENCIÓN DE EJEMPLOS

Recuperar	Concordancias.	Clasificación:	<b>○</b>
	Normal.		•
Agrupación:		Marcas:	<b></b>

Cómo citar el CORPUS

Concordancias.

Pantalla: 1 de 22. Siguiente 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 Ver párrafos

#### CONCORDANCIA

gados de la CEAR presentaron un recurso de habeas corpus, que no fue aceptado. ienso en que el editor de Fonollosa escogió de un corpus poético inmenso, escrito a lo largo de toda su todas las vueltas y vaivenes de recursos y habeas corpus, muertes reiteradas. Personaje sin desperdicio bra tiene una voluntad de convertirse en obligado corpus informativo del especialista en el tema, más q ticia los expedientes relacionados con los hábeas corpus otorgados a los militares involucrados en los ción a la jurisdicción a la que compete el hábeas corpus y la jurisprudencia y la doctrina nos dan la r la justicia militar y la ordinaria, por el hábeas corpus que el juez Braulio Sánchez otorgó a los jefes stra propia historia. Así existe, por ejemplo, un corpus importante de novela picaresca latinoamericana ron que había "desaparecido". Presento el "habeas corpus". Un miembro de la Suprema Corte de Justicia m dictador presentaron un recurso de amparo (habeas corpus) al tribunal de apelaciones y recordaron un di egal. "En mi concepto, dijo, el recurso de habeas corpus interpuesto por los abogados de Ochoa sólo pod udio, ni pericia, se invocó el recurso de "habeas corpus", incorporado al nuevo Código de Procedimiento por sus fisuras normativas. El recurso de "habeas corpus" es inobjetable, democrático y de la más pura rte consistencia legal, la aplicación del "Habeas corpus". El montaje comenzó a marchar y se le fijó un s concedido a los sindicados un recurso de hábeas corpus, gracias al cual recobraron la libertad el pas rlando Bustos, para conceder el recurso de hábeas corpus que permitió la libertad del procurador, Orlan cisamente (L'écriture didascalique). A partir del corpus teórico antes comentado (y ampliado con algún ntido estricto, es pues lícito que se apoye en un corpus teórico, que lo haga suyo, sin tener que profu análisis de los componentes espaciales del amplio corpus seleccionado por Georges Zaragoza así lo ponen completan el libro. Por lo que hace referencia al corpus escogido este está formado por trece obras fra cir, y resumiendo, que Zaragoza ha constituido su corpus por lo que al apartado español toca, con gener nte. A Jesús Rubio corresponde la elaboración del corpus documental, así como la articulación -a todas os (con textos o, si se requiere, con determinado corpus documental) es particularmente idónea cuando s damentales. 1. Análisis cualitativo inicial de un corpus de mensajes representativo del tipo de proceso De la acumulación de datos clínicos ha surgido un corpus de conocimientos, que ha dado base a diversas

### Concordancias

Antecedentes:

→ Rojo, 2015

→ Bolaños, 2015

"(Las concordancias) no solo garantizan el uso y frecuencia de una unidad léxica determinada, sino que esta aparece **contextualizada** y clasificada en géneros textuales, períodos históricos, zonas geográficas, etc."

[Alonso, 2007, p. 19]

## Frecuencias

Palabra	Parte Oración	Frec.	Palabra	Parte Oración	Frec.	Palabra	Parte Oración	Frec.
el	Artículo	188.833	hacer	Verbo	7.629	entre	Preposición	2.811
de	Preposición	119.766	poder	Verbo	7.490	dos	Pronombre	2.764
у	Conjunción	54.644	me	Pronombre	7.067	primero	Pronombre	2.752
que	Conjunción	53.905	pero	Conjunción	6.943	mismo	Pronombre	2.746
en	Preposición	51,338	le	Pronombre	6.933	nos	Pronombre	2.703
а	Preposición	43.173	ir	Verbo	6.269	deber	Verbo	2.671
un/uno	Artículo	01001100	otro	Pronombre	507.00	alguno	Pronombre	100,734
ser	Verbo	41.717	yo	Pronombre	5.274	sólo	Adverbio	2,665
se	Pronombre	34.839	si	Conjunción	4.933	sí	Adverbio	2.613
no	Adverbio	25.508	mi	Pronombre	4.916	hasta	Preposición	2.587
del	Artículo	23.643	año	Sustantivo	4.254	pasar	Verbo	2.532
	Verbo	22.249	1000	Adverbio	4.225	P. Constant	Preposición	2.458
haber		20.330	ya	0.120.2020	4.190	desde	(9,845,983,898,8)	2.454
por	Preposición	20.293	dar	Verbo	4.048	bien	Adverbio	2.434
con	Preposición	18.645	ver	Verbo	3.967	país	Sustantivo	2.309
que	Pronombre	17.895	porque	Conjunción	3.803	pues	Conjunción	2.195
su	Pronombre	17.673	cuando	Conjunción	3.736	así	Adverbio	2.191
lo	Pronombre	14.303	sin	Preposición	3.698	nuestro	Pronombre	2.171
para	Preposición	13.259	muy	Adverbio	000000	ni	Conjunción	200000
al	Artículo	MATERIA.	mucho	Adjetivo	3.547	llegar	Verbo	2.163
este	Pronombre	11.719	sobre	Preposición	3.352	grande	Adjetivo	2.151
como	Adverbio	11.346	saber	Verbo	3.255	donde	Pronombre	2.142
estar	Verbo	10.990	vez	Sustantivo	3.204	usted	Pronombre	2.114
tener	Verbo	10.343	bueno	Adjetivo	3.143	nuevo	Adjetivo	2.108
azeroum.	Verbo	10.094	2000000000000	Verbo	3.129	1000000000	Sustantivo	2.097
decir	1470.5055	8.959	querer		3.100	cosa		2.094
100 march 100 ma	Adverbio	8.506	también	Adverbio	3.038	hablar	Verbo	2.075
más		0.000						
más ese	Pronombre	8.260	aquel	Pronombre	2.983	creer	Verbo	2.044

de *Eduteka* (basado en corpus *Cumbre*)

### **Variante**



### trasatlántico, ca

V. transatlántico.

## Representatividad

• "(...) una muestra **representativa** que permita al lingüista fundamentar su investigación en datos objetivos. Un corpus no puede identificarse con la lengua, sino que es un conjunto de datos que la **representa** de una manera más o menos fiable."

[Rafel y Soler, 2003, p. 42]

"(...) todas las muestras son, de algún modo, tendenciosas.
 (...) En todo momento, los investigadores se tienen que cuestionar cómo fueron obtenidas las muestras y hasta qué punto pueden ser válidas las conclusiones que de ellas se han extraído."

[Torruella y Llisterri, 1999, p. 63]

España: 50% América: 50%

Escritos: 90%

Orales: 10%

## Representatividad

Distribución temporal de los textos del CREA				
1975-1979	10%			
1980-1984	15%			
1985-1989	20%			
1990-1994	25%			
1995-1999	30%			

http://corpus.rae.es/ayuda c.htm# Toc30228257

# Corpus del Español del Siglo XXI (CORPES)

Versión beta

:/parametros-de-seleccion-de-textos

#### Medio

El 90 % de los textos corresponde a la lengua escrita y el 10 % a la lengua oral.

Los materiales escritos proceden en un 40 % de libros, publicaciones periódicas (40 %), material de Internet (7,5 %) y miscelánea (2,5 %).

#### Cogranico

La distribución general del CORPES asigna un 30 % del total a formas procedentes de España y un 70 % a formas procedentes de América.

## Lema - forma - variante

Forma	Lema	Variante
cantó	cantar	cantar
trasatlánticos	transatlántico	trasatlántico

### La web es un corpus...

### Lema



#### rendir - Definición - WordReference.com

www.wordreference.com/definicion/rendir \*

rendir - Significados en español y discusiones con el uso de 'rendir'.

#### rendir - sinónimos y antónimos - WordReference.com

www.wordreference.com/sinonimos/rendir -

rendir - sinónimos de 'rendir' en un diccionario de 200,000 sinónimos online.

#### Rendir - Real Academia Española. Diccionario Usual.

lema.rae.es/drae/srv/search?val=rendir -

rendir. (Del lat. redděre, infl. por prenděre y venděre). 1. tr. Vencer, sujetar, obligar a las tropas, ... Rendir gracias, agradecer; rendir obsequios, obsequiar.

#### rendir - significado de rendir diccionario

es.thefreedictionary.com/rendir - Traducir esta página

Definición de **rendir** en el Diccionario de español en línea. Significado de **rendir** diccionario. traducir **rendir** significado **rendir** traducción de **rendir** Sinónimos de ...

#### Rendir cuentas | Opinión | EL PAÍS

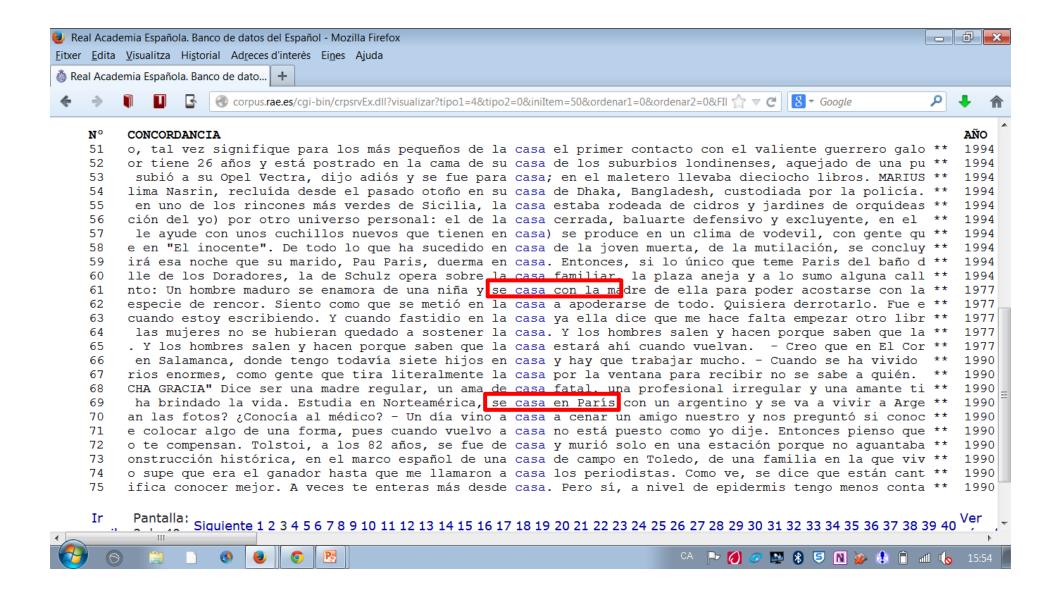
elpais.com > Opinión \*

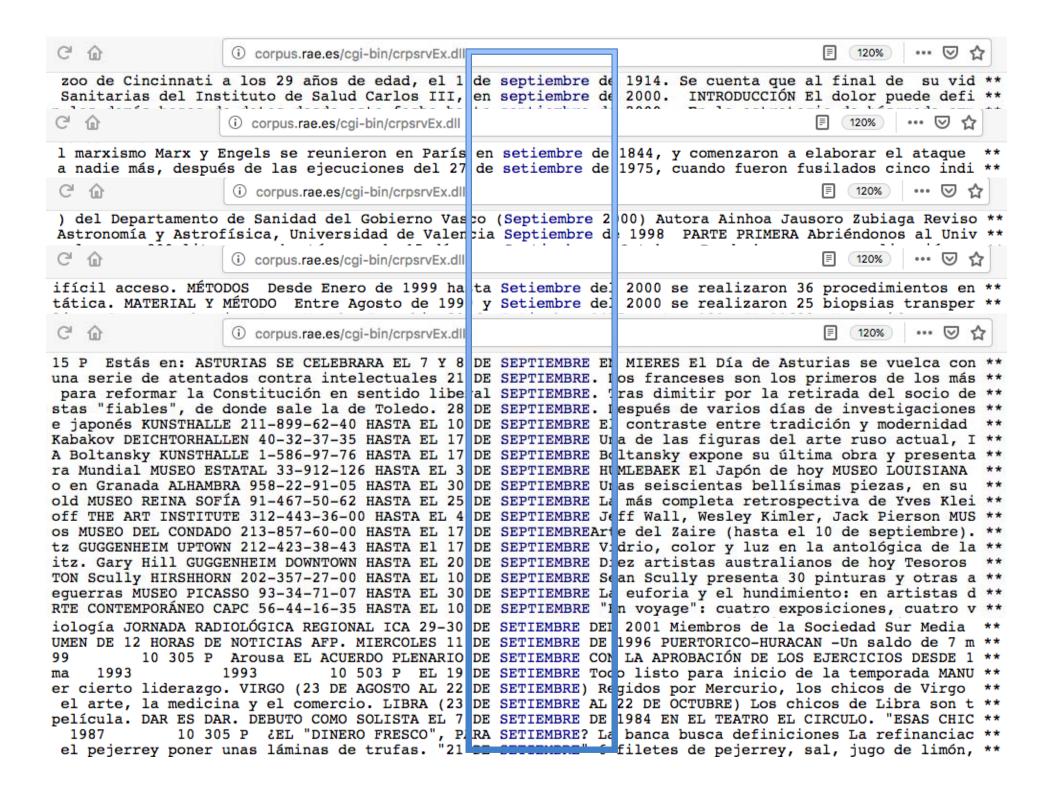
8 de feb. de 2015 - Los simpatizantes de los nuevos partidos deben exigir responsabilidades a sus líderes.

... sin lematizar

"(...) quien sabe extraer información lingüística de Google tendrá el camino más fácil cuando intente manejar un verdadero programa de concordancias como el CREA"

[Sala Caja, 2004, p. 141]







1000 formas más frecuentes

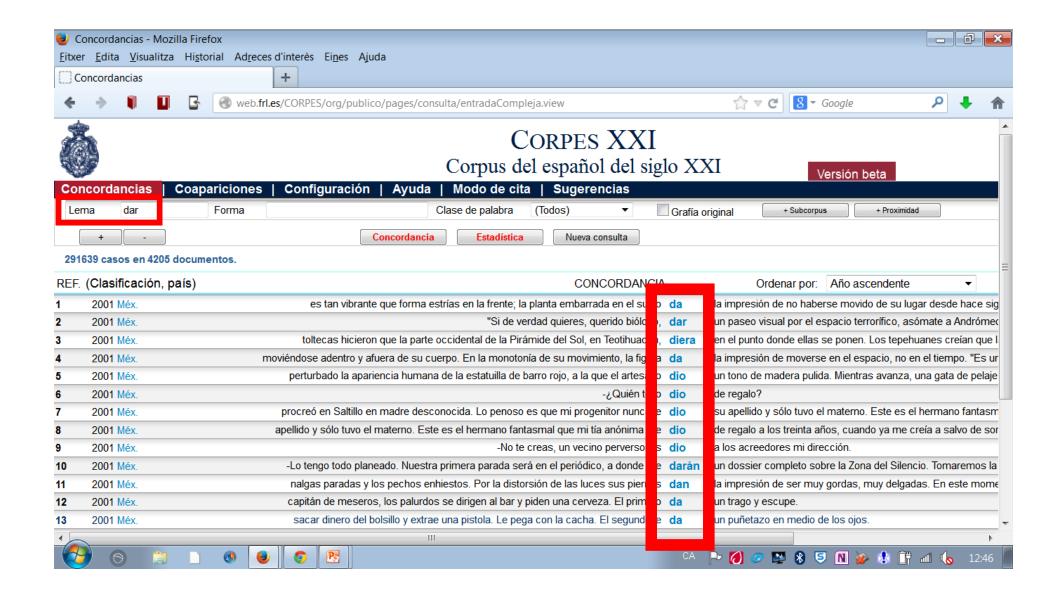
5000 formas más frecuentes

10000 formas más frecuentes

Lista total de frecuencias

Orden	Frec.absoluta	Frec.normalizada
1. de	9,999,518	65545.55
2 la	6,277,560	41148.59 (casos por
3. que	4,681,839	30688.85 millón)
4 el	4,569,652	29953.48
5. en	4,234,281	27755.16
6. y	4,180,279	27401.19
7. a	3,260,939	21375.03
8 los	2,618,657	17164.95
9. se	2,022,514	13257.31
10. del	1,857,225	12173.87
11 las	1,686,741	11056.37
12. un	1,659,827	10879.95
13. por	1,561,904	10238.07
14. con	1,481,607	9711.74
15. no	1,465,503	9606.18
16. una	1,347,603	8833.36
17. su	1,103,617	7234.06
18. para	1,062,152	6962.26
19. es	1,019,669	6683.79
20. al	951,054	6234.03
21 10	866,955	5682.77
22. como	773,465	5069.96
23. más	661,696	4337.33
24. 0	542,284	3554.60
25. pero	450,512	2953.04
26. sus	449,870	2948.84
27. le	413,241	2708.74
28. ha	380,339	2493.07
29. me	374,368	2453.93
30. si	327,480	2146.58
31. sin	298,383	1955.86
A STATE OF THE STA		

### (corpus LEMATIZADOS)

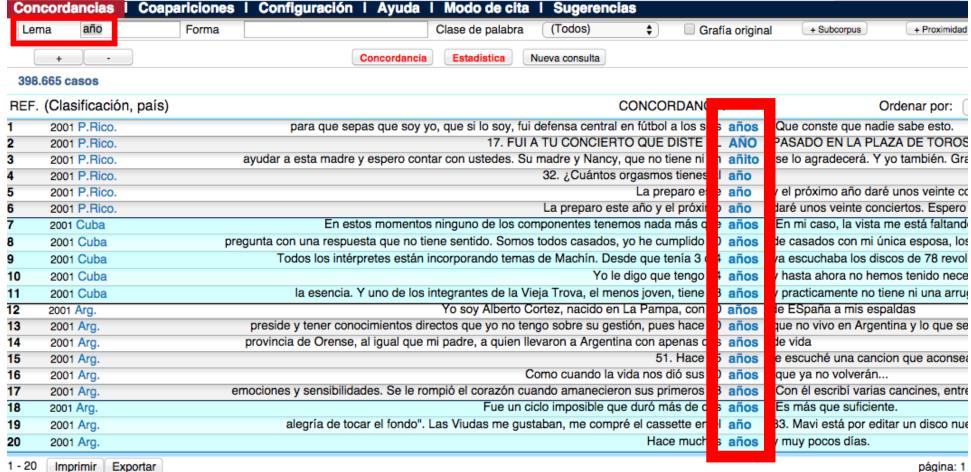






# CORPES XXI Corpus del español del siglo XXI

Versión beta (0.7)







VER CONTEXTO: HACER CLIC EN LA PALABRA (TODAS LAS SECCIONES), N�MERO (UNA SECCIÏ¿½N), O [CONTEXTO] (VARIAS) [AYUDA...]

	CONTENTO	FREC	
1	PONER	538625	
2	PONE	199668	
3	PUESTO	139788	
4	PUSO	128158	
5	PONEN	97754	
6	PONIENDO	62897	
7	PONGO	62138	
8	PUSE	51997	
9	PONGA	51586	
10	PUSIERON	44619	
11	PONEMOS	36802	
12	PONGAN	33888	<b>-</b>
13	PON	28757	
14	PONES	28470	
15	PUESTA	26089	•

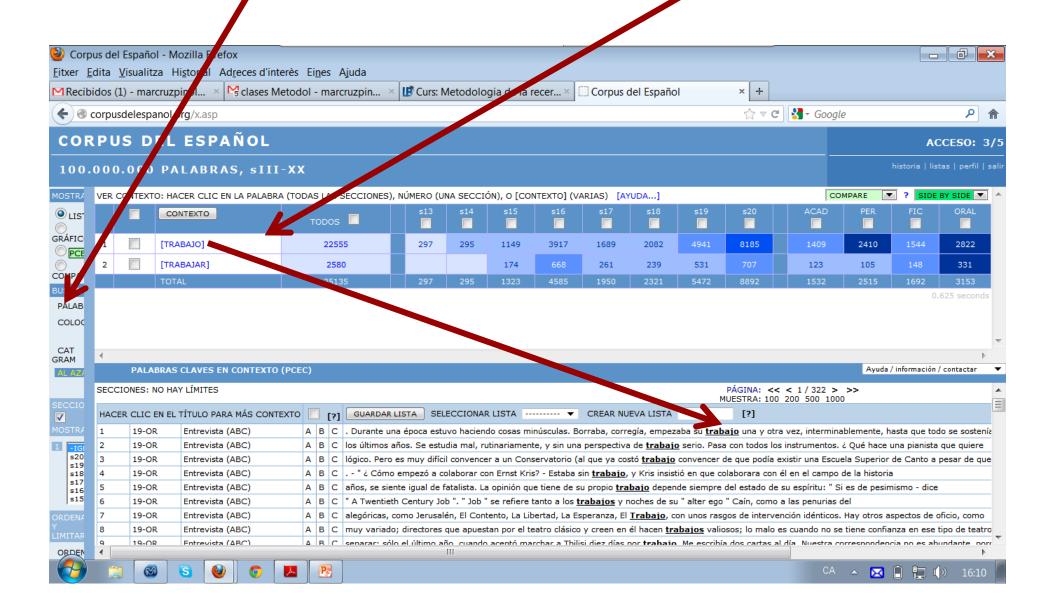
BUSCAR

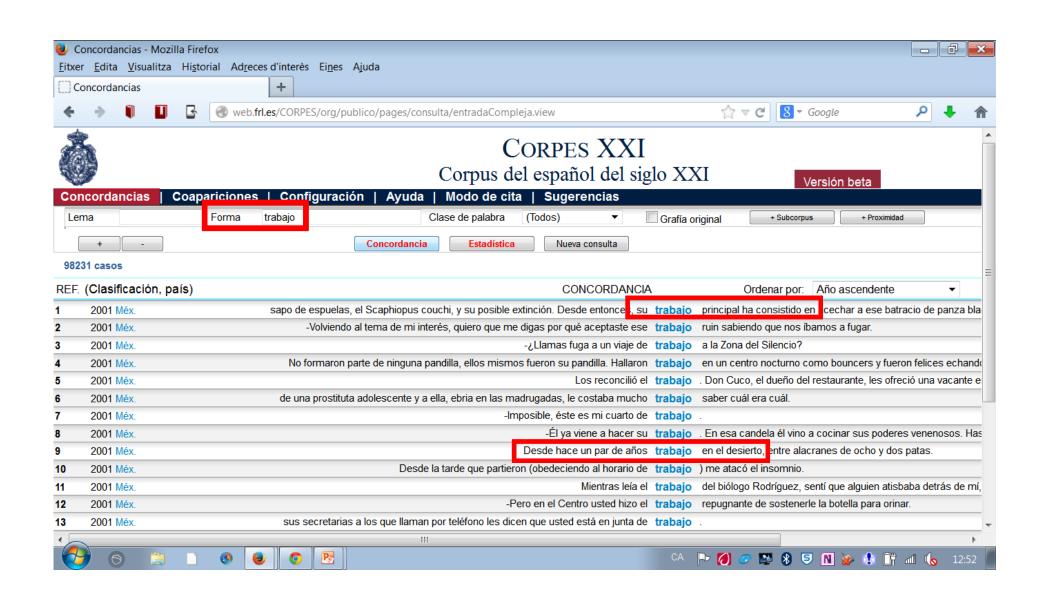
PALABRA(S)

trabajo

AGRUPAR POR

LEMAS





## Anotación













#### 4. 4. Sintaxis de búsqueda más intuitiva y sencilla

En nuestra opinión, algunos sistemas de sintaxis para las búsquedas son innecesariamente complejos, como el sistema CQP a la izquierda. La antigua sintaxis de búsqueda de BYU era mucho más simple, pero aún había demasiados corchetes, puntos, asteriscos, etc., difíciles de escribir con el teclado de teléfono móvil. Hemos simplificado la sintaxis de búsqueda aún más, como se muestra a la derecha. Los usuarios pueden utilizar cualquier combinación de la antigua y la nueva sintaxis. (Para más información, incluyendo los nuevos códigos para clases de palabras, haga clic en LIST [LISTA] en el formulario de búsqueda de un corpus y después en Part of Speech [Clase de palabra].)

Tipo de búsqueda	Sintaxis CQP	Sintaxis de BYU antigua	lueva sintaxis	Ejemplo
Palabras	[word = "nooks"]	nooks	ooks	nooks and crannies
Lema (formas de las palabras)	[lemma = "decide"]	[decide]	ECIDE	DECIDE that it
Clase de palabra	[tag = "NN."]	[nn*]	IOUN	fast NOUN
Sinónimos	No disponible	[=soft]	soft	soft, smooth, quiet
Listas de palabras personalizadas	No disponible	[emailAddress@clothes]	clothes	dress, shoe, sock
Combinación de los anteriores	[lemma = "end" & pos = "VV."]	[end].[v*]	ND_v	end, ends, ended, ending
Combinación de los anteriores	[lemma = "eat"] [tag = "NN."]	[eat] * [nn*]	AT * NOUN	ate the bananas, eat some cake
Combinación de los anteriores	Not possible	[[emailAddress@clothes]]	CLOTHES	dress, dresses, shoe, shoes
Combinación de los anteriores	Not possible	[[=clean]].[v*]	CLEAN_v	cleans, scoured, washing
Combinación de los anteriores	Not possible	[wear] * [=nice] [email@clothes]	VEAR * =nice OCLOTHES	wore some good-looking pants

## La LC y los enfoques lingüísticos

```
funcional ~ formal
```

lengua real ~ diseño mental

habla (parole) ~ lengua (langue)

actuación (performance) ~ competencia (competence)

empirismo ~ racionalismo

observ. datos naturales ~ introspección (trabajo de campo, corpus)

[Pérez Guerra, 1998]

"Las corrientes generativistas tempranas (Chomsky, 1957), que partían de la idea de que el número de enunciados de una lengua es infinito, mantenían que no puede haber ningún repertorio finito de datos lo suficientemente adecuado para la explicitación de los mecanismos de producción lingüística. Según esto, había que buscar el objeto de descripción en la figura de un hablante ideal, el cual posee la competencia lingüística que le permite producir enunciados en su lengua.

(...) Esta orientación...

[Rafel y Soler, 2003, p. 42]

...supuso implícitamente la crítica a la utilización de corpus como base de descripción de la lengua. La evolución posterior de la lingüística, y la especial atención prestada últimamente a la lingüística aplicada, ha hecho ver, sin embargo, que el objetivo de un corpus no es dar una visión total de la lengua, sino ofrecer una muestra representativa que permita al lingüista fundamentar su investigación en datos objetivos. Un corpus no puede identificarse con la lengua, sino que es un conjunto de datos que la representa de una manera más o menos fiable."

## Tipología de corpus

[Sinclair, 1996; Torruella y Llisterri, 1999; Vargas, 2006]

### **Criterios**

- el canal
- la representatividad
- la cronología
- los idiomas
- la anotación

- corpus textuales
  - muestra escrita
  - corpus de lengua oral transcrita
- corpus orales

corpus de referencia

- corpus sincrónico
- corpus diacrónico o histórico
- corpus cronológico o periódico (p/ej: s. XV)
  - corpus monolingüe
  - corpus corpus paralelo (bi-texts)
  - corpus comparable (paired texts)

- corpus simples
- corpus *lematizados* 
  - casa → casa
- corpus anotados con información morfológica
  - casa → casa, Sust., femenino
- corpus analizados sintácticamente (+ treebanks)
  - compró una casa → casa, Sust., femenino, OD
- corpus anotados semánticamente
  - casa → casa, Sust., femenino, OD, [vivienda]

### Corpus lematizados

- paradigmas flexivos
- "lema" en consultas lexicográficas

### Corpus marcados con la categ. Gram

- combinaciones de categorías (Adj+N, p/ej)

### Corpus marcados semánticamente

- hiperonimia, hiponimia...

# Corpus y L2

# Corpus y L2

# Diseño de materiales didácticos

- "(...) [en los manuales analizados] se utiliza un gran número de voces que tienen un alto índice de disponibilidad (...). Pero, junto a ellas, se usan algunas de bajo índice y otras de índice cero (...).
- (...) lexías que sólo son utilizadas una vez en todo el volumen. (...) esa adquisición será más rápida y segura cuanto mayor sea el número de veces que la forma se utilice.
- (...) debe graduarse la presencia de las formas léxicas en los manuales. Tienen que aparecer más frecuentemente aquellas que tienen un mayor índice de disponibilidad y, por tanto, el hablante extranjero va a tener que reproducir o descifrar un mayor número de veces."

[Benítez, 1994, p. 332]

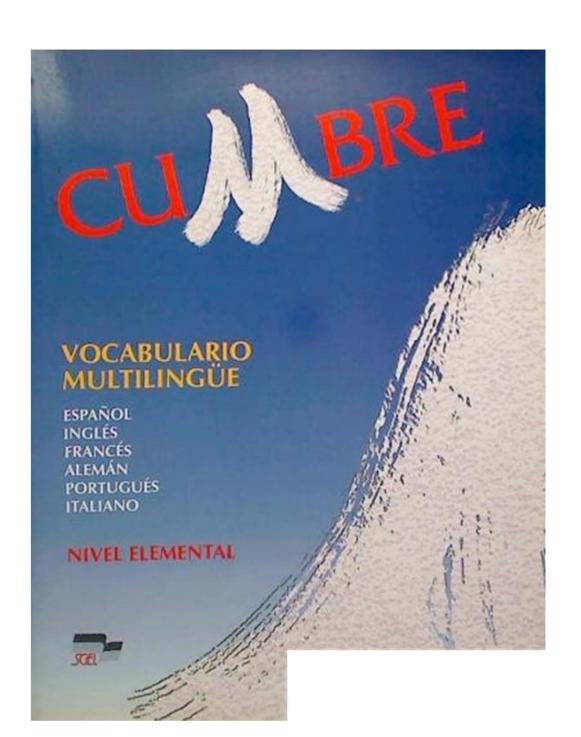
- 1º Conocer la importancia de las palabras de acuerdo con su frecuencia y distribución.
- 2º Decidir qué palabras deben enseñarse o no en cada nivel, en cuáles hay que insistir más y cuáles deben enseñarse para que los alumnos entiendan los materiales que manejan.
- 3º Poseer una información objetiva sobre las palabras y cuándo deben enseñarse.
- 4º Extraer el inventario de palabras del ámbito lingüístico en que se mueven los alumnos para facilitarles la comprensión de su entorno.
- 5º Saber cuáles son las palabras de ámbitos concretos cuando se enseña la lengua con fines específicos.
- 6º Seleccionar los grupos de palabras y formas que deben enseñarse conjuntamente, y en qué orden.
- 7º Determinar el grado de dificultad del léxico de los materiales de lectura y aprendizaje.
- 8° Que quienes confeccionan les materiales pedagógicos puedan seleccionar el léxico que debe aparecer en ellos.

[Alvar Ezquerra, 2004, p. 21]

"A corpus can be a very rich resource for writers of textbooks and other teaching materials because it gives us a detailed view of **how real people speak and write** in everyday situations. It can give us information about vocabulary, grammar, formality and informality, the differences between spoken and written language, how we perform such basic functions as requesting, greeting, and apologizing, how people open and close conversations, how we change the subject, how we interrupt one another, how we ask questions, and so on."

[McCARTHY, 2004<sup>a</sup>, p. 6] [Touchstone]





"Al principio el profesor quizá deba reducir la complejidad de consulta de algunos corpus y presentar actividades en un formato más "tradicional". El papel del profesor en el uso de esta herramienta es el orientar a los alumnos a familiarizarse con ella y deberá saber invitar a los estudiantes a formularse dudas y necesidades de aprendizaje que podrán satisfacer interactuando con dicho corpus"

[Soriano, 2018]

"(...) el **beneficio** que reporta su uso para ti y para tus alumnos es mucho mayor que el **esfuerzo** que supone aprender su manejo"

[González García, 2019[

### Aula 5, Ed. Difusión (B2)

- · CREA: sin lematizar
  - interfaz sencilla
  - formas flexionadas
- ¿ejemplo en mayúsculas?
- otros corpus
  - ¿diacrónicos?
  - ¿lematizados?
  - ¿interfaces?

# Corpus

# DE materiales didácticos

# Corpus DE materiales didácticos

- ejercicios de repaso
- materiales ajustados a los exámenes
- vocabulario por niveles

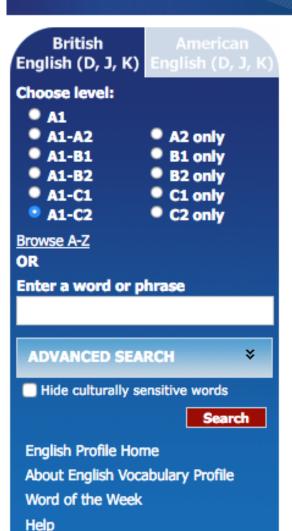


# **English Vocabulary Profile**





Preview version



#### **Help Contents**

Click on a heading to skip to that section.

- 1 Introducing the English Vocabulary Profile
  - 1.1 What is the EVP?
  - 1.2 Why is the EVP important to me?
  - 1.3 How has the EPV been produced?
- 2 British and American English
- 3 Search
  - 3.1 Browsing a list of words from a particular level
  - 3.2 Looking up a word
  - 3.3 Searching with wildcards
  - 3.4 Advanced Search
  - 3.5 Results list
  - 3.6 Single-click
- 4 Understanding dictionary entries
  - 4.1 Different coloured text
  - 4.2 Speaker icons
  - 4.3 Phonetic symbols
  - 4.4 Grammar symbols
  - 4.5 Word family

# Corpus DE materiales didácticos

"(...) The English Vocabulary Profile is based on extensive research using the Cambridge Learner Corpus (CLC). This is a collection of several hundred thousand examination scripts writte by learners fro added to every year. Combined with solid evidence of use in many other sources related to general English, such as examination vocabulary lists, classroom materials and a wide range of course books (...), the CLC confirms what learners can and cannot do at each level."

[English Profile, 2019]

# Corpus y L2

# Estudio de la interlengua











Diccionario de términos clave de ELE

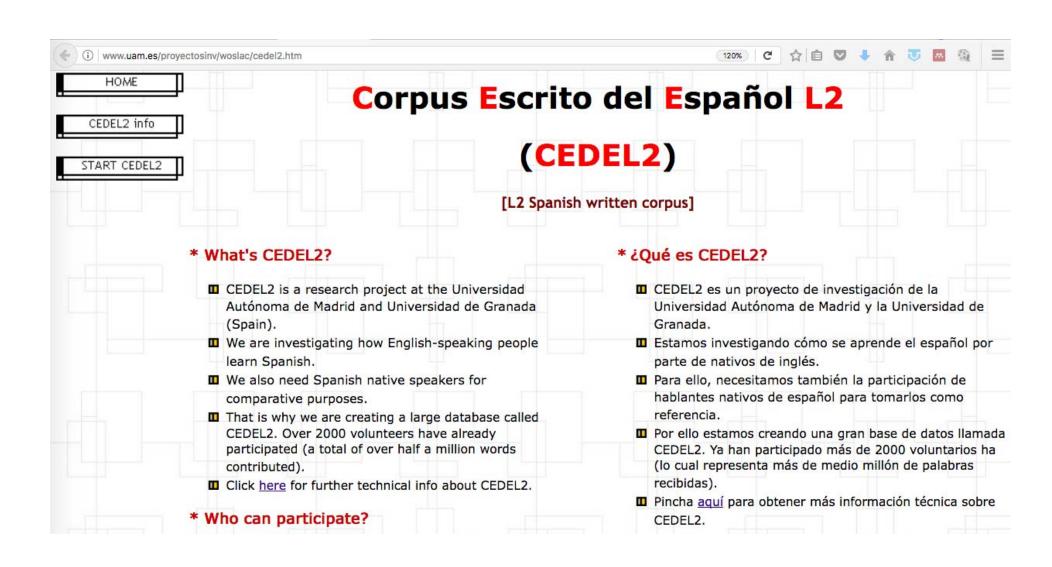
#### Lingüística de corpus

La lingüística de corpus es una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus, esto es, muestras reales de uso de la lengua. En rigor, el término no define una disciplina lingüística, como lo pueden ser la morfología, la sintaxis o la pragmática, sino un enfoque metodológico que es posible adoptar desde disciplinas diversas, que se contrapone a una metodología basada fundamentalmente en la introspección.

En la bibliografía especializada, el término corpus (plural córpora, aunque también se usa la forma corpus) se utiliza en dos sentidos distintos. En una primera acepción, designa la recopilación de material lingüístico hecha con un propósito de investigación concreto, ya sean muestras de oraciones, de enunciados o de textos. Este uso del término es frecuente en el ámbito de la lingüística aplicada, especialmente en las investigaciones de adquisición y aprendización en estudiar algún aspecto de la interlengua de la saprendientes de una L2 puede reunir una serie de producciones (largas o breves, hechas este fin o pensadas originariamente con otro objetivo de aprendizaje), que constituirán su corpus de estudio; se trata de los datos que sirven para ese estudio específico. En una segunda acepción, un corpus es una recopilación extensa de textos (escritos, orales o de ambos tipos) recogidos con el fin de servir como muestra representativa de una lengua, como conjunto de datos lingüísticos reales que reflejen el uso de la lengua (o del tipo de lengua específico) del cual quieren ser representativos. Es a esta segunda acepción a la que se refiere el término lingüística de corpus.

En su concepción, la lingüística de corpus es muy antigua, puesto que son múltiples los trabajos clásicos que han basado sus descripciones en muestras reales de lengua. Este es, por ejemplo, el único método de trabajo que han conocido los estudios de lingüística histórica o de fonética acústica; así mismo, es el método que han seguido muchas de las gramáticas clásicas (corpus básicamente oracionales) y que, en el

→ la importancia del "corpus de control"









#### CEDEL2: Corpus Escrito de Español L2



Corpus "de control"

→ etiquetas: lenguas, errores...



#### Corpus Oral de Español como Lengua Extranjera (ELE)

Entrevistas Errores Búsqueda

- 1. Objetivos
- 2. Participantes
- 3. Tipología de las lenguas maternas
- 4. Grabaciones y datos
- 5. Diseño de la entrevista
- 6. Convenciones de transcripción
- 7. Etiquetado de errores
- 8. Recuentos de palabras y listas de frecuencias
- 9. Referencias
- 10. Acceso a la guía didáctica
- 11. Actividades para profesores



#### 7. Etiquetado de errores

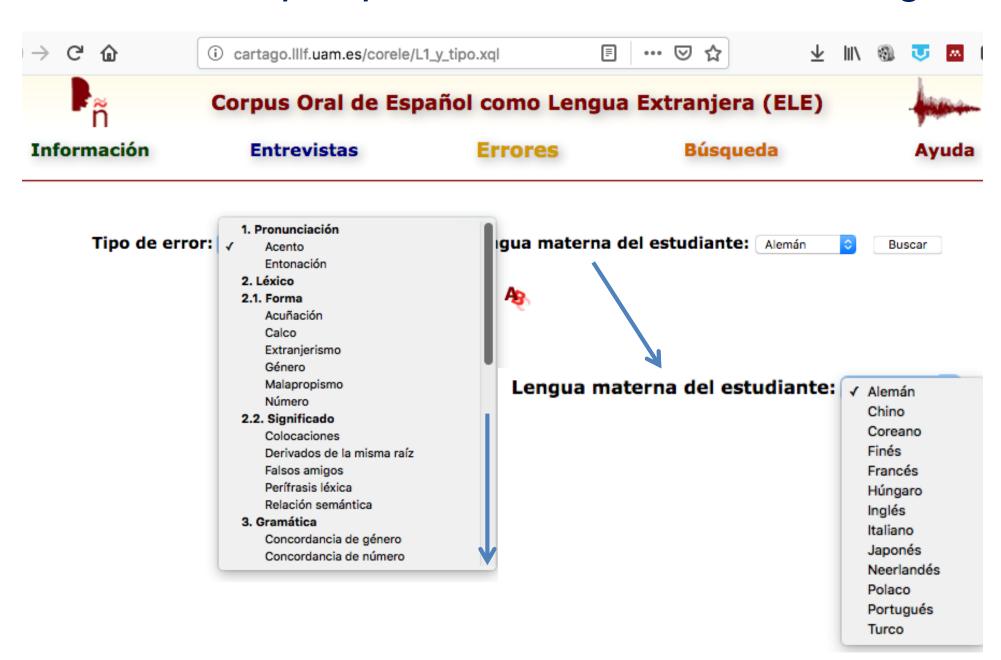
Los errores de producción oral hon sido marcados siguiendo una tipología basada en estudios previos para el inglés (James, 1998; Granger, 2003, Nicholls, 2003) y el español (Fernández, 1997; Bustos Gisbert, 1998; Vázquez, 1999).

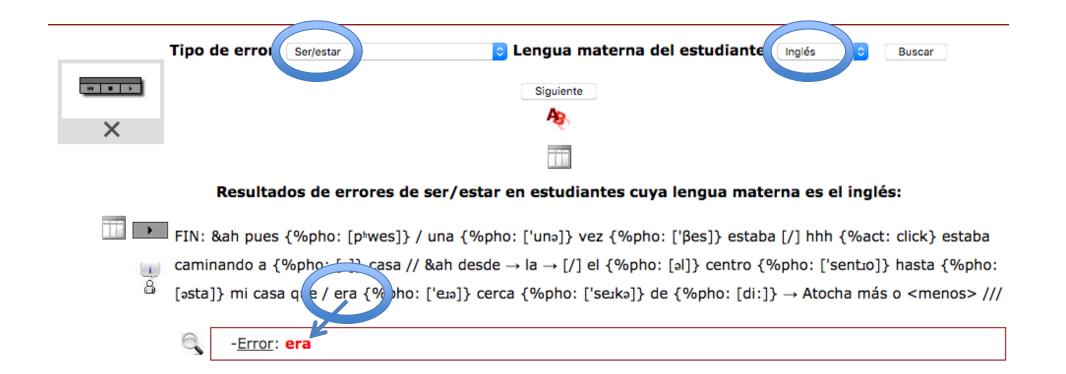
La información y los criterios según los que se han marcado los errores son los siguientes:

• Categoría de la palabra, el sintagma o la cláusula: la clasificación es la siguiente:



Búsqueda **Entrevistas** Errores Seleccione el tipo de búsqueda (solo puede seleccionar uno o dos parámet Nivel lingüístico Categoría Lengua materna Mecanismo de cambio Tipo de error ☐ Nivel de competencia (MCER) Seleccionar





→ etiquetado: lenguas, lemas...

El Corpus de aprendices de español como lengua extranjera (CAES) es un conjunto de textos escritos producidos por estudiantes de español con diferentes grados de dominio lingüistico (niveles A1 a C1 del Marco común europeo de referencia, aplicado al español en el Plan curricular del Instituto Cervantes. Niveles de referencia para el español) y procedentes de seis L1: árabe, chino mandarín, francés, inglés, portugués y ruso. Se trata de una herramienta que permite a los profesionales del campo de ELE (profesores, investigadores, evaluadores, autores de materiales didácticos, responsables y equipos de centros e instituciones lingüísticas, etc.) llevar a cabo investigaciones aplicadas sobre la base de datos sólidos y objetivos, ya que puede proporcionar información sobre dificultades de aprendizaje, errores más comunes, vocabulario más o menos empleado, etc. que se podrá aplicar con facilidad en las aulas o integrar en los textos. Sin negar la importancia que tienen otros proyectos de corpus de este tipo, el CAES viene a cubrir un vacío importante en el área específica de corpus de aprendices de lenguas en la línea de proyectos similares como el ICLE (International Corpus of Learner English) de la Universidad de Lovaina.

Acceda al Corpus de aprendices de español como lengua extranjera (CAES).



	Nivel de español							
	Coincidencias	Estudiantes	Freq. norm					
A1	1 216 / 155 458	386 / 526	7 822/millón					
A2	0/0	0/0						
B1	0/0	0/0						
B2	0/0	0/0						
C1	0/0	0/0						

L1									
	dencias	Estudiantes	Freq. norm.						
Chino mandarin	102 / 18 75	40 / 63	5 491/millón						
Francés	86 / 13 2	38 / 44	6 504/millón						
Inglés	68 / 8 764	18 / 26	7 759/millón						
Portugués	543 / 52 9 2	144 / 167	10 262/millón						
Ruso	43 /								
Árabe	374 / 55 .30	130 / 203	6 772/millón						

A1/Ruso	A1/Ruso	/Ruso Mi padre es informatico, es muy tranquilo, aunque a_veces	esta casado , pero eso no imoportante .
A1/Ruso	A1/Ruso	/Ruso Mi padre	esta casado y tiene dos hijos , mi y mi hermana Katerina .
A1/Ruso	A1/Ruso	/Ruso Kirill vive en Moscú con su madre ,	esta estudiante y soltero .
A1/Ruso	A1/Ruso	/Ruso Me madre llama Taisia , tiene 73 años y	esta jubilada .
A1/Ruso	A1/Ruso	/Ruso Mikhail	esta mariado y vive con su familia en Sankt_Peterburgo .
A1/Ruso	A1/Ruso	/Ruso Ella tiene 55 años ,	esta médico tambien .
A1/Ruso	A1/Ruso	/Ruso Yo voy a llegar tarde, por que hoy	esta partido de futbol en Moscu con mi favorito equipo de futbol .
A1/Ruso	A1/Ruso	/Ruso La madre	esta soltera y vive en Moscú con me .
A1/Ruso	A1/Ruso	/Ruso Mi padre es hubilado , tiene setenta_y_cuatro años .	esta viudo .
A1/Ruso	A1/Ruso	/Ruso	Esta engeniero y trabaja en una empieza .
A1/Ruso	A1/Ruso	/Ruso Irina , mi hermana major ,	está casada , tiene hija , mi sabrina Vita .
A1/Ruso	A1/Ruso	/Ruso Olga tiene veinte_y_seis años , no	está casada y vive con mis padres .
A1/Ruso	A1/Ruso	/Ruso Tenemos muchas intereses comunes , ella	está el guía de la ciudad par á mí .
A1/Ruso	A1/Ruso	/Ruso Mi padre se llama Slava , tiene 59 años ,	está médico , vive con mi madre que se llama Galina .
A1/Ruso A1/Ruso A1/Ruso	A1/Ruso A1/Ruso A1/Ruso	/Ruso Irina , mi hermana major , /Ruso Olga tiene veinte_y_seis años , no /Ruso Tenemos muchas intereses comunes , ella	está casada , tiene hija , mi sabrina Vita .  está casada y vive con mis padres .  está el guía de la ciudad par á mí .

Estudiante: 1528 Sexo: Hombre

L1: Ruso

Estudios: Otros

Número de meses en países de habla española: 0

Contactos personales en países de habla española: No

Conocimientos de otras lenguas:

Tarea:

Edad: 48

País: Rusia

Edad de inicio en el estudio del español: 47

Número de meses estudiando español: 2

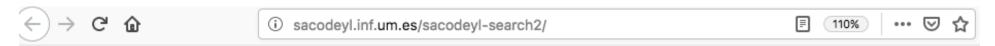
Tienen un hijo, se llama Nikita.

Es mi primo.

Kirill vive en Moscú con su madre, esta estudiante y soltero.

Kirill	vive	en	Moscú	con	su	madre		esta	estudiante	у	soltero	
NPP	VMUP9S	X	NPG	Х	DS+7CS	NCFS	Р	VMUP9S	iccs	С	NCMS	Р
Kirill	vivir	en	Moscú	con	su	madre		estar	estudiante	у	soltero	•

Mikhail esta mariado y vive con su familia en Sankt Peterburgo.



#### Welcome to the SACODEYL search tool!

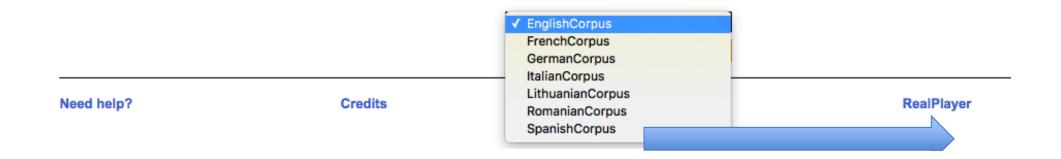
This website lets you search and browse pedagogic corpora in 7 European languages.

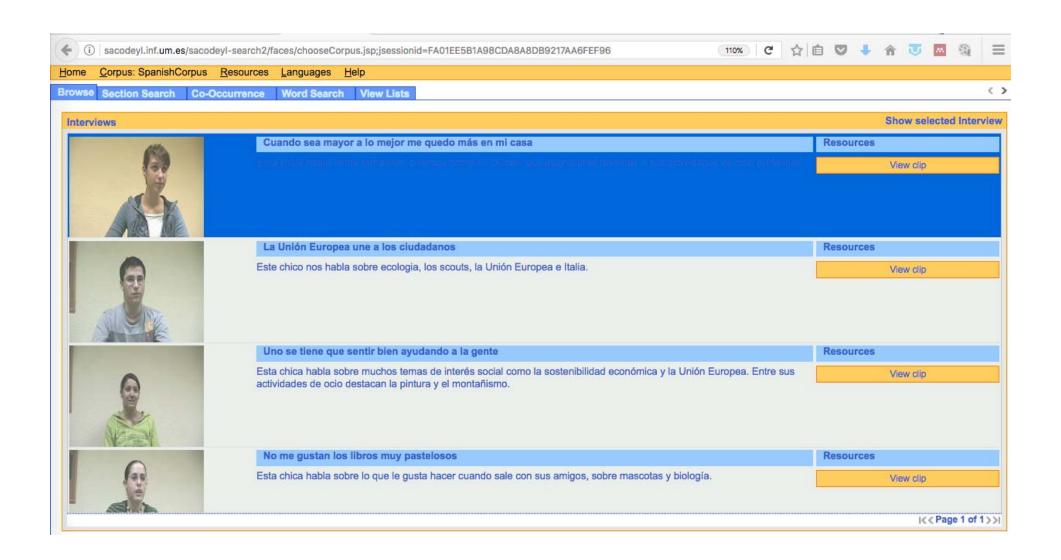
The corpora are based on structured video interviews with pupils between 13 and 18 years of age. The interviews have enriched for language learning purposes. Feel free to explore them on your own! Access to the corpora and the videos watch the videos, please install RealPlayer.

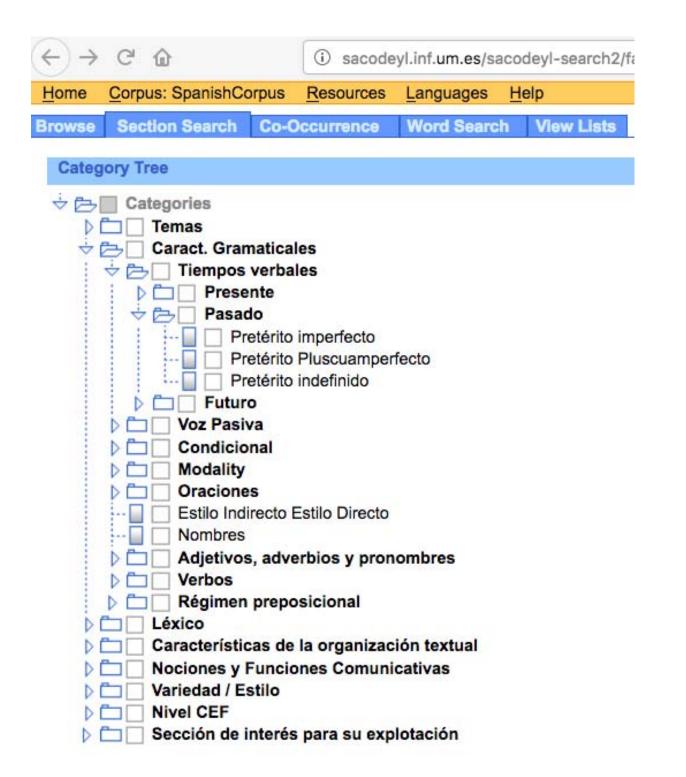
The names in the SACODEYL corpora are fictional and do not represent the real identities of the interviewees. The interviewees and/or the intutors have formally consented to the use of their interviews for educational purposes. NO OTHER USE of the materials is allowed. The SACO SACODEYL videos are provided as an integral part of the SACODEYL system and their use is only envisaged in the context of the SACOL

Permission from copyright holders is needed to create, among others, external course materials or web sites. You need to obtain permission who products in any ways that infringe on the exclusive rights granted to the copyright holders. For all non-educational uses, please contact s

By selecting a corpus and clicking on "Go!" you accept the aforementioned terms and conditions.

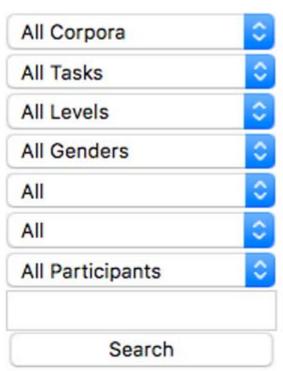








Corpus:
Task:
Level of the participant:
Gender of the participant:
Year(s) of Learning Spanish:
Other Language(s) studied:
Participant ID:
Search String



# Corpus y L2

# Uso didáctico de los corpus de aprendices



#### 11. Actividades para profesores

Actividad 1. ¿Qué corregir?

Actividad 2. La explicación del error

Actividad 3. Los errores de cada grupo de alumnos

Actividad 3.1. Estudiante 1

Actividad 3.2. Estudiante 2

Actividad 3.3. Estudiante 3

Actividad 3.4. Estudiante 4

Actividad 3.5. Estudiante 5

Actividad 3.6. Estudiante 6

Actividad 3.7. Estudiante 7



#### Actividad 1. ¿Qué corregir?

Vas a escuchar a una estudiante narrando una historia a partir de unos dibujos.

- ⇒ Puedes ver los dibujos pulsando abajo en "Mostrar viñetas".
- 1. Escucha una primera vez sin leer la transcripción. ¿Cuál crees que es la lengua materna de la estudiante? ¿Por qué?
- 2. Escucha una segunda vez y anota los errores que corregirías. Justifica tu respuesta (ej. porque impiden la comunicación).

Si realizas la actividad con otro(s) profesor(es), podéis discutir vuestras propuestas de corrección.

#### Mostrar/ocultar viñetas











#### Suggestions for How to Use this Website

This site is useful for:

- Teacher candidates
- Teachers
- · Pos
- Learners of Spanish
  - Linguistics student

#### The corpus may be used for:

- Classroom assignments
- Research purposes
- Practice correcting learner errors
- Self-improvement of Spanish grammar, vocabulary, fluency or conversation management

This site can be used at the undergraduate and graduate level in:

- Spanish second language acquisition courses
- Spanish applied language courses
- · Spanish teacher training courses

introduction

about

proficiency levels: practice

video activity

glossary and links

What level is the Spanish Learner? beginner, beginner-intermediate, intermediate, intermediate-advanced, or advanced? Select a learner and explore they talk about. Answering the questions for each video will help you determine at what level the learner is speaking Spanish.

#### Learners

Abigail

Alexi

Allison

Amy

Anthony

Brenda

Brian

Cindy

Claire

Elizabeth

Emily

Grace

Hannah

Heather

Jacob

James

Jenny

Jessica

John

Julie

#### **Topics**

Hometown

Past

Sports

Politics

Other Countries

Family

Comparisons

Major

Apartment



vocabulary

fluency

conversation management

generalizations

evaluate

## Corpus de nativos y L2

# Las concordancias como input

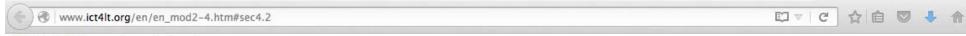
### https://sites.google.com/view/corpus-ele-jaen-2019





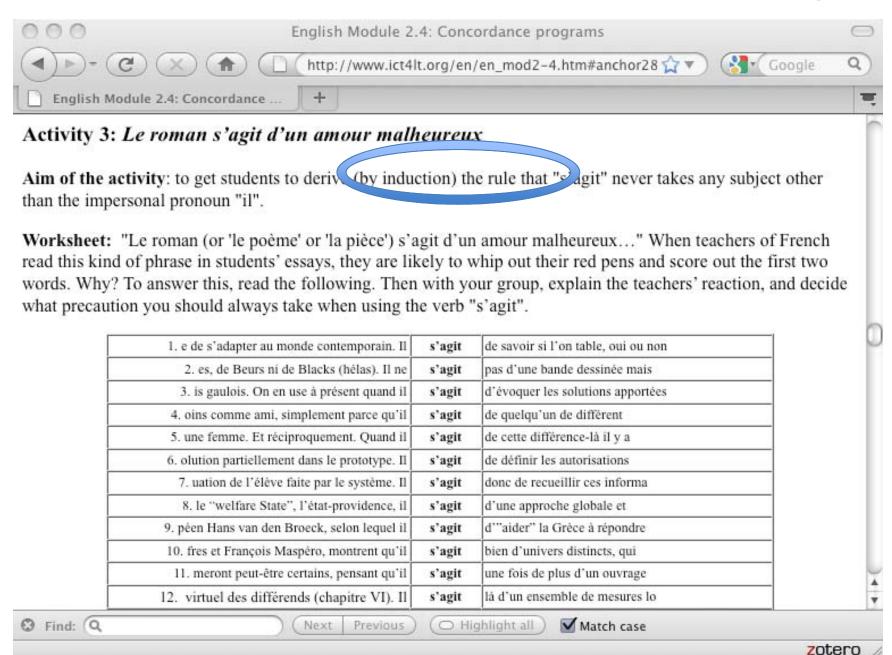
- 1. Llenar espacios en blanco en concordancias
- 2. Completar el contexto más amplio de las líneas de concordancias
- 3. Análisis del tono o estilo de diferentes textos
- 4. Utilización libre del programa de concordancias como ayuda a la escritura, la corrección de redacciones o la comprensión de textos.

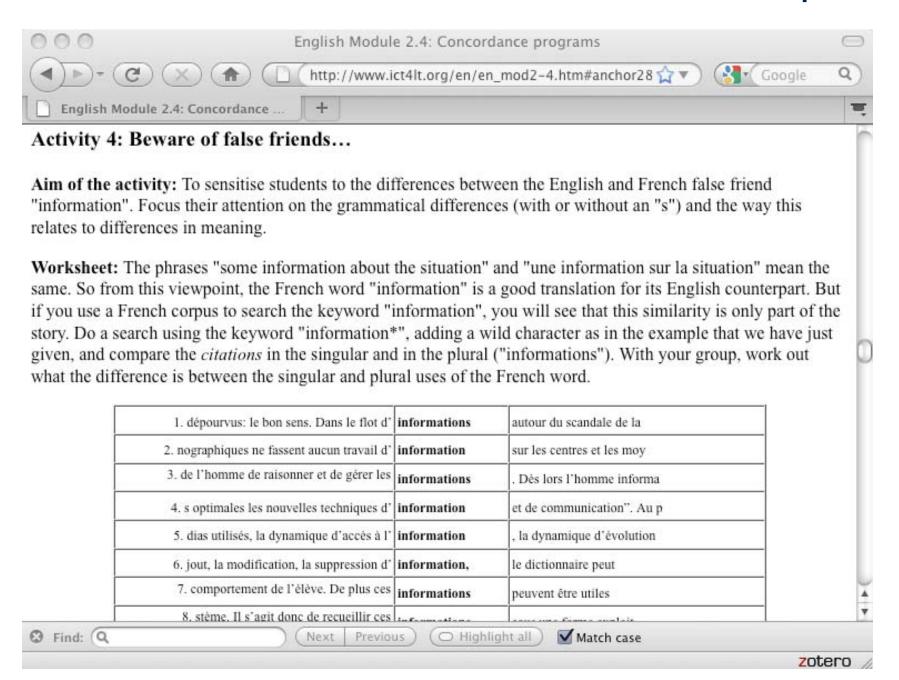
[Honeyfield (1989), McCullough, 2001, p. 136]



#### 4.2 Activities and worksheets

ACTIVITY NAME AND LANGUAGE	PURPOSE OF ACTIVITY	PRESENTATION
100		
1. Guess the mystery word (F)	beginners, lexis	paper,online (LAN)
2. Donc on peut dire que (F)	style, usage	paper, online (LAN)
3. S'agit (F)	derive a rule, grammar	paper, online (LAN)
4. Beware false friends (F)	lexis	paper, online (LAN)
5. Coffee or tea? (F)	cultural differences	paper, online (LAN)
6. Les Anglais et les Britanniques (F)	political correct, usage, lexis	paper, online (LAN)
7. Changing lifestyles (F)	cultural differences, usage	paper, online (LAN)
8. dürfen and müssen (G)	lexis, usage	paper
9. Preposition am (G)	usage	paper
10. Syntax of adverbs (E)	driving a rule, grammar, syntax, statistics, usage	online (WWW)
11. reason + because (E)	correctness, usage	online (WWW)
12. Students' own writing (E)	variety, usage	online, standalone PC
13. L'aigle noir (F)	literary analysis	online (LAN)
14. Mariana (E)	literary analysis	online (LAN or WWW)





### RESULTADOS ESPERADOS

- Aportar datos empíricos al campo de investigación del ABD en la enseñanza y aprendizaje de ELE.
- Introducir a este grupo de estudiantes en el mundo de los corpus lingüísticos. Se espera confusión y recelo en algunos estudiantes pero se espera convencer a la mayoría de las ventajas de este enfoque.
- Elaborar un banco de actividades de ABD para uso de profesores de ELE.

## Corpus y L2

¿Qué buscar? ¿En qué corpus?

## ¿En qué corpus?

## Muchos corpus textuales de nativos en línea

## ¿Útiles? ¿Difíciles?

- → Depende
  - Conceptos básicos sobre LC
  - Tipos de corpus
  - Tipos de interfaces
  - Proyectos (beta)

# → Conocer para escoger

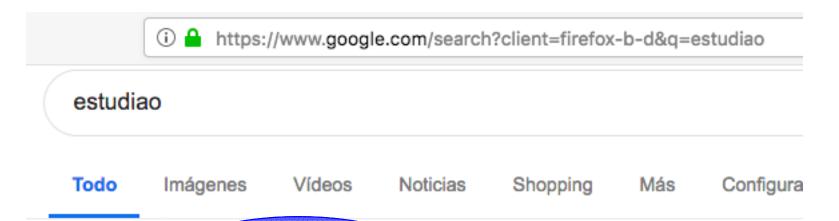
## ¿ Qué buscar?



### La palabra X, ¿se dice?

"estudiao", ¿se dice?

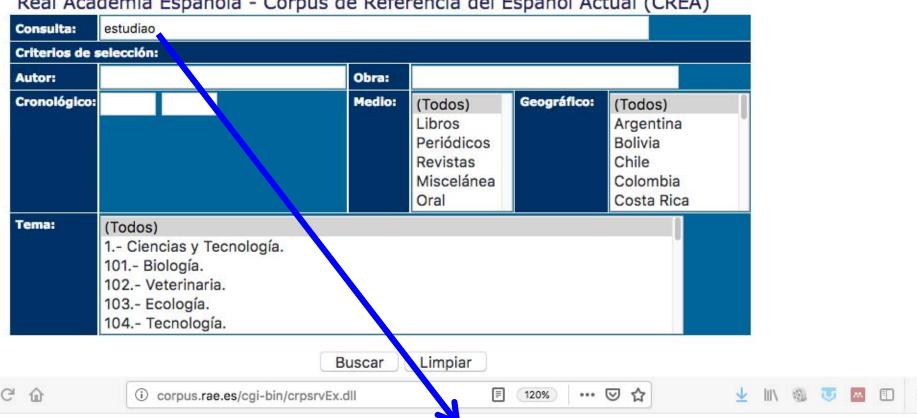
→ Google



Aproximadam (nte 290.000 resultados (0.20 segundos)



Real Academia Española - Corpus de Referencia del Español Actual (CREA)



#### No CONCORDANCIA

- libros!... Eso te sale o no te sale y, como no es estudiao, ú ni te enteras bien y muchas veces coge d \*\*
- bro y algunos hasta de poesías, yo creo que había estudiao y todo. Bueno, pues esa calcumanía de mujer, \*\* 2

### Corpus Leeds

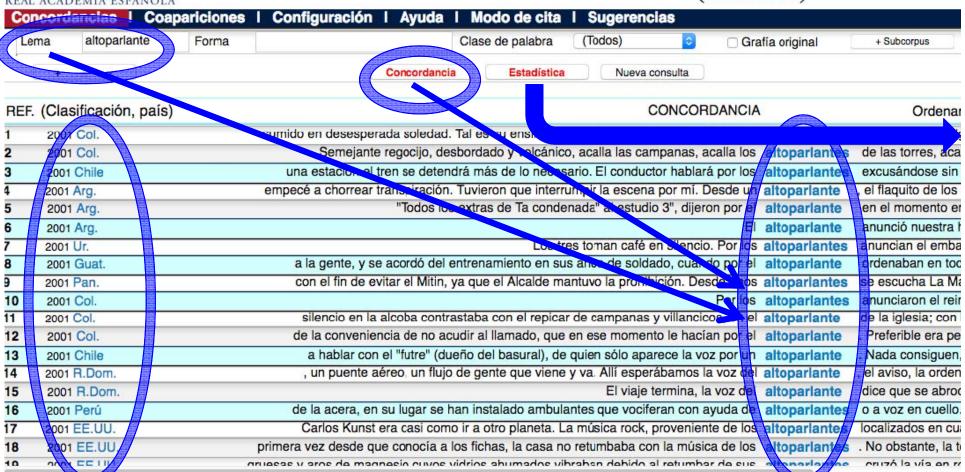


### La palabra X, ¿dónde se dice?

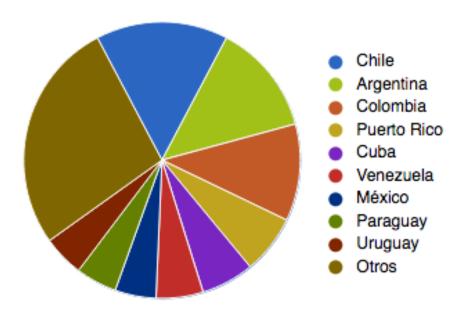
### La palabra "altoparlante", ¿dónde se dice?



## Corpus del Español del Siglo XXI (CORPES)



#### Distribución País



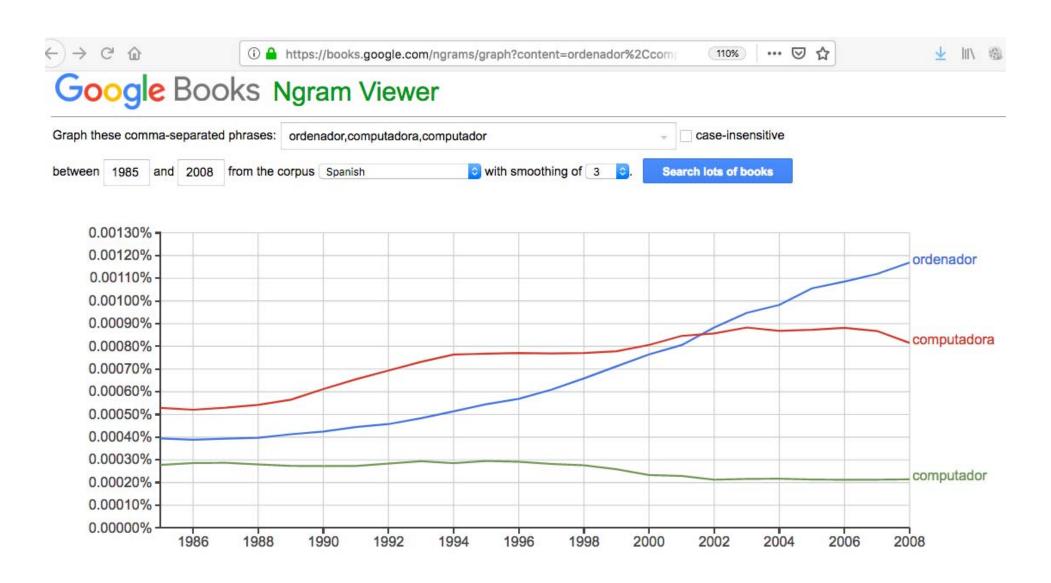
#### Distribución Zona



### Qué palabra se usa más, ¿X o Y?

### Qué palabra se usa más:

→ "ordenador", "computadora" o "computador"



### La palabra X, ¿cómo se usa?

La palabra "aunque", ¿cómo se usa?



i https://www.linguee.es/espanol-ingles/search?source=a





### aunque

#### Diccionario español-inglés

although conj 🜓 🛈

Aunque estoy a dieta, de vez en cuando me como un helado.

while conj (1)

though conj 🜖 🛈

Podemos cenar pizza esta noche, aunque preferiría pasta.

even if conj 🜒 🛈

Voy de campamento este fin de semana aunque nieve.

even though conj 🜒 🗓

Aunque el libro es viejo, está en buen estado.

albeit conj (1)

Finalmente llegaron los invitados, aunque muy tarde.

manag fraguantas

consulted.

daccess-ods.un.org



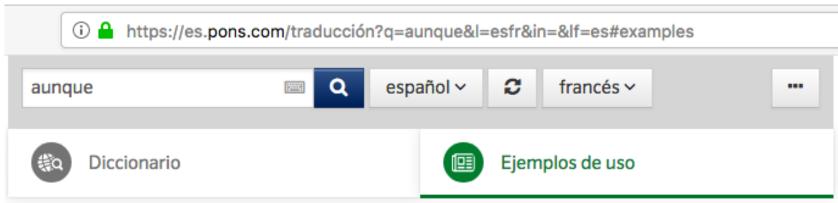
su presentación, los servicios interesados son consultados

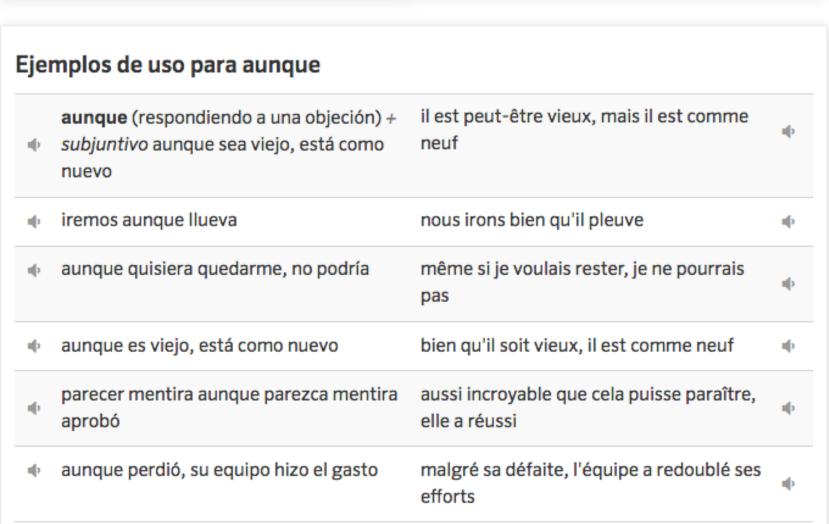
sistemáticamente.

For simplicity, the term "bank" is used, although partnerships may

they are submitted, the departments concerned are systematically

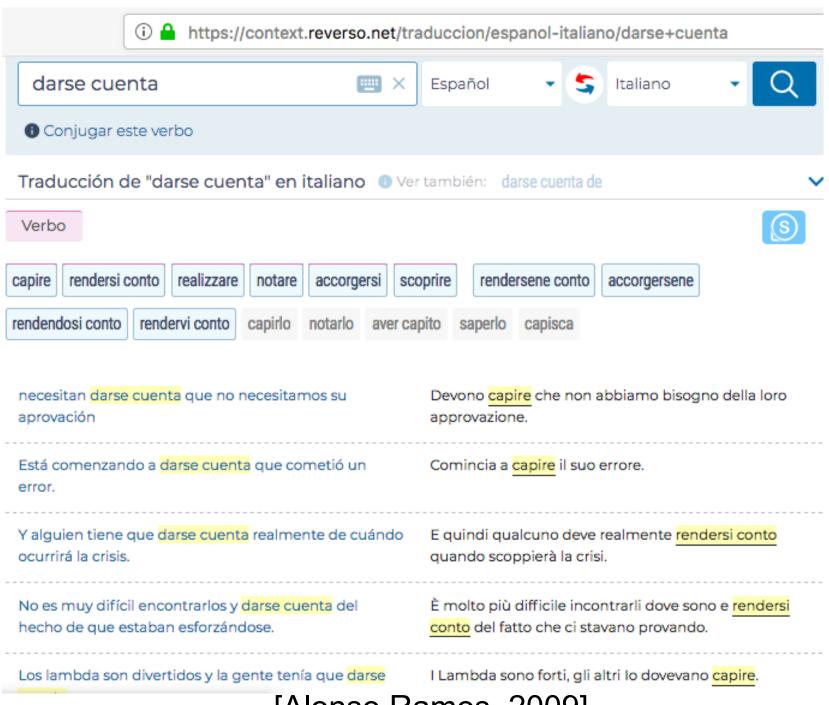
daccess-ods.un.org





### La <u>expresión</u> X, ¿cómo se dice <u>en otras lenguas</u>?

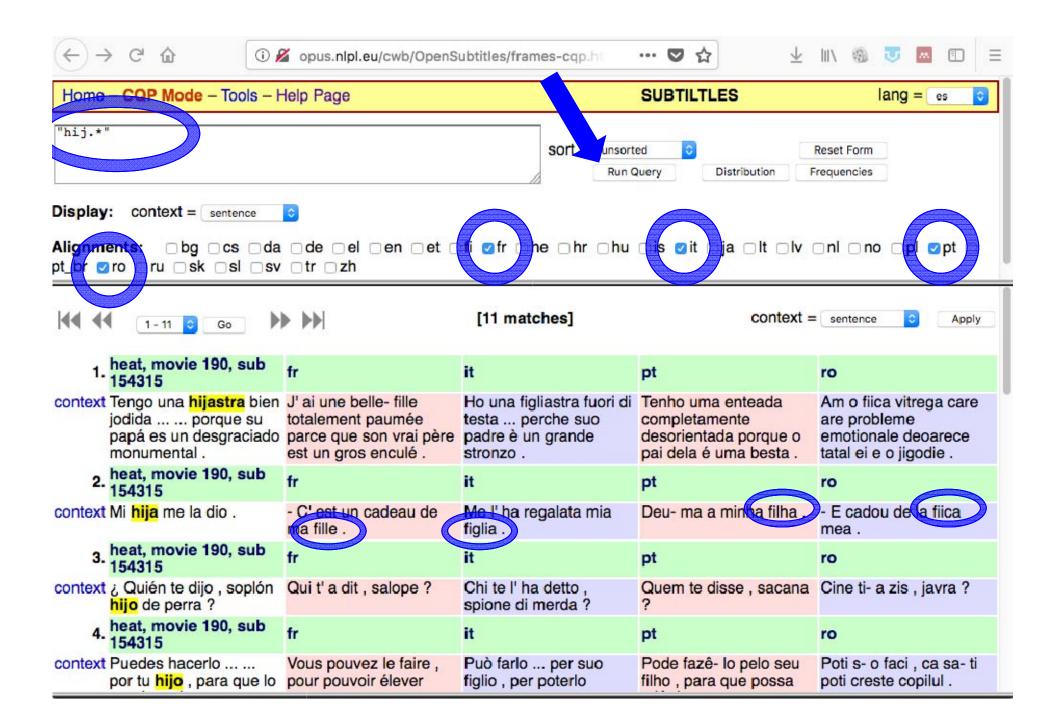
La expresión "darse cuenta de...", ¿cómo se dice en otras lenguas?

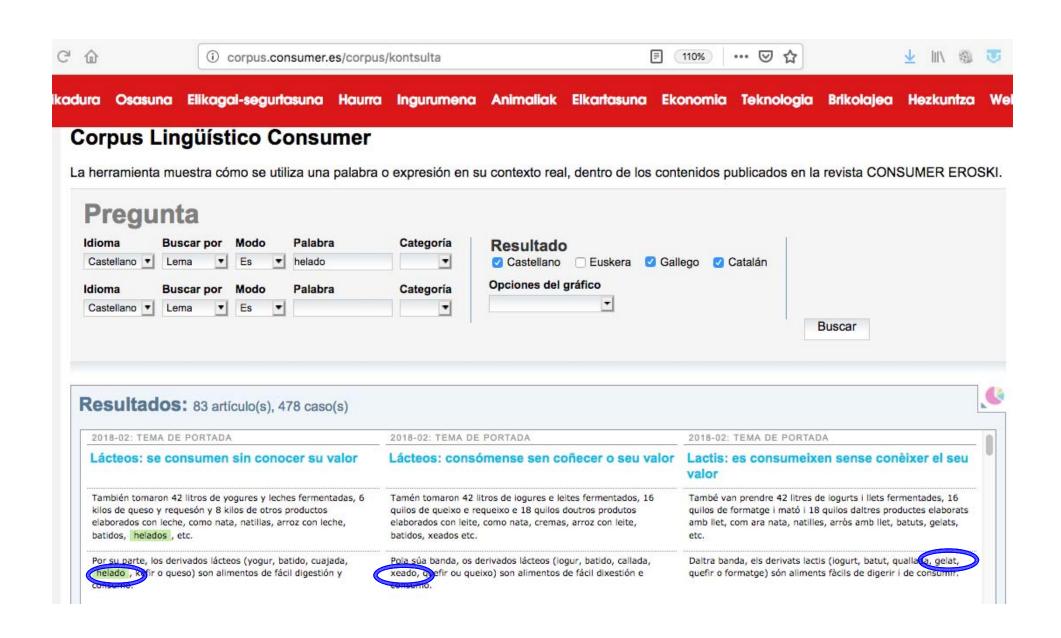


[Alonso Ramos, 2009]

# La palabra X, ¿cómo se dice en otras lenguas parecidas?

La palabra "hijo", ¿cómo se dice en otras lenguas románicas?





La palabra X, ¿cómo se usa en la lengua oral?

La palabra "mola", ¿cómo se usa en la lengua oral?

#### i valesco.es/?q=palabras



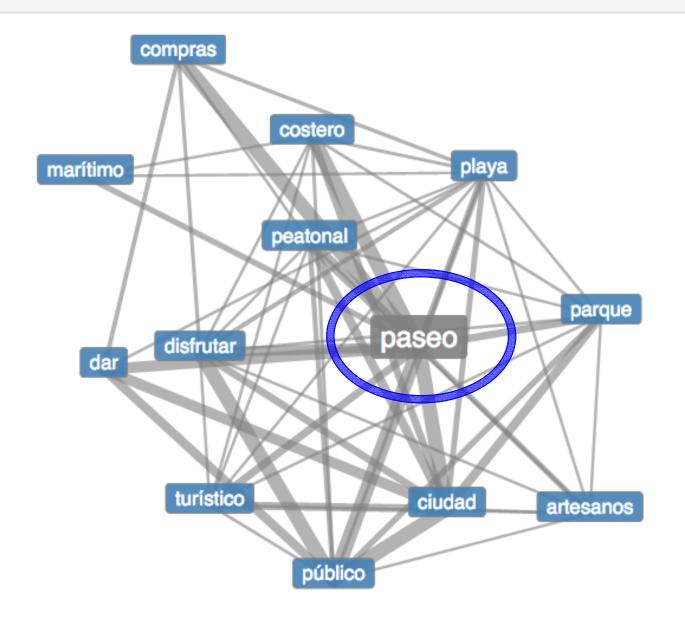


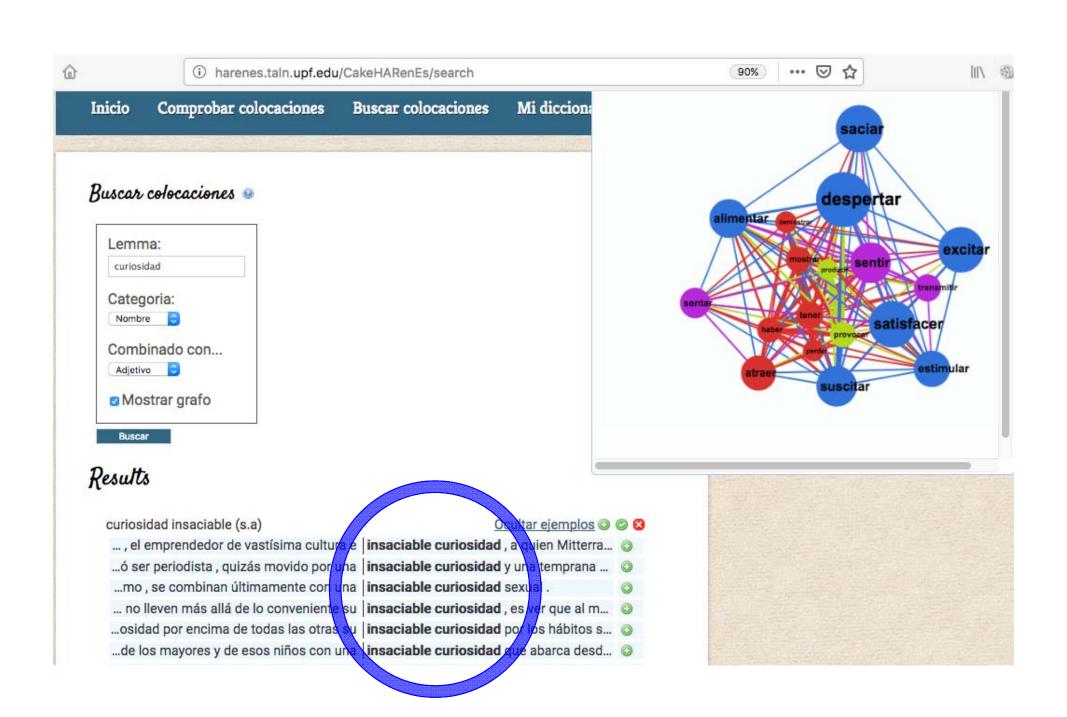
* Palabra (campo obligatorio)	Alargamiento vocálico	Orden palabra en intervención	Orden de la intervención	Texto en la intervención
mola	no	002	0009	C: cómo mola nano (( ))
mola	no	005	0010	B: a mí no me mola NAda
mola	no	003	0011	C: no te mola nada ¿el qué?§
mola	no	002	0015	C: §no mola nada hoy sucesos† °tío° mola más†=
mola	no	007	0015	C: §no mola nada hoy sucesos† *tío* mola más†=
mola	no	002	0017	C: =Levante mola más§
mola	no	006	0019	A: §((en el levante sí que mola))§
mola	no	007	0067	B: [tomás ese/ tomás a] mí me mola porque es un tio [que/ aunque no tenga muy buen nivel/ se lo curra nano]
				C. 8 v to(do) el rato

¿Qué palabras se suelen combinar con la palabra X?

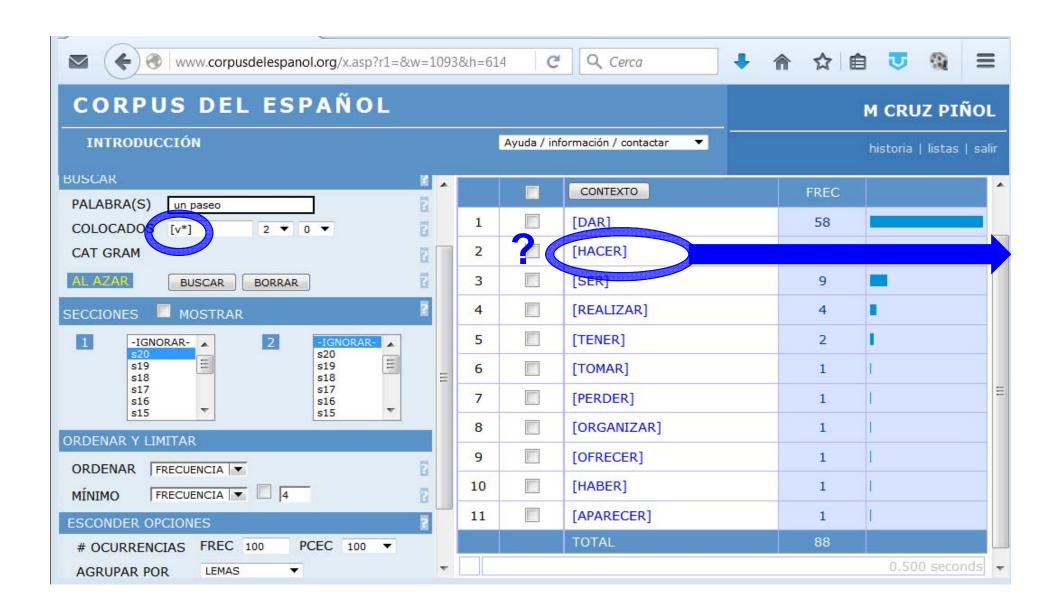
¿Qué palabras se suelen combinar con la palabra "paseo"?







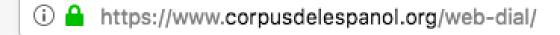
¿Qué <u>verbos</u> se suelen combinar con "un paseo"?





# ¿Qué combinaciones N+negro son más frecuentes?





## Corpus del Español: Web/Dialects





**SEARCH** 

**FREQUENCY** 

List Chart Collocates Compare KWIC						
negro	Word/phrase [POS]					
_N*	Collocates noun.ALL					
+ 4 3 2 1 0	0 1 2 3 4 +					
Find collocates Res	et					





## Corpus del Español: Web/Dialects (i) 🖹 🕹 🗳 🎬 🍳









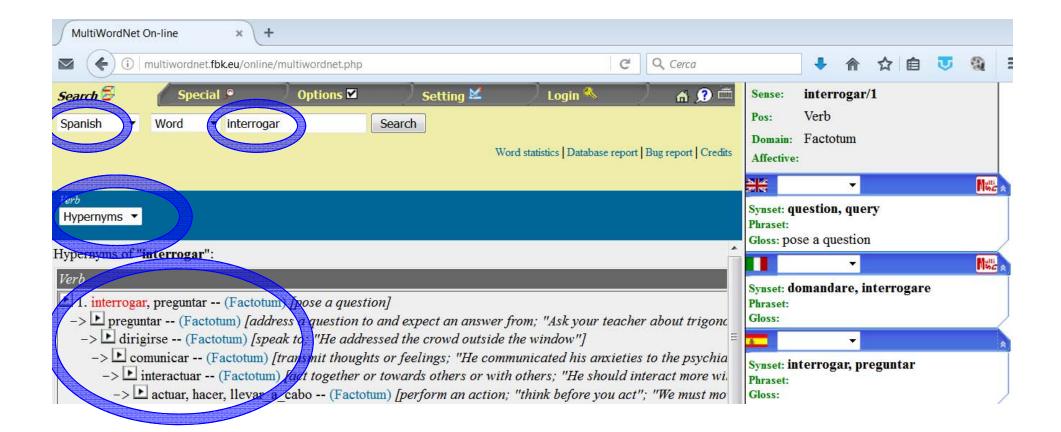




	SEARCH		FREQUENCY	CONTEXT
1		COLOR		6731
2		AGUJERO		5580
3		MERCADO		3782
4		HUMOR		2717
5		DINERO		1179
6		PELO		1141
7		FONDO		990
8		ORO		913
9		VESTIDO		868
10		CHOCOLATE		867
11		GATO		838
12		CINE		822
13		нимо		754
14		PUNTO		746
15		CISNE		669
16		TRAJE		662
17		CABELLO		660
18		ΤÉ		643
19		PRESIDENTE		641
20		GÉNERO		599

# La palabra X, ¿qué relaciones semánticas mantiene (hiperónimos/hipónimos)?

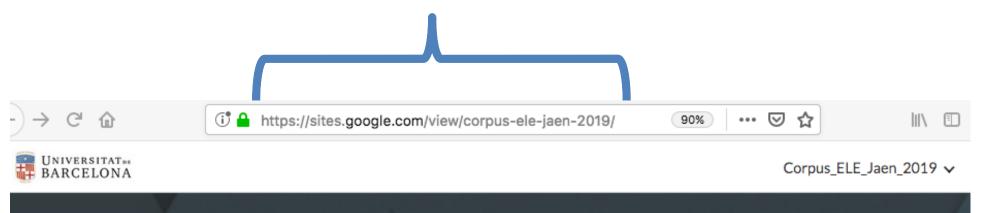
La palabra "interrogar", ¿qué relaciones semánticas mantiene?



### A modo de síntesis

- Conocer para escoger
- No es complicado (la web como corpus)
- Conocimientos elementales sobre LC
- Verificar si ocurre
- Corpus como juez
- Concordancias como input
  - Lengua en uso
  - Aprendiz explorador
  - Ejercicios (huecos...)
  - Aprendizaje autónomo

### https://sites.google.com/view/corpus-ele-jaen-2019/



#### Referencias

#### Corpus y Webs

"Lingüística de corpus y español como L2/LE: de la investigación al aula" (InLéxico 2019)

Mar Cruz Piñol (abril 2019)

https://sites.google.com/site/marcruzpinolub/