

TALLER  
I+D+I

# SKETCH ENGINE: CONSULTA Y GENERACIÓN DE CORPUS HISTÓRICO-DIALECTALES EN ESPAÑOL

**Miguel Calderón Campos**  
(Universidad de Granada)

**21.05.2026**  
09:00 - 12:00H.  
15:00 - 18:00H.

**22.05.2026**  
09:00 - 12:00H.

Aula R.S.40  
Faculté des Lettres et Sciences Humaines (UnINE)  
Espace Tilo-Frey (CH-2000, Neuchâtel)

**unine**  
Université de Neuchâtel  
Institut de langues et  
littératures hispaniques

Proyecto PID2022-136256NB-I00, financiado por  
MICIU/AEI/10.13039/501100011033 y por FEDER, UE



# Sketch Engine: consulta y generación de corpus histórico-dialectales

Miguel Calderón Campos

calderon@ugr.es

Proyecto VITA VERBORUM: PID2022-136256NB-I00

financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE



Université de Neuchâtel: Institut de Langues et Littératures hispaniques-  
21-22 de mayo de 2026

# Introducción y definición de corpus

# ¿De qué trata este curso? (1/2)

## Herramientas de análisis textual

Herramientas que permiten analizar  
enormes cantidades de texto y  
**detectar patrones**  
que de otro modo serían invisibles

# ¿De qué trata este curso? (2/2)

Para detectar patrones **en textos** necesitamos tres cosas:

Los **datos** — los textos en formato digital (.txt), organizados en lo que llamamos un *corpus*

Las **herramientas** — plataformas como **Sketch Engine**, Voyant Tools, AntConc, LancsBox

Saber **cómo hacer la consulta**

Nota: para trabajar con texto plano (txt) se recomienda BBEdit (Mac) y Notepad++ (Windows)

# ¿Qué vamos a aprender?

Buscar en corpus  
ya existentes (predeterminados)

explotar corpus de gran tamaño como **esTenTen23**, **CORPES XXI**, etc. para encontrar patrones lingüísticos, ejemplos de uso real y combinaciones usuales de palabras

Construir nuestro  
propio corpus

convertir los textos que ya tenemos o que necesitamos para nuestra investigación (textos en PDF, Word, txt) en un corpus sobre el que hacer búsquedas automáticas.

# ¿Qué es un corpus?

**1**

## **Una colección de textos**

seleccionados con un criterio claro: las obras completas de un autor, los discursos del rey, artículos de opinión sobre un tema...

**2**

## **Representativos**

de lo que queremos estudiar: el corpus tiene que contener los textos adecuados para responder a nuestra pregunta de investigación.

**3**

## **En formato electrónico**

lo que nos permite hacer búsquedas automáticas sobre grandes cantidades de texto: miles - millones de palabras.

# Antes de seguir: una pregunta previa

¿Qué textos necesitáis para vuestra investigación?

- ¿Los tenéis ya en formato electrónico?
- ¿Qué patrones os interesan o qué preguntas os gustaría poder hacerles?
- ¿Qué tipo de análisis os ayudaría en vuestra tesis/artículo?

*Tened estos textos y preguntas en mente a lo largo de todo el taller  
Empezamos con una introducción al manejo de corpus predeterminados*

# Corpus predeterminados

*esTenTen23 · Spanish Trends · Timestamped JSI web corpus 2014-2021 Spanish*

Corpes XXI

# ¿Qué es un corpus predeterminado?

Además de construir nuestro propio corpus, podemos trabajar con corpus ya existentes, creados y mantenidos por instituciones o equipos de investigación.

La ventaja es inmediata: contienen millones de palabras y están listos para usar. Inconveniente: pueden no adaptarse a nuestras necesidades

**En este taller trabajaremos con corpus predeterminados, alojados en dos plataformas distintas.**

# Los corpus que vamos a usar

## Sketch Engine

- esTenTen23: casi 29 000 millones de palabras extraídas de la web en español (2020-2023)
- Spanish Trends - Timestamped JSI spanish

## CORPES XXI — Real Academia Española (acceso libre)

- Corpus del Español del Siglo XXI: 500 millones de palabras, con selección textual cuidada y datos de uso por países

# Primeras búsquedas en CORPES XXI

*Lema - Forma - Concordancia - Corpus como complemento de los diccionarios - Inventarios de variantes - Coapariciones - Distribución geográfica - Frecuencia absoluta y normalizada*

# CORPES XXI

El *Corpus del Español del Siglo XXI* es el corpus de referencia de la RAE. Contiene **algo más de 500 M de palabras** de textos escritos y orales contemporáneos producidos en todos los países hispanohablantes.

1

## Acceso libre

No necesita registro ni instalación

2

## Textos bien seleccionados

Prensa, literatura, ensayo, textos orales...

3

## Distribución por países

Compara el uso en todo el mundo hispanohablante

# ¿Cómo calcular el tamaño de CORPES XXI y de los subcorpus nacionales?

Basta con hacer una consulta vacía en la página principal, sin tocar nada

## Corpus del Español del Siglo XXI (CORPES)

**Búsqueda**

---

Elementos gramaticales ▾

Lema

Forma

 Categoría gramatical

+

**Sensibilidad** ☐ Acentos ☐ Mayúsculas

**Filtros**

---

**Resultado**

---

Tipo de resultado  
Estadísticas ▾

---

Limpiar

Buscar

# Lema y forma

Antes de hacer nuestra primera búsqueda, necesitamos entender una distinción fundamental.

## Lema

La palabra tal como aparece en el diccionario.

*zozobrar, zafio, amanuense, pequeño*

## Forma

Cualquiera de las variantes de ese lema.

*zozobrando, zozobraba, zozobra... / zafio, zafias...*

IMPOTANTE: ver en CORPES XXI qué variantes se recogen bajo el lema *pequeño*

# Página principal de CORPES XXI (dic. 2025)

## Corpus del Español del Siglo XXI (CORPES) <sup>Versión 1.4</sup>

### Búsqueda

Elementos gramaticales ▾

Lema

Forma

Categoría gramatical

+

Sensibilidad

☐ Acentos

☐ Mayúsculas

### Filtros

### Resultado

Tipo de resultado ▾

Concordancias

Concordancias por página ▾

20

Limpiar

Buscar

# Tipos de resultados disponibles en CORPES XXI

Estadísticas

Concordancias



Documentos

Coapariciones

Inventarios

Tipo de resultado

Concordancias



## Ejemplos del lema *zozobrar* en CORPES XXI

<input type="checkbox"/>	ⓘ	LE2001_0018	Esp.	2001	...olino o mar gruesa que hiciera	<b>zozobrar</b>	la frágil barca de mi existencia y ...
<input type="checkbox"/>	ⓘ	LE2001_0020_012	Esp.	2001	...arbón, y cuando su dedo índice	<b>zozobra</b>	por fin en el terreno anegadizo d...
<input type="checkbox"/>	ⓘ	LE2001_8975	Esp.	2001	...s, como un descomunal galeón	<b>zozobrado</b>	en medio de la oscuridad envene...
<input type="checkbox"/>	ⓘ	LE2001_8981	Esp.	2001	...dedicarla a esto), en ocasiones,	<b>zozobra</b>	. «Me siento poco valorada, ya n...
<input type="checkbox"/>	ⓘ	PA2001_1585	Chile	2001	...én después que su embarcación	<b>zozobró</b>	en las costas de una pequeña isla...
<input type="checkbox"/>	ⓘ	LA2002_0021	Méx.	2002	...galos del 6 de enero. Entonces	<b>zozobró</b>	la esperanza y acepté resignada ...
<input type="checkbox"/>	ⓘ	LA2002_0083	Col.	2002	...ulación de un navío a punto de	<b>zozobrar</b>	arroja. Así hemos acallado a lo la...
<input type="checkbox"/>	ⓘ	LA2002_0090_002	Hond.	2002	...fuerzos de la tripulación por no	<b>zozobrar</b>	, emitió un todavía más devastad...
<input type="checkbox"/>	ⓘ	LA2002_0132_008	Cuba	2002	...anantial de ideas, de pasiones	<b>zozobradas</b>	, de sueños estuprados que viaja...

# ¿Qué es una concordancia?

Una concordancia es una **lista** de todos los **ejemplos** de uso de una palabra en el corpus, mostrados en su **contexto**.

A esto se le llama **KWIC** (Key Word in Context) o **PCEC**: la palabra buscada aparece en el **centro** y podemos ver lo que hay a su izquierda y a su derecha. No nos dice qué *significa* una palabra, sino cómo se *usa* realmente.

...el proyecto había empezado a **zozobrar** por falta de recursos...

...la ilusión empezó a **zozobrar** ante tanta incertidumbre...

*Ejemplo de KWIC — Palabra clave en contexto*

## ¿Qué cosas pueden zozobrar?

### Dar dos sinónimos de *zozobrar*, sin consultar el DLE

Varias formas de hacerlo:

1. Leer concordancias
2. Obtener coapariciones o colocaciones
3. Ir al corpus esTenTen de Sketch Engine y consultar
  1. Word Sketch
  2. Tesauro

# Coapariciones (colocaciones) del lema *zozobrar*

## *Sustantivos* que suelen aparecer en el intervalo -5, +5 en CORPES

Coapariciones

sustantivo ▼

Lema	Cat.	Frec.	MI	T-Score	LL
embarcación	sustantivo	21	12,73	4,58	143,61
barco	sustantivo	21	10,87	4,58	120,03
costa	sustantivo	10	9,3	3,16	47,54
punto	sustantivo	19	7,32	4,35	68,02
agua	sustantivo	10	6,7	3,16	31,96

Podemos ver ejemplos de la combinación de zozobrar + embarcación, zozobrar + barco, etc.

## Tesouro (esTenTen)

	Lempos	Frecuencia <sup>?</sup>	
1	naufragar	52.184	...
2	encallar	25.491	...
3	semihundir	1411	...
4	embarrancar	4551	...
5	fondear	38.864	...
6	escorar	17.225	...
7	atracar	69.825	...
8	faenar	29.133	...
9	adrizar	644	...
10	varar	118.438	...

## word sketch (esTenTen)

objects of "zozobrar"			
<b>yola</b>	20	9,5	...
zozobrar una yola			
<b>barquilla</b>	6	7,5	...
<b>barca</b>	54	6,3	...
zozobrar la barca			
<b>patera</b>	11	6,1	...
<b>embarcación</b>	182	6,0	...
zozobrar la embarcación			
<b>lancha</b>	32	5,4	...
zozobrar la lancha			
<b>barcaza</b>	7	5,3	...
<b>navío</b>	19	5,0	...
zozobrar el navío			
<b>canoa</b>	9	4,9	...
<b>balsa</b>	10	4,3	...
<b>velero</b>	7	4,2	...
<b>barco</b>	106	3,5	...
zozobrar el barco			
<b>nave</b>	74	3,5	...
zozobrar la nave			
<b>bote</b>	19	2,7	...
zozobrar el bote			
<b>buque</b>	22	2,6	...
buque zozobrado			
<b>cimiento</b>	6	1,3	...
<b>madrugada</b>	5	1,2	...
<b>ánimo</b>	7	0,2	...

# Los corpus como complemento del diccionario

## *zozobrar* (1)

### El diccionario dice de *zozobrar*...

hundirse [una  
embarcación]

fracasar o frustrarse [algo]

(M. Seco, O. Andrés, G. Ramos,  
DEA, s.v. *zozobrar*)

### El corpus muestra...

- una empresa zozobra
- un proyecto zozobra
- una ilusión zozobra
- la democracia zozobra
- el gobierno zozobra

***Eso es algo que ningún diccionario recoge con la misma riqueza que un corpus.***

# Los corpus como complemento del diccionario: *zozobrar* (2)

## Uso literal

*Sujeto: embarcaciones*

*yolas, petroleros,  
transbordadores...*

*yola = patera, cayuco*

## Uso figurado

*La imagen del naufragio se extiende a:*

- **Mundo interior:**

*razón, esperanza, cordura, identidad*

- **Mundo institucional:**

*repúblicas, democracias, sistemas, leyes*

- **Relaciones y proyectos:**

*matrimonios, empresas, planes, temporadas*

- **Tiempo y lenguaje:**

*la tarde, el día, las palabras*

# Concordancia de *zozobrar*: usos literales y metafóricos

## Uso literal — sujeto: embarcación

...embarcación mal equipada **zozobró** en esas aguas antes de alcanzar...

*Cuba, 2002 — narrativa*

...de julio, el Urslein **zozobra** cerca a la costa esmeraldeña...

*Ecuador, 2002 — narrativa*

## Uso metafórico — el sujeto escala en abstracción

...de enero. Entonces **zozobró** la esperanza y acepté mi destino.

*México, 2002 — narrativa*

...paralelamente, haga **zozobrar** el equilibrio económico del país.

*España, 2003 — prensa*

...la República, **zozobró** nuevamente el 18 de marzo de 1861.

*Rep. Dom., 2002 — ensayo*

*El sujeto metafórico escala de lo interior (la esperanza) → lo colectivo-económico (el equilibrio) → lo político-institucional (la República)*

# ¿Qué más nos ofrece CORPES XXI?

Además de la concordancia, CORPES XXI nos permite hacer otros tres tipos de consulta:

- 1 Inventario de formas (resultados ordenados por frecuencia)**  
*Qué variantes del lema aparecen en el corpus y cuál es la más frecuente (zozobrar, zozobró, zozobra...)*
- 2 Coapariciones**  
*Qué palabras suelen aparecer en el entorno de nuestra palabra (ver supra coapariciones de zozobrar)*
- 3 Estadística: distribución geográfica, cronológica y temática**  
*Con qué frecuencia aparece la palabra en cada país hispanohablante. Aquí es donde el tamaño del subcorpus empieza a importar...*

## Inventario: variantes del lema *zozobrar* ordenadas por frecuencia (cinco primeros resultados)

inventario (CORPES) – variantes (CDH) – frecuencia (SkE)




Formas	F <sub>Abs.</sub>	F <sub>Norm.</sub>
<b>zozobrar</b>	111	0,25
<b>zozobró</b>	43	0,09
<b>zozobra</b>	38	0,08
<b>zozobraba</b>	17	0,03
<b>zozobraron</b>	7	0,01

recordar: inventario del lema *pequeño*

# Variantes de la forma buscada en los corpus históricos

¿De cuántas formas distintas se escribe **cien** en el siglo XV?

Hacer la búsqueda del lema **cien** en el CDH nuclear

<b>Concordancias</b>		<b>Nómina</b>	<b>Ayuda</b>	<b>Modo de cita</b>
	<input checked="" type="checkbox"/> <b>CDH nuclear</b>		<input type="checkbox"/> <b>S.XII-1975</b>	
			<input type="checkbox"/> <b>1975-2000</b>	
<b>Lema</b>	<input type="text" value="cien"/>	<b>Forma</b>	<input type="text"/>	<b>Clase de palabra</b> <input type="text" value="(Todos)"/> 
				<input type="checkbox"/> <b>Graf</b>
<b>Subcorpus</b> <a href="#">Limpiar</a>				
<b>Título</b>	<input type="text"/>	<b>Autor</b>	<input type="text"/>	<b>Fecha de creación</b> <input type="text" value="1401"/> - <input type="text" value="1500"/>

# Variantes del lema *cien* en el s. XV (CDH nuclear)

## Distribución Forma

Forma	Freq	Fnorm.
çient	239	34,43
çien	74	10,66
cient	52	7,49
cien	39	5,61
Cien	1	0,14
Çient	1	0,14
1 - 6 of 6		página: 1

# Estadística: distribución geográfica del lema *zozobrar* (CORPES)

## *Frecuencia absoluta y frecuencia normalizada*

Frec. Abs.    Frec. norm.

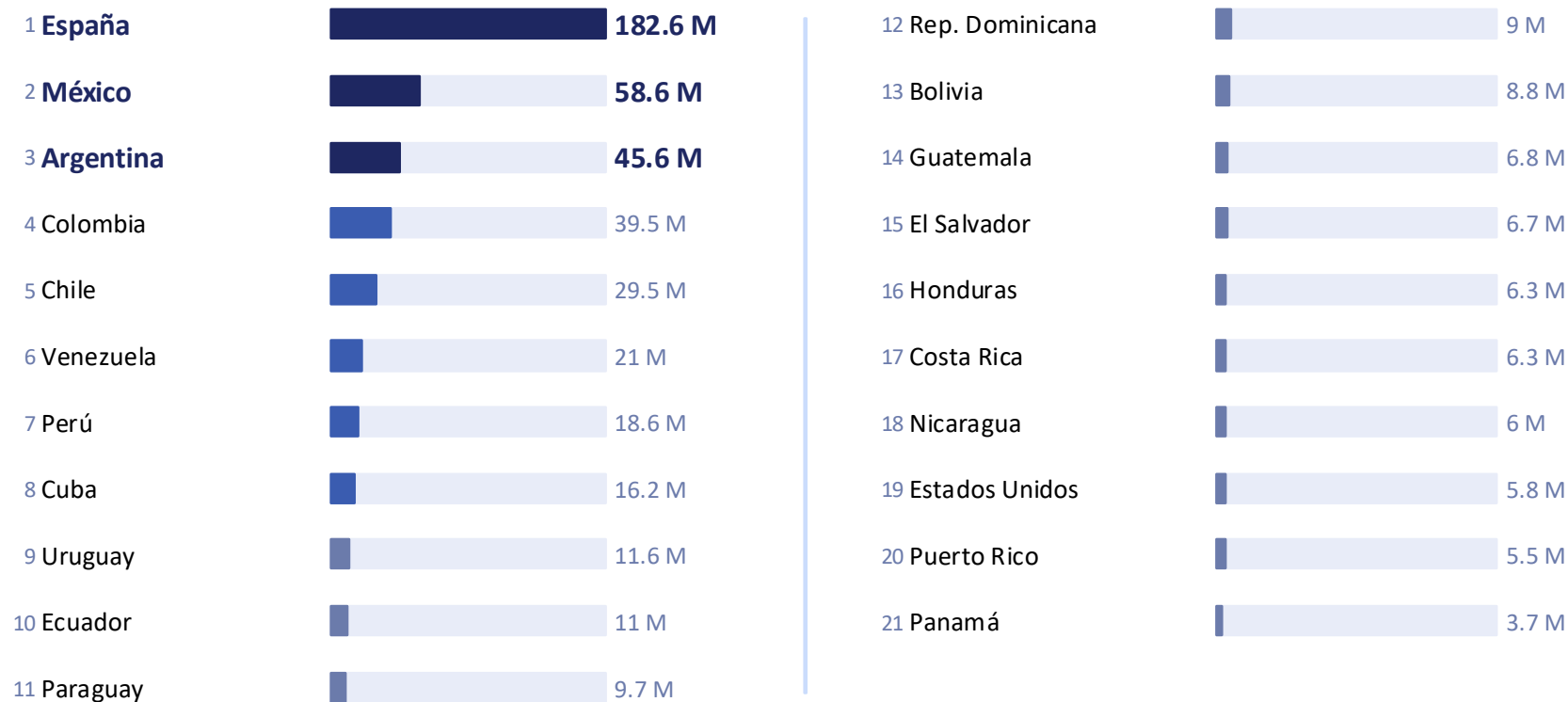
España	91	0,57
Estados Unidos	2	0,39
Guinea Ecuatorial	4	4,46
Honduras	8	1,44
México	22	0,43
Nicaragua	2	0,38
Panamá	4	1,24
Paraguay	5	0,59
Perú	6	0,37
Puerto Rico	5	1,03
República Dominicana	23	2,9

¿En qué país se usa más el verbo *zozobrar*?

¿En España o en la RD?

# Tamaño de los subcorpus nacionales de CORPES XXI

Frecuencia absoluta en millones de palabras (total ≈ 510 M)



Para comparar el uso de una palabra entre países es imprescindible usar la frecuencia por millón, no la frecuencia absoluta.

# Frecuencia absoluta y frecuencia por millón

El problema es que el subcorpus de España es mucho más grande que el de la RD. Para comparar datos de corpus de distinto tamaño necesitamos **normalizar** la frecuencia.

Calcular la frecuencia por millón (frec. relativa, frec normalizada) de *zozobrar* en España y la RD

La medida estándar habitual: frecuencia por millón

$$\left( \text{frecuencia absoluta} / \text{n}^\circ \text{ palabras del corpus} \right) \times 1\,000\,000$$

frec. observada = frec. absoluta / frec. relativa = frec. normalizada = frec. por millón

# El caso de *lindo* en Uruguay y España

Apliquemos la fórmula a la palabra *lindo* en CORPES XXI:

	Frecuencia absoluta	Tamaño del subcorpus	Frecuencia por millón
Uruguay	650	11 609 318	55,98
España	745	182 619 056	4,07

***lindo* es casi 14 veces más frecuente en Uruguay que en España**

*algo que la frecuencia absoluta ocultaba por completo*

# Obtención de los ejemplos a partir de la distribución geográfica.

## Ejemplos antillanos de *zozobrar* ordenados por país

### Concordancias

[Descargar en TSV](#)[Descargar en documento](#)[Descargar en texto](#)[Imprimir](#)

Ordenar por:

País

Cuba	2014	— Uhmmm, ¿y el barco podía	<b>zozobrar</b>	?
Cuba	2020	... una diáspora, está diseñado para	<b>zozobrar</b>	y uno descubre, muy pronto, que l...
Cuba	2024	...obación del paquete de leyes que	<b>zozobró</b>	.
P.Rico.	2006	...incess pudo estar ayer a punto de	<b>zozobrar</b>	en alta mar, después de haber deja...
P.Rico.	2007	...dolfo, conocido como el Vikingo,	<b>había</b>	zozobrado en la tormenta luego de ...

## Ver la distribución y ejemplos de *yola* por países en CORPES XXI

País	F <sub>Abs.</sub>	F <sub>Norm.</sub>
Puerto Rico	33	6,82
República Dominicana	52	6,57
Perú	30	1,86
Ecuador	3	0,31
Panamá	1	0,31
Bolivia	2	0,26
México	13	0,25
España	25	0,15
Chile	3	0,11
Cuba	1	0,07
Venezuela	1	0,05
Argentina	1	0,02
Colombia	1	0,02

ordenado por la  
F<sub>Norm</sub>

# *yola* en el DLE, DEA y el DA

Del fr. *yole*, y este de or. germ.

1. f. Embarcación muy ligera movida a remo y con vela.

SIN.: **embarcación**, **barca**.

DLE, s.v. *yola*

¿Qué matiz falta en la definición? Leer los ejemplos de la diapositiva siguiente

Buscar el primer sinónimo de *yola* en esTenTen (Tesauro)

# Ejemplos del lema *yola* en Puerto Rico (CORPES XXI)

P.Rico.	2001	...s de inmigrantes que cruzan en	<b>yola</b>	el Canal de la Mona, en busca de ...
P.Rico.	2002	.../ la lancha de Cataño porque en	<b>yola</b>	no / porque la hunde / ¿tú sabes? ...
P.Rico.	2005	...o de los cinco capitanes de una	<b>yola</b>	que desembarcó el 3 de diciembr...
P.Rico.	2005	... las 93 personas hacinadas en la	<b>yola</b>	, pagó 30,000 pesos dominicanos...
P.Rico.	2005	...los acompañando al dueño de la	<b>yola</b>	.
P.Rico.	2005	...s ilegales (el equivalente de las	<b>yolas</b>	dominicanas) y por la oleadas de ...
P.Rico.	2005	...onde a menudo son detectadas	<b>yolas</b>	con indocumentados que realiza...

# Sketch Engine: de la concordancia a Word Sketch

*Concordancias: lema, forma, frase, carácter, CQL - Búsquedas por países - Ordenar variantes por frecuencia - Gestión de demasiados ejemplos - Uso de la pleca | - Colocaciones - Word Sketch -*

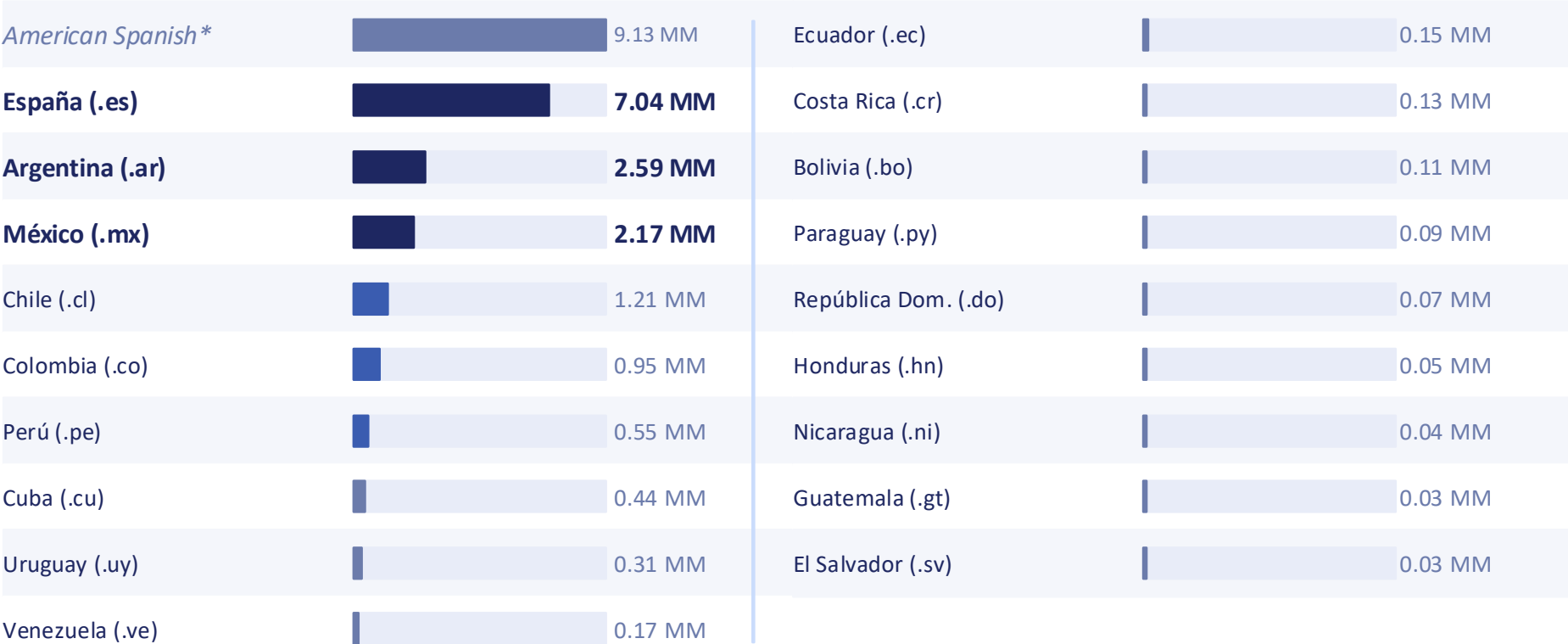
# Entramos en Sketch Engine

Sketch Engine es una plataforma de análisis de corpus. Una vez dentro, seleccionamos nuestro primer corpus: **esTenTen23**, casi 29 000 millones de palabras extraídas de la web en español (años más representados 2020-2022-2023).

Antes de buscar, conviene visitar la **información del corpus**: tamaño, distribución y subcorpus nacionales disponibles.

# Distribución de esTenTen23 por subcorpus nacionales (dominios de internet). Esto nos permite hacer **búsquedas por países**

Miles de millones de palabras (tokens) — total: ~ 29 000 M



\* "American Spanish domains" agrupa dominios no asociados a un país específico. A diferencia de CORPES XXI, esTenTen23 no us a criterios editoriales de selección: refleja la presencia de cada variedad en la web. Falta Puerto Rico como dominio web independiente en esTenTen23

# Concordancia. Tipos de consulta (1/2)

Sketch Engine ofrece dos modos de búsqueda: simple y avanzado  
Usaremos preferentemente el modo **avanzado** porque nos permite elegir más flexiblemente el tipo de consulta:

## Lema

Ejemplos del lema *fome* en el subcorpus chileno

## Forma

Una forma concreta y exacta: ejemplos de la forma *cantamañanas* en el subcorpus de España

# Ejemplos del lema *fome* en Chile (esTenTen23)

Chilean domain .cl		lema <b>fome</b> • 6926		5,74 por millón tokens • 0.000021%													
Detalles				Contexto izquierdo		KWIC		Contexto derecho									
1	<input type="checkbox"/>	jaja.cl	cuuec Ponga	palabras Nuevaas pu Pelolais 1999 jajsojsoajs que pena jajajaja ui la wea	<b>fome</b>	xd que pena las weonas picas a lais que no tienen otra v											
2	<input type="checkbox"/>	rscj.cl	oy viviendo. Muchas veces he escuchado decir que la imagen de Mater es estática, que es	<b>fome</b>	, que no invita a nada, que es como si estuviera esperan												
3	<input type="checkbox"/>	eure.cl	as siempre estamos más abajo que ellos, somos distintos. Nosotros decíamos que eso era	<b>fome</b>	, porque somos todos iguales. (Myriam, 2012) La insegui												
4	<input type="checkbox"/>	fmdos.cl	n del capitulo real ... Pero entre tanto reclame no fluía la historia .... No me gustó el final ...	<b>Fome</b>	!!!#DementeGranFinal Al debe con el final faltó mas acció												
5	<input type="checkbox"/>	fmdos.cl	i días, Repe nunca va a ser entretenido. Una persona que está adentro me dice 'es lo más	<b>fome</b>	que hay". "Aparentemente, los ensayos los están hacien												
6	<input type="checkbox"/>	fmdos.cl	arar. Júzguenme todo lo que quieran, llámenme lastimera, pobre we**, florero, pinta mono,	<b>fome</b>	, busca pantalla y todos los calificativos negativos, pero r												
7	<input type="checkbox"/>	gamba.cl	o diciendo "Qué hora será en Nueva York?" o como piñera "Qué hora será en Venezuela?"	<b>FOME</b>	CTM TE VOY A IR A MATAR COMUNISTA RECULEAO.												
8	<input type="checkbox"/>	gamba.cl	tristas, ya que tanto hablas contra Cuba, pero te vas en pura palabreria chanta culiao. Tan	<b>fome</b>	tu lenguaje propio del hampa, siempre lo mismo comunis												
9	<input type="checkbox"/>	gamba.cl	falta un diccionario ya que su unica neurona la tiene para hacer caca... El conchasumadre	<b>fome</b>	x la xuxa, si era era una ofensa entonces tienes un serio												

fome = aburrido, sin gracia (DLE)

# Ejemplos de la forma *cantamañanas* en España (esTenTen23)

European Spanish dom... x		forma <b>cantamañanas</b> • 1046		0,15 por millón tokens • 0.0000032% i											
<input type="checkbox"/> Detalles		Contexto izquierdo		KWIC		Contexto derecho									
1	<input type="checkbox"/> i armas.es	mpo de sobra a los asaltantes de robar y matar y quitarse de en medio. Políticos		<b>cantamañanas</b>		y corruptos, eso es lo que sois. Si quieren que no se usen a									
2	<input type="checkbox"/> i piomoa.es	lo desenreden otros. Vaya panda. Don jaque me prece, por un lado, un señorito		<b>cantamañanas</b>		y por otro un admirador de los terroristas, siempre que sean									
3	<input type="checkbox"/> i piomoa.es	n Kolakowski viviendo Franco, contra la beatería "antifascista" de unos laboristas		<b>cantamañanas</b>		. Ello aparte, y pese a lo que cree el oportunista Treglown, c									
4	<input type="checkbox"/> i piomoa.es	istocracia de nada, son un atajo de traidores, chorizos, ricachos sin escrúpulos y		<b>cantamañanas</b>		. Pero volviendo al principio, todo pasa por crear excelencia									
5	<input type="checkbox"/> i huarte.es	es de Huarte. COLECTIVO ARTÍSTICO ASOCIACIÓN ARTÍSTICA Y CULTURAL		<b>CANTAMAÑANAS</b>		Cantamañanas es un colectivo que centra su labor en la org									
6	<input type="checkbox"/> i huarte.es	ECTIVO ARTÍSTICO ASOCIACIÓN ARTÍSTICA Y CULTURAL CANTAMAÑANAS		<b>Cantamañanas</b>		es un colectivo que centra su labor en la organización anual									
7	<input type="checkbox"/> i cualia.es	s de quienes venden duros a peseta. ¿A quién me refiero? Pues a los arribistas y		<b>cantamañanas</b>		que nos convencen de que somos demasiado frágiles –derr									
8	<input type="checkbox"/> i jotdown.es	ser auténticos ¿No debería haber arrasado VOX con su estilo ultra de tertuliano		<b>cantamañanas</b>		? Mateo Renzi "el renovador" tiene un partido que es un rej									
9	<input type="checkbox"/> i laliamos.es	naginarías, afectadas por las más peregrinas dolencias, se enfrentan a un doctor		<b>cantamañanas</b>		y a sus dos enfermeras, que les prescribirán terapias entre l									

cantamañanas = persona informal y que no merece crédito (DEA)

# Concordancia. Tipos de consulta (2/2)

## Frase

Un sintagma o grupo de palabras: *una manga de*, en el subcorpus argentino (ver diapositiva *infra*)

## Carácter

Búsqueda por cadenas de caracteres: *bsc* en subcorpus peruano\*

## CQL

Expresiones regulares para búsquedas complejas, que veremos más adelante

\*elijo corpus peruano para hacer la consulta más rápidamente, reduciendo el tamaño

# Tipo de búsqueda: **frase** (secuencia de palabras) *una manga de*, en subcorpus argentino

## CONCORDANCIA

Spanish Web 2023 (

BÁSICO

AVANZADO

AI SEARCH

LEA

Tipo de consulta ?

simple

lema

**frase**

forma

caracter

CQL

Frase

una manga de

✓ A = a ?

Subcorpus ?

Argentinian domain .ar ▼



Macro ?

ninguno

# Fragmento de la concordancia de *una manga de* en Argentina

Argentinian domain .ar

frase una manga de

739

0,29 por millón tokens • 0.0000022%

Detalles

Contexto izquierdo

KWIC

Contexto derecho

1

ole.com.ar

un peso de Diego. En cambio todo el entorno que estuvo al lado suyo ahora son

una manga de

tránsfugas, vividores, traidores hijos de mil puta". E

2

cmtv.com.ar

¡es! ¿Cómo no querés que los cague si son unos boludos de mierda? ¡Son todos

una manga de

garcas! ¡Este país está lleno de ladrones! ¿Yo?... ¡

3

fabio.com.ar

s? Yo no conozco muchos nenes yankis, pero los latinos que veo por la calle son

una manga de

lieros y gritones insoportables. O sera que soy un v

4

scielo.org.ar

al Consulado que el 5 de noviembre "descargó en los estramuros de esta ciudad

una manga de

Piedras qe. enteramente ha dexado sin el menor fr

5

senasa.gob.ar

reando la situación. También se solicita a los productores que, en caso de avistar

una manga de

langostas, denuncie rápidamente la presencia a la:

6

copenoa.com.ar

la expresión, pero es real) como para vender a un regalado precio nuestro país a

una manga de

giles vestidos con traje y que le importan un comin

7

elmendo.com.ar

YO. Y NO ENTIENDO LA MANGA DE INFELICES Q LES PARECE CHISTOSO

UNA MANGA DE

GILES Q CREEN HACER BUENAS COSAS, ESPI

8

elmendo.com.ar

asaron por las pelotas, ya que los intereses coloniales eran más importantes que

una manga de

incivilizados montadores de camellos. La Gran Siri

9

enorsai.com.ar

uesta explicarle a mis a mis hijos que una señora decidió jugar con las vacunas,

una manga de

irresponsables, nos arruinó la vida, porque prioriza

10

latinta.com.ar

Jé lo hicieron? Una chance es que las fuerzas de seguridad estén integradas por

una manga de

tontos que cuatro meses después no logró conseg

liero = mentiroso; garca = alguien que te caga, estafador, sinvergüenza;

# Ordenar el inventario de frecuencias por la primera palabra a la derecha: ¿Qué es lo más frecuente después de *una manga de...* en Argentina?

Argentinian domain .ar ✕

simple una manga de • 1447  
0,56 por millón tokens • 0.0000044% ⓘ

⚡

## FRECUENCIA

BÁSICO AVANZADO LEARN 🎓

Selecciona un atributo y su posición en la concordancia: ?  
lema 🔍

contexto izquierdo      contexto derecho

6 5 4 3 2 1 KWIC ▼ 1 2 3 4 5 6

+

La kwic en este caso es el sintagma *una manga de*. Buscamos qué aparece a la derecha (distancia +1, es decir, primera posición del contexto derecho)

# Inventario de frecuencias a distancia +1 de *una manga de* en el subcorpus argentino

Lemma		Frecuencia
1	<input type="checkbox"/> langosta	83
2	<input type="checkbox"/> ladrón	69
3	<input type="checkbox"/> vago	40
4	<input type="checkbox"/> inútil	32
5	<input type="checkbox"/> delincuente	30
6	<input type="checkbox"/> corrupto	26

una manga de langostas = una nube de langostas = una plaga de langostas

# Análisis: *Una manga de* en el español rioplatense

Corpus argentino (.ar) de esTenTen23 — 1 418 ocurrencias

## Sentido mayoritario

### Cuantificador colectivo peyorativo

Introduce un grupo de personas al que el hablante juzga negativamente.

Equivale a «una panda de...», «una sarta de...» o «una caterva de»

Registro coloquial e informal, con fuerte carga emocional e irónica.

*Sentido recto/técnico (minoritario): manga como tubo, conducto o etapa deportiva*

## Ejemplos del corpus (esTenTen23, .ar)

son una manga de tráfugas y vividores

son todos una manga de garcas

una manga de giles vestidos con traje

una manga de infelices

una manga de irresponsables

una manga de tontos

una manga de crotos

una manga de rufianes

una manga de locos

crotos= desarrapados

# Tipo de búsqueda: **caracter**

## Buscamos la secuencia *bsc* en el subcorpus peruano (1/2)

CONCORDANCIA

Spanish Web 2023 (esTenTen23)

Obtén más almacenamiento +

Peruvian domain .pe

caracter bsc • 4004  
7,35 por millón tokens • 0.000012%

Detalles

Contexto izquierdo

KWIC

Contexto derecho

Frecuencia

# Ordenar por frecuencia de los lemas que contienen *bsc* en subcorpus peruano (2/2)

Peruvian domain .pe x

caracter **bsc** • 4004  
7,35 por millón tokens • 0.000012% i

**FRECUENCIA**

BÁSICO AVANZADO LEARN

Selecciona un atributo y su posición en la concordancia: ?

lema

contexto izquierdo 6 5 4 3 2 1 KWIC 1 2 3 4 5 6 contexto derecho

+

Agrupar por primera columna

IR

1	<input type="checkbox"/>	absceso	1522
2	<input type="checkbox"/>	obsceno	906
3	<input type="checkbox"/>	oscuro	321
4	<input type="checkbox"/>	obscenidad	152

## ***Balotaje.*** Cómo gestionar demasiados ejemplos

***balotaje*** ‘segunda votación, cuando no hay mayoría suficiente en la primera’ es un préstamo del francés *ballotage*

Al buscar el lema ***balotaje*** en esTenTen23 obtenemos dos datos inmediatos: la **frecuencia absoluta** y la **frecuencia por millón**.

# Búsqueda de la forma *balotaje* en Sketch Engine

## CONCORDANCIA

Spanish Web 2023 (esTenTen23)



BÁSICO

AVANZADO

LEARN

Tipo de consulta ?

simple

lema

frase

forma

Categoría gramatical

**cualquier**

adjective

adverb

conjunction

Forma

**balotaje**

✓ A = a ?

busca tanto mayúsculas  
como minúsculas

Frec. absoluta y  
relativa de  
la forma *balotaje*

forma **balotaje** • 20.828

0,63 por millón tokens • 0.000063%



búsqueda realizada  
15/02/2026

# Demasiados ejemplos: ¿cómo gestionarlos?

Una búsqueda en esTenTen23 puede devolver miles de ejemplos. Sketch Engine ofrece dos soluciones:

## Muestra aleatoria

El sistema selecciona un número determinado de ejemplos al azar. Es el procedimiento más riguroso para un análisis estadísticamente representativo.

## Buenos ejemplos

Sketch Engine selecciona los ejemplos más claros y típicos, los que mejor ilustran el uso central de la palabra. Útil para descripción lexicográfica o ilustrar un uso en un trabajo.

# Reducir el número de ejemplos para facilitar la consulta

reducir 20 828 ej.  
a 200, manteniendo  
la representatividad  
(o el número que  
queramos)



Contexto de

**Muestra aleatoria**

# Seleccionar buenos ejemplos automáticamente: GDEX

**GDEX** (*Good Dictionary EXamples*) ordena automáticamente los ejemplos, de mejor a peor, según criterios de legibilidad y utilidad:

1. La frase es lo suficientemente corta y autosuficiente
2. no contiene referencias externas que la dejen incompleta (*él lo hizo allí*)
3. no incluye vocabulario excesivamente técnico ni contenido controvertido

# GDEX de *balotaje*

Y el **balotaje** cultural es entre kirchnerismo y macrismo.

Ni aquí ni en otros países con sistema de **balotaje** .

Y tomó posición frente al **balotaje** del próximo domingo.

Clamor por su postulación y confianza camino al **balotaje** .

Que no le fue suficiente para llegar al **balotaje** .

Como dije desde un principio no llegaron ni al **balotaje** .

Resultado obtenido tras evaluar 2000 ejemplos (por defecto, evalúa 300)

## ***Balotaje*. Ejemplo de uso de la pleca |**

Muchas palabras tienen variantes ortográficas: *balotaje* y *ballotage*, *vermú* y *vermut*. Si buscamos solo una, perdemos los ejemplos de las demás.

La **pleca (|)** funciona como el operador 'o': busca una variante o la otra. Hacemos varias búsquedas en una sola consulta

**balotaje | ballotage**

Devuelve todos los ejemplos de ambas variantes y permite ordenarlos por frecuencia.

# Frecuencia de *balotaje* | *ballotage* en Uruguay

(6 items, 2791 Frecuencia total)

	Word	Frecuencia
1 <input type="checkbox"/>	balotaje	2544
2 <input type="checkbox"/>	ballotage	151
3 <input type="checkbox"/>	Balotaje	76
4 <input type="checkbox"/>	BALOTAJE	12
5 <input type="checkbox"/>	Ballotage	7
6 <input type="checkbox"/>	BALLOTAGE	1

## ¿ A partir de la definición del DLE, puedes decir a qué sustantivos se aplica *vívido*?

1. **adj. poét. vivaz** (ll eficaz, vigoroso).

SIN.: **vivaz, vivo, expresivo, realista<sup>1</sup>, verosímil, elocuente.**

2. **adj. poét. vivaz** (ll de ingenio agudo).

escribir  
5 sustantivos  
a los que se  
pueda aplicar  
el adj. *vívido*

### Sinónimos o afines de «vívido, da»

- **vivaz, vivo, expresivo, realista<sup>1</sup>, verosímil, elocuente.**

El DEA define *vívido* como ‘que **evoca vivamente** la realidad’

# ¿Qué puede ser *vívido* en español?

De las definiciones analizadas, parece deducirse que *vívido* significa 'intenso, que se percibe o recuerda con fuerza'

¿puede ser *vívido* un olor, un sabor, un color?

Para salir de dudas, vamos a buscar las colocaciones (coapariciones de *vívido* en esTenTen23)

# Colocaciones de *vívido* en un intervalo -3, +3

	Forma	Coocurrencias ?	C
1	<input type="checkbox"/> nítidas	178	
2	<input type="checkbox"/> descripciones	450	
3	<input type="checkbox"/> alucinaciones	160	
4	<input type="checkbox"/> pesadillas	219	
5	<input type="checkbox"/> retrato	591	
6	<input type="checkbox"/> nítidos	91	
7	<input type="checkbox"/> recuerdos	1160	
8	<input type="checkbox"/> sueños	1229	
9	<input type="checkbox"/> nítida	112	
10	<input type="checkbox"/> realistas	242	

	Forma	Coocurrencias ?	C
11	<input type="checkbox"/> brillantes	361	
12	<input type="checkbox"/> vibrantes	89	
13	<input type="checkbox"/> coloridas	84	
14	<input type="checkbox"/> colorida	79	
15	<input type="checkbox"/> colores	2360	
16	<input type="checkbox"/> realista	265	
17	<input type="checkbox"/> imaginación	469	
18	<input type="checkbox"/> descripción	561	
19	<input type="checkbox"/> Sueños	69	
20	<input type="checkbox"/> vívido	32	

# Análisis de las colocaciones de *vívido* (1/2)

¿Puede un olor o un sabor ser *vívido*?

Lógicamente sí: si *vívido* significa ‘intenso, que se percibe o recuerda con fuerza’, nada impediría decir  
*\*olor vívido* o *\*sabor vívido*.

Sin embargo, el corpus muestra que el adjetivo está casi exclusivamente restringido al dominio **visual** y al de la **memoria**  
(*recuerdo vívido, sueño vívido, imagen vívida*).

Para los otros sentidos, el español recurre a otros adjetivos:  
*penetrante, intenso, pronunciado, fuerte*.

# Análisis de las colocaciones de *vívido* (2/2)

Esto ilustra una distinción fundamental en lingüística de corpus:

**Restricción semántica:** lo que una palabra *puede* significar según el diccionario

**Restricción colocacional:** con qué palabras *de hecho* se combina en el uso real

Los diccionarios describen el significado.

Los corpus revelan el comportamiento combinatorio real.

# Word Sketch

## el perfil combinatorio de una palabra

**Word Sketch** es la herramienta más potente de Sketch Engine. Organiza las combinaciones de una palabra por **función gramatical**: qué sustantivos la modifican, qué verbos rigen, qué adverbios la acompañan... Es el **retrato combinatorio completo** de una palabra: no solo *con qué* aparece, sino *cómo y en qué función*.

*Ninguna otra búsqueda ofrece una visión tan completa del comportamiento combinatorio de una palabra.*

# Cuál es el perfil combinatorio del sustantivo *silencio*



The screenshot shows the 'WORD SKETCH' web application interface. On the left is a dark blue sidebar with various icons. The main area has a light blue background. At the top, there's a search bar with 'Spanish Web 2023 (esTenTen23)' and a magnifying glass icon. Below the search bar are four tabs: 'BÁSICO', 'AVANZADO' (which is selected), 'VISUALIZAR LISTA', and 'LEARN' with a graduation cap icon. The 'Búsqueda ?' field contains the word 'silencio'. Below it, the 'Categoría gramatical ?' dropdown is open, showing 'auto' as the selected option, with other options being 'adjective', 'adverb', 'noun', and 'verb'. To the right of the dropdown, the 'Subcorpus ?' dropdown is set to 'ningún (todo el corpu...' with a lock icon and a plus sign. Below that, 'Collocates in each relation ?' is set to '12'. 'Frecuencia mínima ?' is set to 'auto'. 'Puntuación mínima ?' is set to '0'. There is a checkbox for 'Traducir ?' which is currently unchecked. At the bottom left, there's a 'Tipos de texto ?' dropdown. A red 'IR' button is at the bottom right.

WORD SKETCH

Spanish Web 2023 (esTenTen23)

BÁSICO AVANZADO VISUALIZAR LISTA LEARN

Búsqueda ?  
silencio

Categoría gramatical ?  
auto  
adjective  
adverb  
noun  
verb

Subcorpus ?  
ningún (todo el corpu...  
Collocates in each relation ?  
12  
Frecuencia mínima ?  
auto  
Puntuación mínima ?  
0  
☐ Traducir ?

Tipos de texto ?

IR

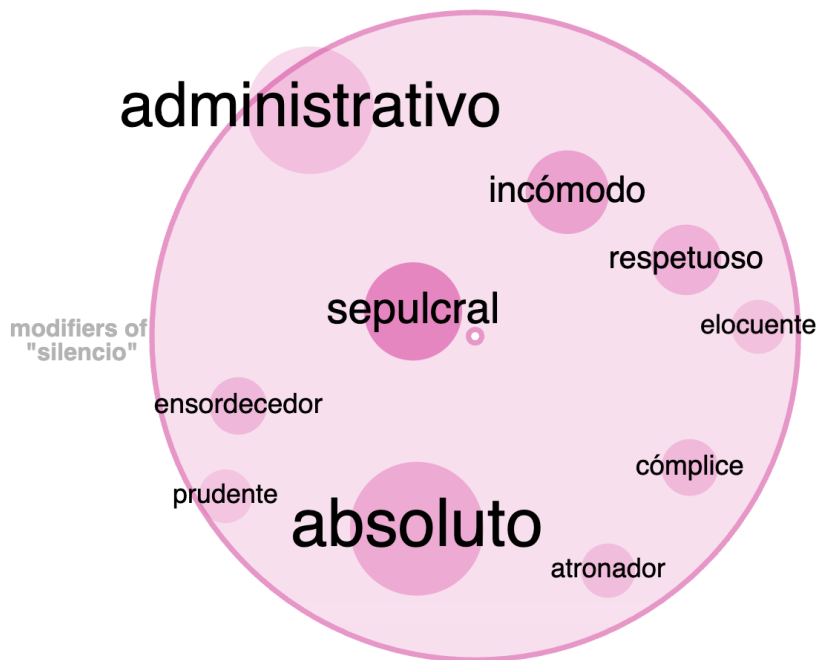
se identifican todos  
estos patrones:

silencio sepulcral  
guardar silencio  
soledad y silencio  
un minuto de silencio  
el silencio de los corderos

# Word Sketh

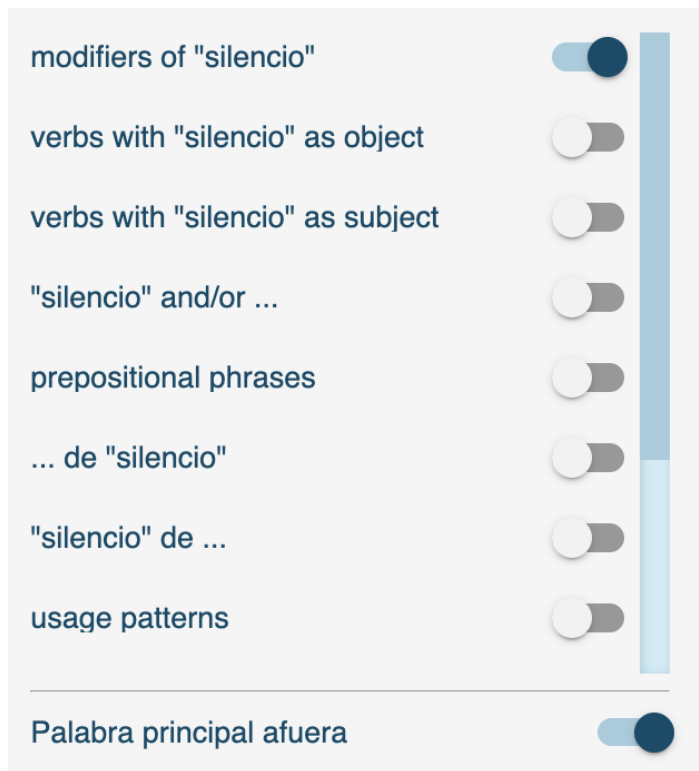
## Posibilidad de representar gráficamente las colocaciones

silencio



<b>sepulcral</b>	6965	10,1
un silencio sepulcral		
• denso en: sex ?		
<b>incómodo</b>	4476	8,7
silencio incómodo		
• denso en: culture & entertainment ?		
<b>absoluto</b>	15.275	8,4
silencio absoluto		
• denso en: travel & tourism ?		
• denso en: sex ?		
• denso en: fiction ?		
<b>respetuoso</b>	2712	8,1
un respetuoso silencio		
• denso en: religion ?		
• denso en: blog ?		
<b>ensordecedor</b>	1392	7,8
un silencio ensordecedor		
<b>cómplice</b>	1340	7,7
silencios cómplices		
<b>atronador</b>	1143	7,5
silencio atronador		
• denso en: blog ?		
<b>administrativo</b>	13.566	7,3

# Cómo se ha hecho el gráfico de *silencio*



## Frecuencia y fuerza (tipicidad) de una colocación silencio sepulcral (tipicidad)- silencio administrativo (frecuencia)

Frecuencia = cuántas veces aparece la combinación *silencio sepulcral* en el corpus  
en el gráfico se corresponde con el tamaño de la letra

Tipicidad o fuerza de la colocación = Se mide con un índice denominado LogDice,  
cuya puntuación máxima es 14

En el gráfico se corresponde con la cercanía al punto central

Una puntuación baja significa que las palabras (o una de las palabras)  
de la colocación también se combinan frecuentemente con otras muchas palabras

Metafóricamente, la tipicidad mide la **fidelidad** (la fuerza de la atracción)  
de los miembros de la colocación

# ¿Qué puede ser *halagüeño* en español?

***halagüeño***. El diccionario nos dice que significa ‘que halaga’ o ‘que promete algo favorable’. Pero ¿qué sustantivos califican los hablantes como halagüeños? Word Sketch da una respuesta inmediata y objetiva:

perspectiva

panorama

resultado

balance

pronóstico

Y casi siempre en contextos negativos:

**poco halagüeño · nada halagüeño**

ver frec.  
de poco y nada  
halagüeño

1

# Dominio semántico de la prospección temporal

Los colocativos de mayor puntuación pertenecen al campo de la **proyección hacia el futuro**

El adjetivo no se distribuye libremente por el léxico: selecciona casi con exclusividad sustantivos que denotan anticipación o evaluación de escenarios venideros.

sirve para evaluar lo que está por venir

Sustantivos colocativos  
con mayor puntuación

**futuro**  
**porvenir**  
**horizonte**  
panorama  
perspectiva  
previsión  
pronóstico  
predicción  
expectativa  
presagio

## 2

## El uso predominante es la negación

**Paradoja:** aunque *halagüeño* es etimológicamente un adjetivo de valoración positiva («que halaga, que da esperanzas»), los datos muestran que sus modificadores más frecuentes son cuantificadores negativos.

Las concordancias más representativas en posición predicativa son:

*«perspectivas no son halagüeñas»*

*«panorama no es muy halagüeño»*

*«futuro no es nada halagüeño»*

### Función pragmática dominante:

eufemismo de la amenaza — el hablante elige «no halagüeño» en lugar de «malo» o «catastrófico» para suavizar una evaluación negativa.

Modificador	Frecuencia	Score
nada	3.005	6,9
poco	2.701	3,6
menos	685	1,6
demasiado	643	2,4

*Cuantificadores negativos con mayor frecuencia y score*

## 3

## Registro formal-periodístico: marca diafásica

uso preferente en dominios como *news*, *economy & finance & business*; escasa presencia en registros coloquiales

aunque *muy* tiene una frecuencia bruta altísima (3 301), su puntuación es baja (0,8).

afinidad combinatoria con adverbios en -mente

### Adverbios colocativos

**moderadamente** logDice: 2,1  
→ analítico

**escasamente** logDice: 1,8  
→ analítico

**francamente** logDice: 1,5  
→ valorativo

**muy** score: 0,8  
→ coloquial (frec. 3301)

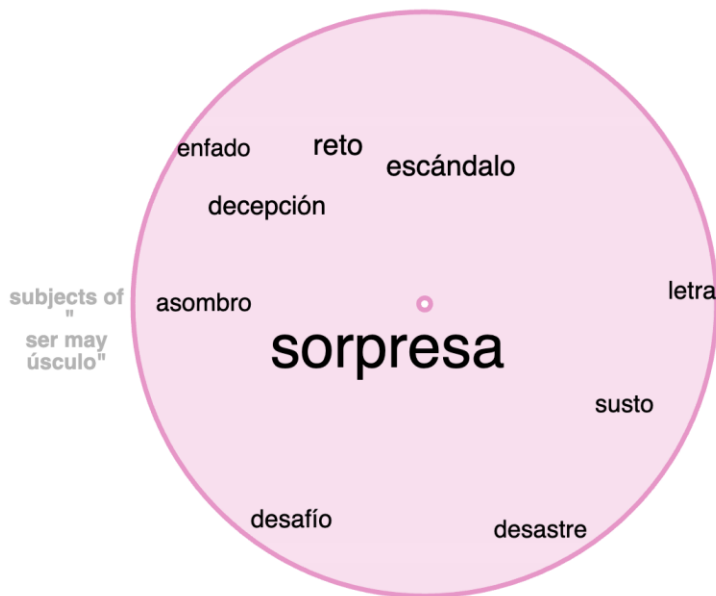
# Práctica de Word Sketch: mayúsculo · equitativo · empedernido

Para cada una, buscar las colocaciones o hacer el Word Sketch en esTenTen23:

- ¿Con qué sustantivos se combina habitualmente?
- ¿Hay sustantivos con una tipicidad muy alta, que casi exigen esa palabra?
- ¿Coincide lo que dice el corpus con lo que dice el diccionario?

# Perfil combinatorio de *mayúsculo*

mayúsculo



Número de colocados (10)

Menos

Más

modifiers of "mayúsculo"



"mayúsculo" and/or ...



prepositional phrases



subjects of "ser mayúsculo"



verbs before "mayúsculo"



nouns modified by "mayúsculo"



usage patterns



Palabra principal afuera



Mostrar círculos de palabras



Indicar puntuación con círculos



Tamaño del fuente dinámico



Relaciones separadas

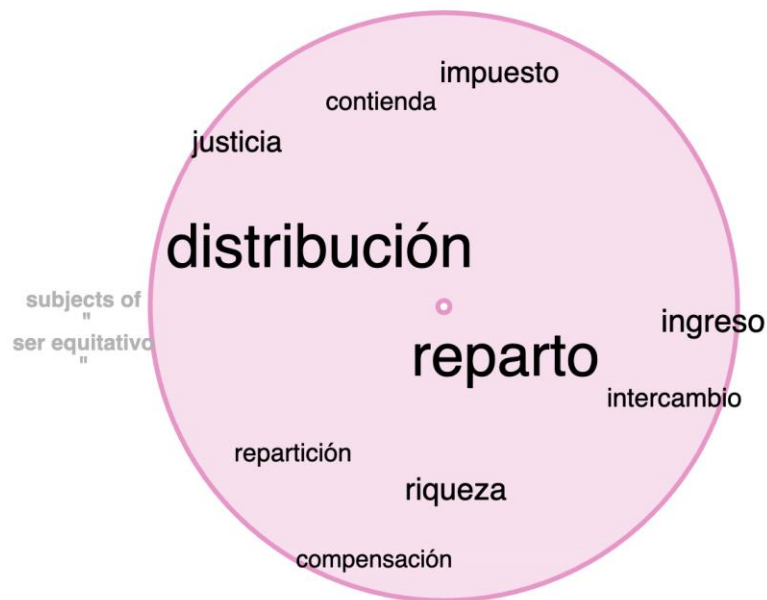


Mostrar nombres de relaciones



# Perfil combinatorio de *equitativo*

equitativo



Número de colocados (10)

Menos

Más

modifiers of "equitativo"



"equitativo" and/or ...



prepositional phrases



subjects of "ser equitativo"



verbs before "equitativo"



nouns modified by "equitativo"



usage patterns



Palabra principal afuera



Mostrar círculos de palabras



Indicar puntuación con círculos



Tamaño del fuente dinámico



Relaciones separadas



Mostrar nombres de relaciones



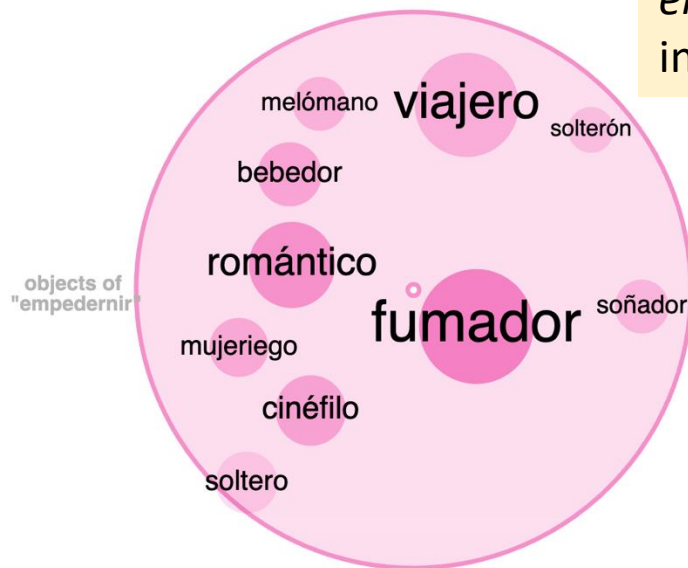
Ángulos iguales



# Perfil combinatorio de empedernido

## OJO: lema: *empedernir* en esTenTen23

empedernir



no aparece el lema  
*empedernido*, sino el  
infinitivo *empedernir*

Número de colocados (10)

Menos

Más

usage patterns



objects of "empedernir"



"empedernir" and/or ...



prepositional phrases



wh-words following "empedernir"



adjectives after "empedernir"



Palabra principal afuera



Mostrar círculos de palabras



Indicar puntuación con círculos



Tamaño del fuente dinámico



Relaciones separadas



Mostrar nombres de relaciones



Ángulos iguales



## Subcorpus nacionales de esTenTen23: volvemos a *lindo*

¿Coinciden las frecuencias por millón de *lindo* en España / Uruguay  
en CORPES XXI y esTenTen23?

	CORPES XXI	esTenTen23
Uruguay	55,98	61,5
España	4,07	11,54

**Cuando dos corpus distintos apuntan en la misma dirección, el resultado es  
más fiable.**

# Corpus propios y palabras clave

*La novela incógnita*

# Material complementario

Material necesario para este bloque:  
archivo: novela\_incognita.pdf

# ¿Por qué crear un corpus propio?

**Corpus propio** = núcleo de la investigación      **Corpus predeterminados** =  
corpus de control

# Cómo crear un corpus propio en Sketch Engine

El proceso tiene tres pasos:

1

## Reunir los textos

Sketch Engine acepta casi cualquier formato: PDF, Word, `txt`, HTML...

2

## Subir los textos

La plataforma los procesa automáticamente: los etiqueta morfosintácticamente y los prepara para hacer búsquedas.

3

## Hacer búsquedas

Exactamente igual que sobre cualquier corpus predeterminado.

*Límite en la versión estándar: 1 millón de palabras.*

# Creación de un corpus propio (1/7)

Estamos creando un corpus de 1 solo texto, lo que no suele ser habitual. Lo hacemos así solo para explicar el proceso y ejemplificar el potencial de la extracción de palabras clave. Se añaden más textos simplemente arrastrándolos a la ventana (ver diapositiva siguiente)

Crea su propio corpus de textos de la web o de tus propios documentos.

Nombre

Novela\_incognita

Tipo de corpus

☒ Corpus de un idioma  
☐ Corpus multilingüe

Idioma

Spanish

Descripción

Espacio de almacenamiento usado 6,738,960 de 9,000,000 palabras (74%)

Funcionalidades disponibles ▾

ATRÁS

SIGUIENTE

# Creación de un corpus propio (2/7)

**CORPUS:** Novela\_incognita (Spanish)

CREAR CORPUS > AÑADIR TEXTOS > COMPILAR



**Encuétrame textos en el internet**

Te descargamos textos pertinentes de páginas web.



**Tengo textos**

Sube tus archivos (.txt, .pdf,...) o pega textos

← SUBIR ARCHIVOS

[o pega texto](#)



**Selecciona un archivo o arrastra aquí.**

máximo de archivos: 100  
máximo tamaño de archivo: 500MB

Puedes subir: .csv, .doc, .docx, .htm, .html, .ods, .pdf, .tar.bz2, .tar.gz, .tel, .tgz, .tmx, .txt, .vert, .xlf, .xliff, .xml, .zip

# Creación de un corpus propio (3/7)

## CONTENIDO DEL CORPUS

Carpeta	Palabras
 upload	~26.238 

Espacio de almacenamiento usado 6.765.198 de 9.000.000 palabras (75%)

[ATRÁS](#)

[SIGUIENTE](#)

# Creación de un corpus propio (4/7)

CREAR CORPUS

Novela\_incognita



[Obtén](#)

CORPUS: Novela\_incognita (Spanish)

[CREAR CORPUS](#) > [AÑADIR TEXTOS](#) > [COMPILAR](#)

Listo

Haz clic en COMPILAR para finalizar.

[AÑADIR MÁS TEXTOS](#)

[COMPILAR](#)

Opciones expertas ▾

Log ▾

# Creación de un corpus propio (5/7)

## Compilado

Compilación terminada. El corpus está listo.

AÑADIR MÁS TEXTOS

RECOMPILAR

TABLERO DE CORPUS

# Creación de un corpus propio (6/7)



## TABLERO

Novela\_incognita



### NOVELA\_INCOGNITA

CORPUS INFO

GESTIONAR CORPUS



#### Word Sketch

Colocaciones y combinaciones de palabras



#### Diferencia Sketch

Comparar dos palabras a través de colocaciones



#### Tesauro

Sinónimos y palabras parecidas



#### Concordancia

Ejemplos de uso en contexto

# Creación de un corpus propio (7/7)

## CUENTA

Tokens	29.309
Palabras	26.238
Oraciones	415
Documentos	1

## LEYENDA DE ETIQUETAS

adjective	A.*
adverb	R.*
conjunction	C.*
determiner	D.*
noun	N.*
numeral	Z.*
preposition	S.*
pronoun	P.*

# La novela incógnita

## *Un experimento de corpus*

Hemos subido una novela a Sketch Engine sin revelar su título ni su autor. El objetivo es responder a dos preguntas usando la lista de palabras clave de la novela:

**¿De qué trata la novela?**

**¿De dónde es su autor?**

# ¿Qué es una palabra clave? El índice keyness

Una palabra es **clave** cuando su frecuencia en nuestro corpus es inusitadamente alta comparada con la que tiene en un corpus de referencia (corpus de control)

## Corpus de estudio

Nuestra novela incógnita

## Corpus de control

esTenTen23 - el español general

El índice keyness mide esa diferencia. Cuanto más alta es la puntuación, más característica es la palabra de nuestro texto respecto al español general.

# Simple maths: fórmula para calcular el índice keyness (palabra clave)



Palabras clave  
Extracción de terminología

Palabras con una frecuencia inusitadamente alta en el corpus de estudio  
respecto de la que tiene en el corpus de control

Corpus de estudio (corpus focus, de enfoque) / corpus de control (de referencia)  
novela incógnita (nuestro corpus de estudio) / esTenTen23 (corpus de control)

ÍNDICE KEYNESS (fórmula para hallar las palabras claves en  
Sketch Engine)

valor por defecto de  $n$   
 $n = 1$

$$\frac{fpm_{focus} + n}{fpm_{ref} + n}$$

# El parámetro n: constante de suavizado

$$( \text{fpm\_focus} + n ) / ( \text{fpm\_ref} + n )$$

¿Qué hace n en la fórmula?

- **n=1** → "qué palabras *raras* tiene mi corpus que no tiene el otro"
- **n=100** → "qué palabras *normales* usa mi corpus mucho más que el otro"

En SkE, modificar el valor de "Enfocarse" (por defecto = 1)

Es conveniente probar con 10, 100, 1000, etc. Pasamos de palabras "raras" (es decir, muy difíciles de encontrar en el corpus de control) a palabras "normales" más frecuentes en mi corpus que en el de control

# Las palabras clave de la novela incógnita

Al aplicar el índice *keyness* obtenemos una lista de palabras ordenadas por su puntuación: las primeras son las más características de la novela respecto al español general.

**Vamos a analizarlas una a una con ayuda de CORPES XXI.**

# Palabras clave de la novela\_incógnita

## PALABRAS CLAVE

Novela\_incognita



Obtén más almacenamiento +



PALABRAS ÚNICAS ✓

MULTI-WORD TERMS ✓



corpus de referencia: Spanish Web 2018 (esTenTen18)

(ítems: 2994)

Lema
1 moya ...
2 negroide ...
3 pupusas ...
4 mugroso ...
5 tolín ...
6 sicópata ...
7 vega ...
8 regordete ...
9 miramonte ...
10 cerote ...

Lema
11 cochinada ...
12 embrutecedor ...
13 sombreroos ...
14 vomitivo ...
15 prostíbulo ...
16 bernhard ...
17 edi ...
18 asco ...
19 ivo ...
20 zancudo ...

Lema
21 diarreico ...
22 sombrerudo ...
23 trabanino ...
24 defecación ...
25 tchaikovski ...
26 velorio ...
27 asquerosidad ...
28 asqueroso ...
29 famélico ...
30 sordidez ...

Lema
31 chunches ...
32 abominable ...
33 podás ...
34 seboso ...
35 imbécil ...
36 proceres ...
37 excremento ...
38 energúmeno ...
39 mugre ...
40 colitis ...

Lema
41 montreal ...
42 orín ...
43 engendro ...
44 inmundo ...
45 cervecería ...
46 imbecilidad ...
47 zoquete ...
48 detestable ...
49 repulsivo ...
50 maleante ...

## Palabras clave que nos ayudan a localizar geográficamente la novela

### Combinar utilidades de esTenTen23 y CORPES XXI

pupusas

cerote

mugroso

podás

zancudos

chunches

# Objetivo de la investigación

Determinar de qué trata la novela  
y en qué país se desarrolla a partir  
del análisis de las palabras clave

Vamos a usar dos herramientas:

Sketch Engine para crear el corpus  
(novela\_incógnita)

CORPES XXI para obtener  
información complementaria (distribución  
geográfica) de las palabras clave

## Ver ejemplos de *pupusas* en la novela *incógnita* (corpus de enfoque)

2	negroide	...	12	embrutecedor	...	22	sombren
3	pupusas						in
4	mugroso						ic
5	tolín						ov
6	sicópata						)



Concordancia (corpus de enfoque)



Concordance with macro



Concordancia (corpus de referencia)



# Fragmento de la concordancia de pupusas en la novela\_incógnica

¡ tortillas grasosas rellenas de chicharrón que la gente llama **pupusas** , como si esas pupusas me produjeran a mí algo más que diarrea, ¡  
as de chicharrón que la gente llama pupusas, como si esas **pupusas** me produjeran a mí algo más que diarrea, como si yo pudiera disfr  
a boca ese sabor verdaderamente asqueroso que tienen las **pupusas** , Moya, nada más grasoso y dañino que las pupusas, nada más su  
nen las pupusas, Moya, nada más grasoso y dañino que las **pupusas** , nada más sucio y perjudicial para el estómago que las pupusas, n  
usas, nada más sucio y perjudicial para el estómago que las **pupusas** , me dijo Vega.</s><s>Sólo el hambre y la estupidez congénitas pu  
comer con semejante fruición algo tan repugnante como las **pupusas** , sólo el hambre y la ignorancia pueden explicar que estos sujetos  
orancia pueden explicar que estos sujetos consideren a las **pupusas** como su plato nacional, Moya, escúchame bien, nunca se te vaya a  
loya, escúchame bien, nunca se te vaya a ocurrir criticar las **pupusas** , nunca se te vaya a ocurrir decir que se trata de una comida repug  
ños viven en Estados Unidos soñando con sus repugnantes **pupusas** , deseando tan ardientemente comer sus diarreicas pupusas que h

# Información variada sobre *pupusas*

III. (Del nahua *puxahuac*, cosa fofa o esponjada).

1. f. *Gu, Ho, ES, Ni*. **Tortilla** cocida o frita hecha de harina de maíz o arroz, *rellena generalmente de chicharrón molido, queso o flor de loroco*.

(Dicc. de americanismos, s.v. *pupusa*)

**una sola palabra** clave permite identificar la novela como centroamericana, probablemente salvadoreña



# Fragmento de la concordancia de cerote en la novela\_incógnica

/ sólo me percataba de que en cada frase incluían la palabra **cerote** , me dijo Vega.</s><s>Nunca he visto gente con más excre  
en la boca que la de este país, Moya, no en balde la palabra **cerote** es su principal metililla de lenguaje, no tienen en la boca o  
l metililla de lenguaje, no tienen en la boca otra palabra que **cerote** , su vocabulario se limita a la palabra cerote y sus derivados  
otra palabra que cerote, su vocabulario se limita a la palabra **cerote** y sus derivados: cerotísimo, cerotear, cerotada.</s><s>Incr  
e de nombre Juancho a quien miro por primera vez me llame **cerote** con familiaridad, detesto con especial intensidad que un fer  
tero negroide a quien acabo de conocer me diga a cada rato **cerote** , que me llame cerote como si yo fuese una porción de exc  
acabo de conocer me diga a cada rato cerote, que me llame **cerote** como si yo fuese una porción de excremento humano expe  
y a su negroide amigo ferretero se les pudo ocurrir llamarme **cerote** constantemente y con la mayor familiaridad, en momentos

# Información sobre *cerote* en el DLE

1. m. Mezcla de pez y cera, o de pez y aceite, que usan los zapateros para encerar los hilos con que cosen el calzado.

SIN.: *cerapez*.

2. m. coloq. **miedo** (llangustia por un riesgo).

SIN.: *miedo, pánico, temor, pavor, repelús, canguelo, canguis, culillera, culillo, culío*.

3. m. coloq. *C. Rica, El Salv., Guat. y Nic.* Excremento sólido.

4. m. *El Salv. y Nic. U.* como ofensa o insulto.

5. m. coloq. *Ven.* Suciedad acumulada en algunas partes del cuerpo.

# Datos complementarios de *cerote*, obtenidos en internet

1. También es una forma de saludar a los amigos (cheros).
2. Dependiendo de la entonación es un saludo de amigos o una ofensa.
3. excremento recién hecho.

palabra con muchos significados ambiguos.

*1. Hey! que paso cerote!. (expresión al encontrarse con un amigo).*

*2. Si seguís molestando te rompere la trompa cerote.*

**Sinónimos :**

**Saludo**

**Ofensa**

**Excremento**

Enviado por : Guillermo Martinez (San Salvador, El Salvador) 2008-09-09 06:01pm

cerote, Guatemala y Honduras; utilizado también como vocativo:  
"dispará, pues, cerote, si es que tenés güevos" (CORPES XXI,  
Honduras)

# mugroso / mugriento en CORPES XXI

- Buscar *mugroso* (como lema) en corpes XXI y determinar su extensión geográfica.
- Comparar la frecuencia de *mugroso* con la de *mugriento*

Zona	F <sub>Abs.</sub>	F <sub>Norm.</sub>
México y Centroamérica	166	2,01
Andina	39	1,17
Caribe continental	45	0,85
Antillas	12	0,44
Río de la Plata	23	0,39
Chilena	6	0,23
España	13	0,08

Zona	F <sub>Abs.</sub>	F <sub>Norm.</sub>
Guinea Ecuatorial	2	2,23
Río de la Plata	126	2,18
Chilena	40	1,56
Andina	47	1,41
España	217	1,37
Antillas	35	1,3
Caribe continental	64	1,21
México y Centroamérica	79	0,96

# *Podás, con tilde, en CORPES XXI (1/2)*

## Búsqueda

---

Elementos gramaticales ▼

Lema

Forma

podás



Sensibilidad



Acentos



Mayúsculas

# Podás en CORPES XXI (2/2)

País	F <sub>Abs.</sub>	F <sub>Norm.</sub>
Nicaragua	16	3,05
Costa Rica	8	1,45
El Salvador	6	1,03
Guatemala	5	0,84
Honduras	4	0,72
Colombia	10	0,28
Bolivia	2	0,26
Argentina	2	0,05
España	2	0,01

# Fragmento de la concordancia de *zancudos* en El Salvador (CORPES XXI)

...neta. Ni siquiera las moscas o los	<b>zancudos</b>	, paradigma de terquedad, se pudie...
... vez como quien aparta chingones	<b>zancudos</b>	que se han atrevido a incursionar e...
...stados percibe que la causa es "el	<b>zancudo</b>	".
...edad provocada por "la picada del	<b>zancudo</b>	". Aunque nos encontramos alguno...
...el humo de los buses, el calor, los	<b>zancudos</b>	y el ANDA, a la legua se les nota qu...
...do cerca de alguien, o me pica un	<b>zancudo</b>	... y ... (llora.)
..."Es la única forma de espantar los	<b>zancudos</b>	y los jejenes", dice, pero, tal y com...
...Los manglares están atestados de	<b>zancudos</b>	y jejenes. Fumar puros es una de la...

# Frecuencia del lema *zancudo* en CORPES XXI

País	F <sub>Abs.</sub>	F <sub>Norm.</sub>
Honduras	86	15,49
El Salvador	72	12,41
Nicaragua	42	8
Filipinas	1	7,14
Guatemala	35	5,88
Colombia	139	4,02
Venezuela	62	3,39
Perú	42	2,61
Costa Rica	13	2,36
Chile	58	2,27
Ecuador	15	1,57

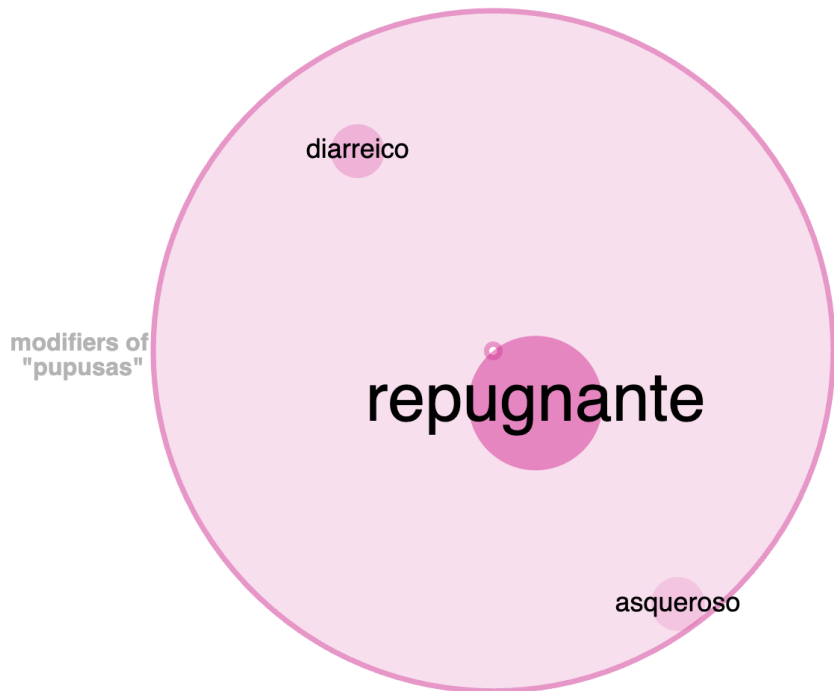
# Definición de *chunches* (DLE) y frecuencia en CORPES XXI

1. *m. C. Rica, Guat., Hond., Nic. y Pan.* Objeto cuyo nombre se desconoce o no se quiere mencionar.

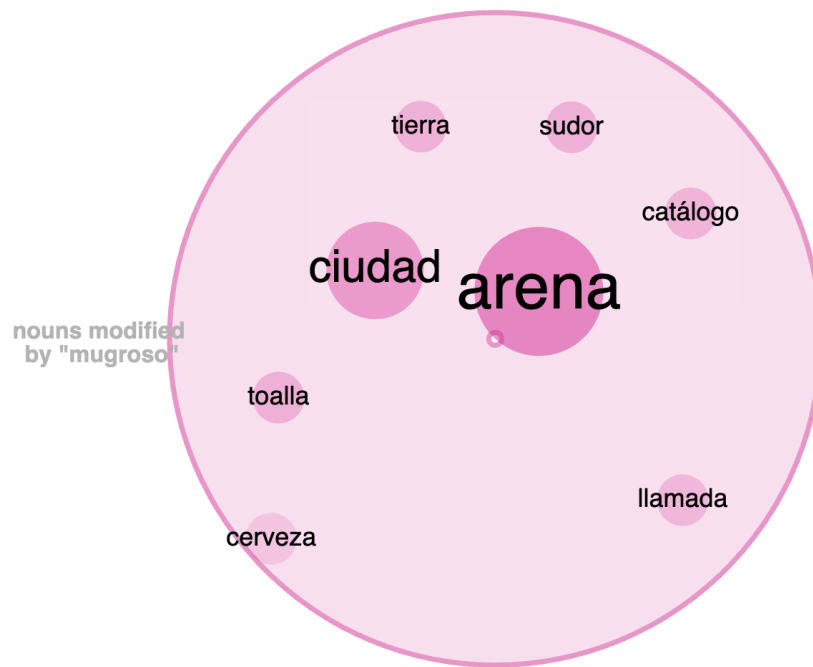
País	F <sub>Abs.</sub>	F <sub>Norm.</sub>
Nicaragua	18	3,51
Costa Rica	14	2,63
Guatemala	13	2,26
Honduras	4	0,76
El Salvador	4	0,73
México	8	0,16
Venezuela	1	0,05
España	1	0

# Representación gráfica de palabras claves

pupusas



mugroso



# Combinaciones clave en la novela\_incógnica

PALABRAS ÚNICAS ✓

EXPRESIONES MULTIPALABRAS ✓

↔ corpus de referencia: Spanish Web 2018 (esTenTen18) (items: 3432)

Término	Término	Término	Término
1 pasaporte canadiense ...	14 casa amurallada ...	27 semen cristalizado ...	40 compañero de colegio ...
2 página social de los periódicos ...	15 casa de la colonia ...	28 cóctel de conchas ...	41 universidad privada ...
3 página social ...	16 granada de fragmentación ...	29 muelle destartalado ...	42 palabra cerote ...
4 degradación del gusto ...	17 proceres de la patria ...	30 par de whiskies ...	43 partida de zoquetes ...
5 gente de sociedad ...	18 llamado proceres ...	31 música folclórica latinoamericana ...	44 porción de excremento ...
6 venta de la casa ...	19 manifestación del espíritu ...	32 carrera de historia ...	45 semen cristalizado en las baldosas ...
7 hermano marista ...	20 ciudadano canadiense ...	33 colitis nerviosa ...	46 centenar de sujetos ...
8 colonia miramonte ...	21 aventura sexual ...	34 propio camarada ...	47 amigo negroide ...
9 hermano ivo ...	22 administrador de empresas ...	35 caja repleta ...	48 causa vomitiva ...
10 tío edi ...	23 sala de migración ...	36 alteración nerviosa ...	49 porción de excremento humano ...
11 cuento famélico ...	24 tortilla grasosa rellena ...	37 grupo de rock ...	50 gente impuntual ...
12 telenovela mexicana ...	25 música llorona ...	38 espectáculo artístico ...	
13 sicópata criminal ...	26 tortilla grasosa ...	39 música folclórica ...	

## Conclusiones (1/2)

Buscamos cada palabra en CORPES XXI: ¿en qué países hispanohablantes aparece y con qué frecuencia?

pupusas

alimento típico de El Salvador

cerote

Guatemala, Honduras y El Salvador; uso como vocativo coloquial muy expresivo

mugroso

mucho más frecuente que *mugriento* en Centroamérica

## Conclusiones (2/2)

**podás**

voseo con tilde en subjuntivo,  
diferencia Centroamérica del Río de la Plata

**zancudos**

mosquito; distribución claramente centroamericana

**chunches**

cosas, trastos; concentrado en México y Centroamérica

**Todas las palabras apuntan al mismo lugar: El Salvador**

# El misterio resuelto

## El asco: Thomas Bernhard en San Salvador (1997)

*Horacio Castellanos Moya (El Salvador)*

Seis palabras clave fueron suficientes para identificar el origen geográfico y la temática de la novela, sin haberla leído.

Origen geográfico identificado

Temática aproximada

Idiolecto y preferencias léxicas del autor

# Crear corpus automáticamente en Sketch Engine desde la web

*Gastronomía peruana (ceviche, tiradito, causa limeña, lomo saltado)*

# Material complementario

Material necesario para este bloque:  
archivo: `gastronomia_peruana.txt`

# Seleccionar palabras claves para generar corpus en la web

paso 1

CREAR CORPUS

prueba5



CORPUS: prueba5 (Spanish)

CREAR CORPUS > AÑADIR TEXTOS > COMPILAR



Encuéntrame textos en el internet

Buscar y descargar automáticamente textos relevantes

paso 2

← TEXTOS DE INTERNET

Tipo de entrada

☒ Búsqueda web ?

☐ Las URL ?

☐ Sitio web ?

tiradito, ceviche, lomo saltado,  
causa limeña



# Palabras clave del corpus gastronomía-peruana

Lema	Lema	Lema	Lema	Lema
1 aji ...	11 nikkei ...	21 maracuyá ...	31 acurio ...	41 cebollines ...
2 ceviche ...	12 aji ...	22 sour ...	32 palta ...	42 salsa ...
3 huancaína ...	13 ceviches ...	23 mariscos ...	33 carapulcra ...	43 brasa ...
4 choclo ...	14 camarón ...	24 filete ...	34 puré ...	44 pescado ...
5 limeño ...	15 anticuchos ...	25 camarones ...	35 tacu ...	45 marisco ...
6 chaufa ...	16 cilantro ...	26 pulpo ...	36 centolla ...	46 apanar ...
7 pisco ...	17 chicharrón ...	27 lomo ...	37 tallarín ...	47 bandejas ...
8 rocoto ...	18 tiradito ...	28 corvina ...	38 peruano ...	48 chalaquita ...
9 cebiche ...	19 camote ...	29 fettuccini ...	39 mayonesa ...	49 piscomar ...
10 tirado ...	20 chifa ...	30 yuca ...	40 chalaca ...	50 tallarin ...

# n-gramas clave del corpus gastronomía-peruana

Término		
1	ají amarillo	...
2	causa limeña	...
3	lomo saltado	...
4	cocina peruana	...
5	leche de tigre	...
6	ají de gallina	...
7	restaurante peruano	...
8	gastronomía peruana	...
9	arroz con choclo	...
10	pasta de ají	...

Término		
11	comida peruana	...
12	punto de ají	...
13	papa a la huancaína	...
14	huevo duro	...
15	zarza criolla	...
16	causa de camarones	...
17	pasta de ají amarillo	...
18	arroz chaufa	...
19	pisco sour	...
20	pescado a la plancha	...

Término		
21	salsa de soya	...
22	ají limo	...
23	queso fresco	...
24	salsa criolla	...
25	trozo de pescado	...
26	salsa especial	...
27	jugo de limón	...
28	vacuno a la plancha	...
29	limón de pica	...
30	salsa golf	...

Término		
31	plato frío	...
32	cocina nikkei	...
33	cebolla morada	...
34	pollo a la brasa	...
35	puré de papas	...
36	dragón en salsa	...
37	diente de dragón en salsa	...
38	jugo de limón de pica	...
39	salsa huancaína	...
40	diente de dragón	...

# Word Sketch del lema *tirado* tiradito de...

tirado



**Cebiche** preparado con  
pescado cortado en  
láminas muy finas  
(Dicc. americanismos, s.v.  
*tiradito*)

**Tiradito** : se trata de una versión del ceviche... (fragmento tomado de la concordancia)

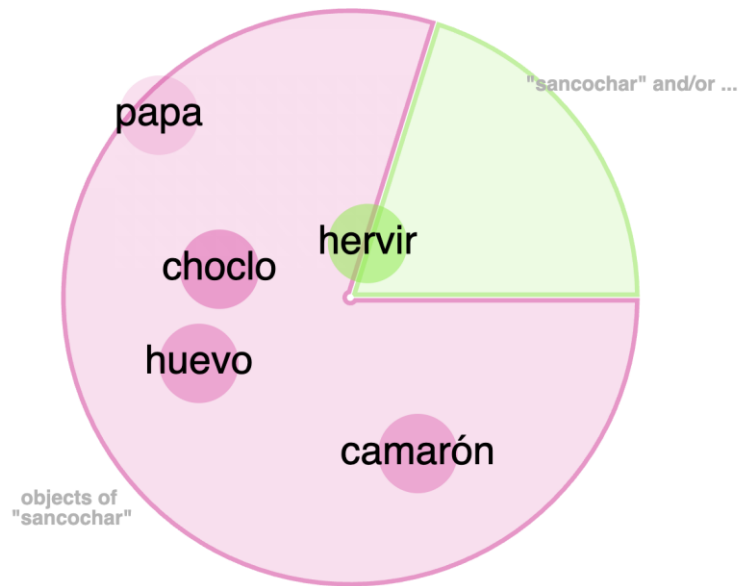
## *Sancochar en el Diccionario de americanismos*

tr. *Gu, Ho, ES, Ni, CR, Pa, Cu, RD, PR, Co, Ve, Ec, Pe, Bo, Ch, Py, Ur.* Cocer, especialmente verduras y carnes, con sal en agua hirviendo. pop + cult. ([salcochar](#)).

Mx. Sofreír un alimento.

# *Sancochar.* Información procedente del corpus gastronomía-peruana

sancochar



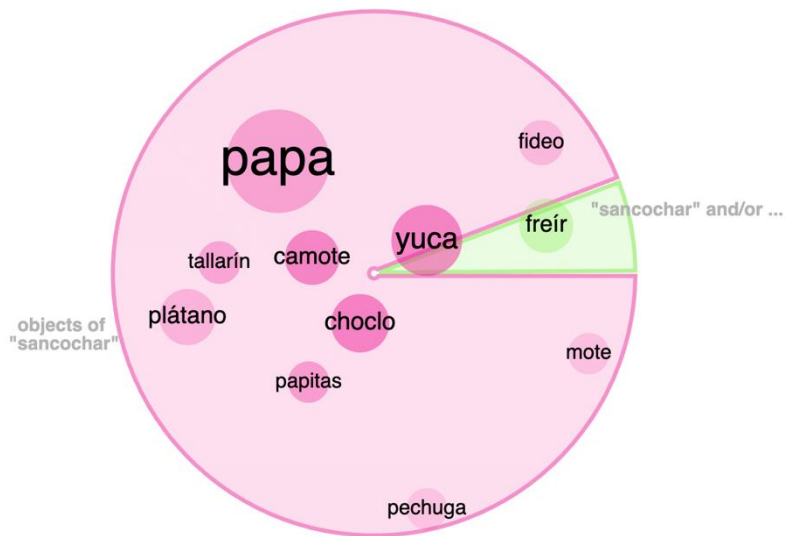
# Podemos mejorar la información de un corpus pequeño consultando Word Sketch (lema = sancochar) en esTenTen23

## sustantivos que preceden al participio *sancochado* en el subcorpus de Perú

Peruvian domain .pe

sancochar como verbo 1096x

### sancochar



Mero con camote dulce y **1choclo** **sancochado** . El Cabrito a la Nortef  
puesta por una porción de **1choclo** **sancochado** desgranado con sangr  
queso, ensalada de fruta, **1choclo** **sancochado** , etc. Asimismo, los ali

Servir con arroz, frejoles o **1yucas** **sancochadas** . L  
nos invitaban algún plátano, **1yuca** **sancochada** o i  
il, pimienta y cebollita. Su **1yuquita** **sancochada** qu  
ve con su porción de arroz, **1yucas** **sancochadas** y c

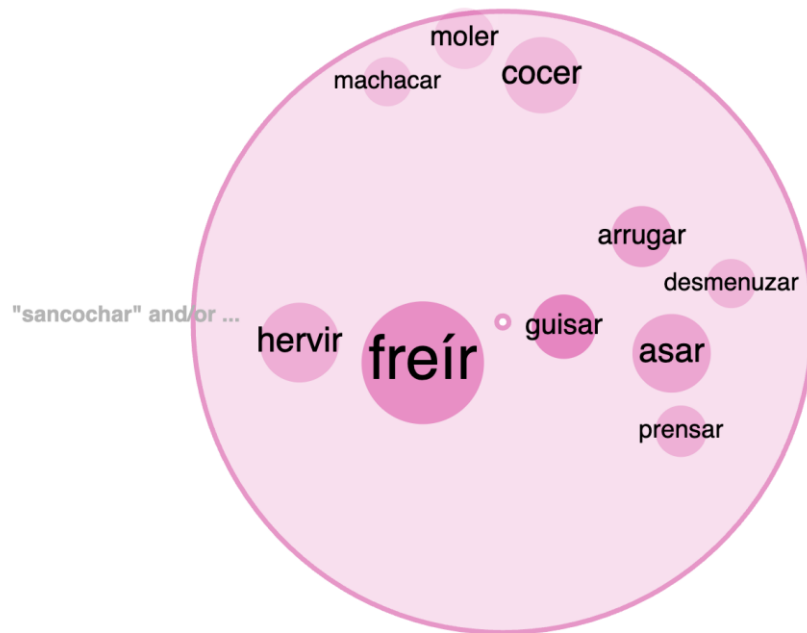
# sancochado y...

## Word Sketch obtenido de esTenTen23 (corpus completo)

### "sancochar" and/or ...

<b>guisar</b>	13	6,9	...
seca sancochada y guisada con			
<b>freír</b>	79	6,6	...
frita o sancochada			
<b>arrugar</b>	11	5,7	...
<b>asar</b>	24	5,6	...
asadas o sancochadas			
<b>hervir</b>	25	5,2	...
sancochado o hervido			
<b>prensar</b>	6	5,2	...
<b>desmenuzar</b>	5	4,9	...
<b>cocer</b>	22	4,7	...
cocidas o sancochadas			

### sancochar



# Tesauro (diccionario de sinónimos) en Sketch Engine (esTenTen23)

La **detección automática de sinónimos** se basa en la teoría de la **semántica distribucional**, que afirma, en esencia, que las palabras que aparecen en los mismos contextos tienden a tener significados similares.

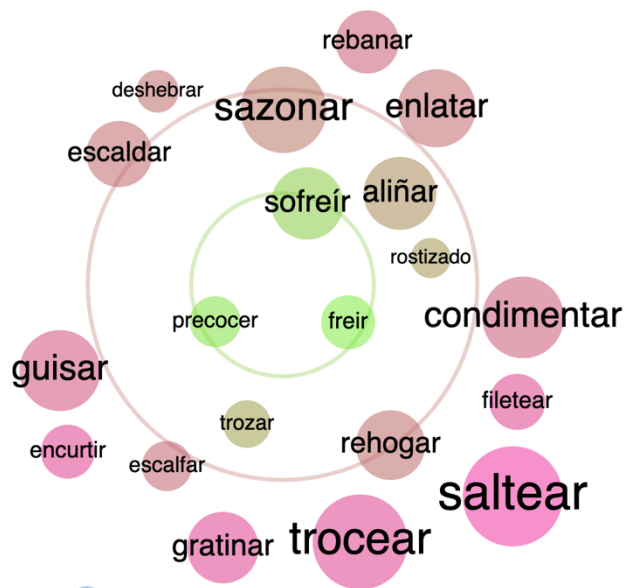
En sentido estricto, el Tesauro no presenta sinónimos, sino palabras que aparecen muy frecuentemente en los mismos contextos que el lema buscado. Muchas de ellas son sinónimos

# Tesauro (diccionario de sinónimos) en Sketch Engine (esTenTen23)

gráfico de burbujas  
sancochar

## Representación gráfica del Tesauro de *sancochar*

# nube de palabras



# Causa limeña

II. (Del quech. *kawsay*, sustento de la vida).

1. f. *Pe.* Puré de **papas** con **ají** amarillo y limón, acompañado de lechugas y aceitunas, que se come frío como entrada. pop.



Receta de Causa lim...

# frase = causa limeña de (esTenTen23)

buscar en esTenTen23 la frase **causa limeña de** y luego frecuencia por primera forma a la derecha

 **FRECUENCIA**

BÁSICO

**AVANZADO**

LEARN 

Selecciona un atributo y su posición en la concordancia: ?

forma



contexto izquierdo

6

5

4

3

2

1

KWIC

▼

1

2

3

4

5

6

contexto derecho

Word	
1	<input type="checkbox"/> pollo
2	<input type="checkbox"/> atún
3	<input type="checkbox"/> Perú
4	<input type="checkbox"/> papa
5	<input type="checkbox"/> Pollo
6	<input type="checkbox"/> Atún
7	<input type="checkbox"/> pulpo
8	<input type="checkbox"/> langostinos
9	<input type="checkbox"/> mejillones
10	<input type="checkbox"/> centolla
11	<input type="checkbox"/> las
12	<input type="checkbox"/> sardinillas

# Tesauro y Word Sketch Difference

## Otros recursos de Sketch Engine

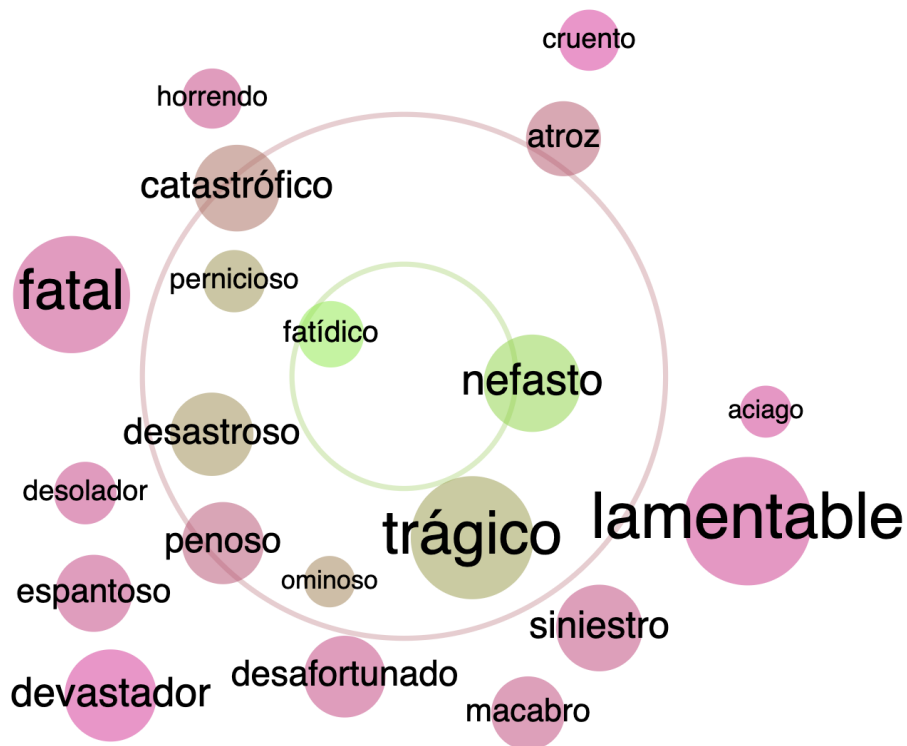
*La sinonimia no siempre es perfecta*

# Tesouro

- Hallar los sinónimos de funesto
- Hallar los sinónimos de vívido
- Hallar las diferencias entre vívido y nítido
- Hallar las diferencias entre monje y fraile
- horrible / fatal

# Tesaurus: sinónimos de *funesto* (esTenTen23)

funesto



## Word Sketch difference (comparar dos palabras)

**Word Sketch Difference** es una ampliación de la función **word sketch**.

Genera **word sketches para dos palabras** y los compara, lo que facilita observar las diferencias en su uso.

Esta función resulta especialmente útil para

**sinónimos cercanos,**

**antónimos y**

**palabras pertenecientes al mismo campo semántico.**

# Tesouro: diferencias entre *vívido* y *nítido*

		subjects of "ser vívido/nítido"		
vívido	alucinación	9	0	...
	descripción	35	0	...
	pesadilla	7	0	...
	relato	29	0	...
	imaginación	12	0	...
	sueño	122	14	...
nítido	recuerdo	72	70	...
	impresión	11	66	...
	imagen	115	1266	...
	foto	18	266	...
	pantalla	8	173	...
	sonido	14	410	...

# Adjetivos de *monje* no compartidos por *fraile*

modificadores del  
lema *monje* que nunca  
aparecen con el lema  
*fraile*

modifiers of "monje/fraile"			
clarisa	1737	0	...
tibetano	2196	0	...
zen	1124	0	...
agustinas	546	0	...
taoísta	401	0	...
cluniacense	268	0	...
ursulina	228	0	...
ortodoxo	534	0	...
shaolin	214	0	...
hindú	213	0	...
bizantino	208	0	...
lesbiano	145	0	...

**Más allá de búsquedas de palabras concretas:**

**Comodines y operadores lógicos en Sketch Engine**  
**Filtrar resultados de una concordancia**

# ¿Qué resultados obtendré si busco...? El punto (.)

*El punto (.) es el comodín más básico.*

*¿Eres capaz de predecir qué formas recuperará cada una de estas búsquedas en Sketch Engine?*

**beren.ena**

**. . . ú**

**.iménez**

→ *Anota tu respuesta antes de comprobarla en el corpus*

# El punto (.): cualquier carácter, exactamente uno

El punto ocupa exactamente una posición: sustituye cualquier letra, tilde, dígito o signo. No puede ser «nada»: siempre representa un carácter.

Búsqueda	Devuelve (entre otras formas)	Lógica
beren.ena	berenjena, berengena, berenXena...	<i>Un carácter entre «beren» y «ena»</i>
...ú	menú, Perú, bacú...	<i>Cualquier forma de 4 letras acabada en ú</i>
.iménez	Giménez, Jiménez, Ximénez, Liménez...	<i>Un carácter inicial distinto antes de «iménez»</i>

# ¿Qué resultados obtendré si busco...? Conjuntos y alternativas

*Estos operadores permiten especificar qué caracteres son posibles en una posición. ¿Producen el mismo resultado todas las búsquedas? ¿Cuál de las 4 búsquedas puede ofrecer resultados distintos de guión, guion?*

`gui[oó]n`

`gui(o|ó)n`

`guion|guión`

`gui[^aeiu]n`

# Conjuntos, alternativas y exclusiones

[ ]

## Conjunto de caracteres

gui[oó]n: uno de los caracteres dentro del corchete ocupa esa posición. Recupera guion y guión.

( | )

## Alternativa entre grupos

gui(o|ó)n: la barra separa alternativas que pueden ser de más de un carácter. Mismo resultado que [ ] en este caso.

|

## Alternativa de cadenas completas

guion|guión: la barra opera sobre toda la expresión a cada lado. Útil cuando las variantes difieren en más de una posición  
[rescibir|reçiuir|recebir|rescevir]

[ ^ ]

## Exclusión de caracteres

gui[^aeiu]n: recupera cualquier carácter menos los que siguen a ^: por tanto *guion* y *guión*. Podría recuperar también *guidn*, *guítn*, etc. poco probables pero posibles si hay erratas en el corpus.

# Qué resultados obtendré si utilizo el signo ?

*El signo ? siempre modifica al elemento inmediatamente anterior. ¿Qué implica eso en cada caso?*

cual (es) ?quier

magrebíe?s

ex-?marido

→ *Anota tu respuesta antes de comprobarla en el corpus*

# El signo ?: el elemento anterior es opcional: lo inmediatamente anterior aparece cero o una vez

El ? convierte en opcional el elemento que lo precede: puede aparecer una sola vez o no aparecer en absoluto. Nunca más de una vez.

Búsqueda	Elemento opcional	Devuelve
<code>cual(es)?quier</code>	(es) → el grupo puede estar o no	cualquier · cualesquier
<code>magrebíe?s</code>	e → la vocal puede estar o no	magrebís · magrebíes
<code>ex-?marido</code>	- → el guion puede estar o no	exmarido · ex-marido

# Cuál es la diferencia entre .\* y .+

*Ambos capturan secuencias de longitud variable, pero con una diferencia crucial. Compara los resultados de cada columna*

*Con .\**

**trans.\***

**.\*trans.\***

*Con .+*

**trans.+**

**.+trans.+**

*¿En qué caso recuperamos «trans» sola? ¿Y cuándo aparece trans exclusivamente en posición interior, es decir, no aparece en posición inicial ni final?*

# La diferencia entre .\* y .+

**.\*** = cero o más caracteres

*(puede recuperar la secuencia buscada como palabra)*

**.+** = uno o más caracteres

*(siempre hay algo más: nunca recupera la secuencia buscada como palabra aislada)*

Búsqueda	Incluye	Excluye
trans.*	trans, transporte, translúcido...	—
trans.+	transporte, translúcido...	<i>trans</i> como palabra aislada
.*trans.*	trans, transporte, retransmitir, Megatrans	—
.+trans.+	retransmitir, intrascendente...	<i>trans</i> aislado, palabras que empiecen o terminen en <i>trans</i>

# Desafío: una sola búsqueda para todas las variantes

*Esta palabra tiene cuatro grafías distintas, todas documentadas en corpus de español.*

vermú

vermut

vermout

vermouth

**¿Cómo capturarías las cuatro variantes en una sola expresión de búsqueda?**

# Solución para las variantes de *vermú*

**verm (ú | ut | out | outh)**

Descomposición de la expresión:

**verm**

Secuencia inicial fija, compartida  
por todas las variantes

**(ú | ut | out | outh)**

Terminaciones alternativas:  
una de las cuatro opciones

otra opción: `vermo?[uú]t?h?`

# Filtrar una concordancia por su contexto

*¿Cómo buscar vos solo en los casos en que aparece con formas del paradigma voseante?*

Palabra buscada:

**vos**



Términos en el contexto:

**te**

**/**

**tuyo**

---

Antes de buscar: seleccionar «Filtrar contexto de lema» en las opciones de concordancia

**todo**

**te y tuyo**

Solo ejemplos en que aparecen ambas formas a la vez en el contexto

**cualquier**

**te o tuyo**

Ejemplos en que aparece al menos una de las dos formas

**ninguno**

**NI te NI tuyo**

Ejemplos de vos en los que no aparece ninguna de las dos formas

# Cómo activar el filtro en Sketch Engine

*Pasos: buscar vos → Concordance → Filtrar contexto de lema → introducir te y tuyo*

BÁSICO **AVANZADO** AI SEARCH LEARN

Tipo de consulta ②

**simple**

lema

frase

forma

caracter

CQL

Simple

**VOS**

Subcorpus ②

ningún (todo el corpus...)

Macro ②

ninguno

Filtrar contexto ② ^

☐ No filtrar

☒ Contexto de lema

☐ Contexto de categorías gramaticales

Solo mantener líneas con

todo de te tuyo dentro 5 Tokens lizquierda y der...

→ Menú de concordancia > Filtrar contexto de lema

# Resultados: vos con te y tuyo en el contexto

*El filtro reduce la concordancia a los ejemplos lingüísticamente relevantes para el paradigma voseante*

isador.</s><s>Bueno, ¿qué <sup>1</sup> te voy a decir a	vos	?</s><s>También tuviste lo <sup>2</sup> tuyo .</s><s>Alerta en l
nen.</s><s>Pero después lo que <sup>1</sup> te pasa a	vos	, es <sup>2</sup> tuyo .</s><s>Yo tenía muchas cosas que sanar
os proyectos.</s><s> <sup>1</sup> Te falta confianza en	vos	mismo, pero dentro <sup>2</sup> tuyo hay una personalidad capa
n?</s><s>Se muere un familiar <sup>2</sup> tuyo y solo	vos	sabés qué <sup>1</sup> te puede pasar con esa información", exp
sto no es <sup>2</sup> tuyo .</s><s>Es nuestro.</s><s>	Vos	<sup>1</sup> te bajaste", señaló un internauta.</s><s>Nacy Dupli
arcados que fue un pedido <sup>2</sup> tuyo .</s><s>A	vos	solito no <sup>1</sup> te da la nafta... ni en la economía, ni en la p
uento cómo funciona para que no <sup>1</sup> te pase a	vos	ni a los <sup>2</sup> tuyos .</s><s>¿Qué es y cómo funciona el e
ren hacer del otro lado, vos hacés el <sup>2</sup> tuyo ,	vos	hacé el tuyo y <sup>1</sup> te voy a decir algo, yo pasé por mucha

⚠ El filtro opera sobre una ventana de contexto modificable ( $\pm 5$  palabras por defecto): te o tuyo deben aparecer cerca de vos