

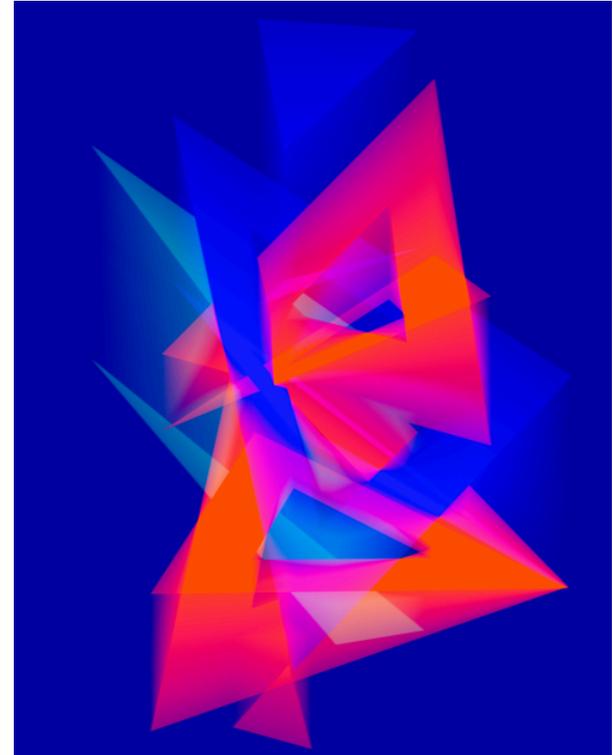
Culturomics, analítica cultural y cultumetría: Aplicaciones básicas con NGram & Wikipedia

Cursos de doctorado:

Herramientas computacionales para el análisis y visualización de datos en Ciencias Sociales y Humanidades

Influ Science

Daniel Torres-Salinas
Universidad de Granada



Escuela Internacional de Posgrado
Actividades Formativas 2025

Índice

| | |
|---|-----------|
| Aproximaciones teóricas al estudio cuantitativo de la cultura | 2 |
| 1. Culturomics | 3 |
| 2. Cultural analytics | 11 |
| 3. Cultumetría crítica | 13 |
| Herramientas para una medición de la cultura: n-gram y wikipedia | 15 |
| 1. Inside Books Ngram Viewer | 16 |
| 2. Herramientas Wikipedia | 34 |
| 3. Ejercicios prácticos | 41 |
| Bibliografía básica | 43 |

Resumen

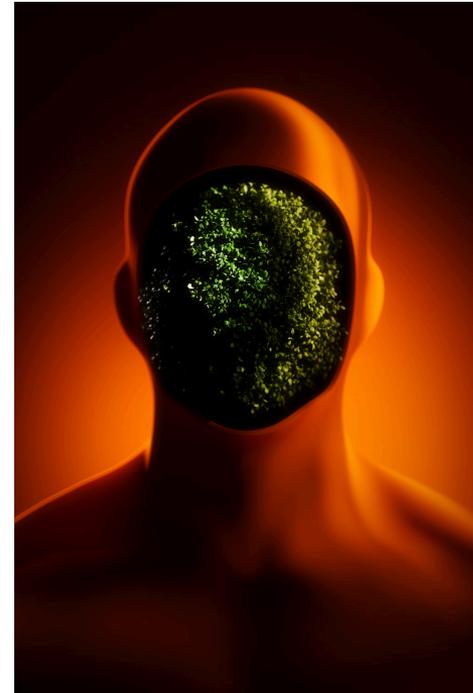
El curso aborda dos herramientas digitales abiertas y gratuitas fundamentales en el ámbito de la investigación digital: Google Books Ngram Viewer y los paquetes de análisis de estadísticas de Wikipedia. Mediante estos recursos, se propone introducir y enseñar a utilizar dichas plataformas para rastrear la evolución del lenguaje, así como la evolución e impacto de diversas tendencias culturales, la popularidad y atención que reciben determinados temas académicos o populares. Se busca responder a múltiples preguntas de investigación mediante el análisis cuantitativo de grandes volúmenes de datos textuales, adoptando marcos teóricos como la culturomía, la analítica cultural y la cultumetría. En resumen, este curso ofrece una inmersión en el uso de herramientas digitales para el análisis textual y cultural, destacando su relevancia en la investigación contemporánea y en la comprensión de las dinámicas culturales a lo largo del tiempo.

Profesor: Daniel Torres Salinas / **Organiza:** Escuela Internacional de Posgrado - UGR, Actividades formativas

Duración: 4 Horas - Contenidos teóricos y prácticos / **doi:** 10.5281/zenodo.15172621

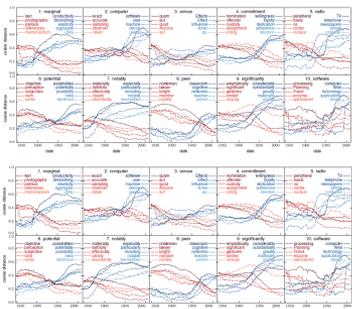
Versión: 2. Publicado el 08/04/2025 / **Licencia:** [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)

Aproximaciones teóricas al estudio cuantitativo de la cultura



1. Culturomics

Cinco características muy básicas de la culturomics



1

Puede definirse como una disciplina que **utiliza métodos cuantitativos** para analizar **vastas cantidades de datos culturales digitales**, con el fin de explorar **patrones, tendencias y evoluciones** en el ámbito cultural a lo largo del tiempo

2

Se apoya en la premisa de que **los libros y otros materiales digitales son un reflejo de los cambios culturales y lingüísticos** de la sociedad. Al analizar millones de documentos digitalizados, la culturomía busca **identificar tendencias que son reflejo de los cambios en la sociedad, la política, la economía**

3

Amplía las fronteras de la investigación cultural e introduce nuevas metodologías para el estudio de la evolución cultural, permitiendo un **análisis más sistemático y empírico** que era anteriormente difícil o **imposible de realizar con métodos convencionales**.

4

Interdisciplinariedad: existe **una gran intersección de la culturomía con otras disciplinas científicas**, como la sociología, la historia y la ciencia política, para estudios culturales más integrados. Asimismo mantiene **estrecha relación la Ciencia de Datos**

5

Existe un campo que se ha denominado como la **Culturomics** a partir de este paper publicado en la revista Science "[Quantitative Analysis of Culture Using Millions of Digitized Books](#)" (Michel et al., 2011)

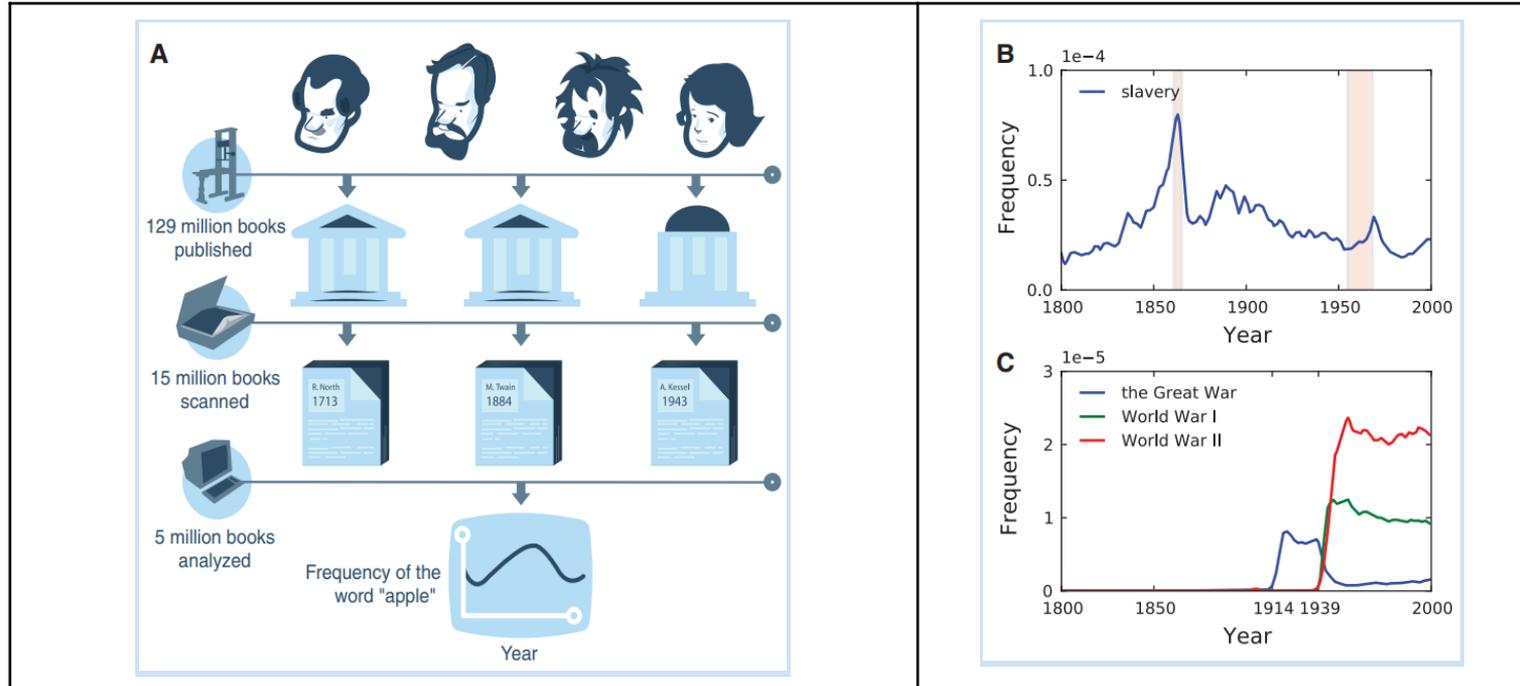
Elementos metodológicos para el análisis cuantitativo de la cultura: Culture Analytics y Culturomics

| | | | | |
|---|--|--|--|--|
| <p>^^Corpus Digitalizado Masivo: grandes bases de datos de textos digitalizados, como libros, periódicos, revistas y otros materiales escritos, que sirven como fuente</p> | <p>**Herramientas de Procesamiento de Lenguaje Natural (PLN): Empleo de tecnologías y algoritmos de PLN para analizar, entender y extraer información de los textos en el corpus.</p> | <p>**Análisis de N-Gramas: Estudio de secuencias de palabras (n-gramas) para identificar y seguir la frecuencia de aparición de términos específicos a lo largo del tiempo.</p> | <p>^^Técnicas Estadísticas y Matemáticas: Aplicación de métodos estadísticos para analizar y interpretar los patrones encontrados en los datos, permitiendo identificar tendencias y correlaciones.</p> | <p>^^Visualización de Datos: Creación de representaciones gráficas de los resultados del análisis, como gráficos de tendencias temporales, mapas de calor y redes semánticas, facilitando la interpretación y comprensión de los hallazgos.</p> |
| <p>^^Estudios Longitudinales: Realización de análisis a lo largo de extensos periodos temporales, permitiendo observar la evolución de la cultura y del lenguaje a través de décadas o incluso siglos.</p> | <p>^^Acceso abierto y Reproducibilidad: Promoción del acceso abierto a los datos y herramientas utilizados, así como la transparencia en los métodos de análisis, para facilitar la verificación y replicación de los estudios.</p> | <p>^^Ética y privacidad: Consideración de aspectos éticos relacionados con el uso de datos digitalizados y el respeto a la privacidad y derechos de autor de los materiales analizados.</p> | <p>^^Interpretación Crítica: se enfatiza la importancia de una interpretación crítica y contextual de los resultados, teniendo en cuenta los factores socioculturales, históricos y políticos relevantes.</p> | <p>^^Preocupación por la representatividad y el sesgo: Existe una preocupación compartida por desarrollar metodologías que permitan interpretaciones justas y equitativas de los datos culturales.</p> |

**** Técnicas específicas de las culturomics que se centra en el lenguaje >> N-Gram**

^^El resto son características compartidas con la Culture Analytics >> Wikipedia

Enfoque metodológico de la propuesta original de la culturomics de Michel et al. 2011



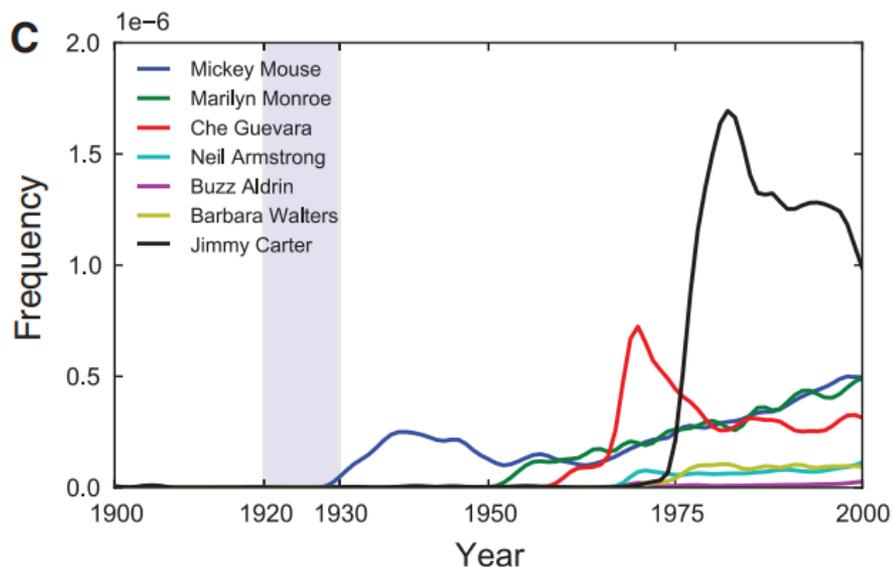
Los análisis culturónomicos estudian millones de libros al mismo tiempo. (A) Fila superior: aproximadamente 129 millones de ediciones de libros han sido publicadas desde el advenimiento de la imprenta. Segunda fila: Bibliotecas y editoriales proveen libros a Google para su digitalización (medio izquierda). Más de 15 millones de libros han sido digitalizados. Tercera fila: Cada libro se asocia con metadatos. Cinco millones de libros son seleccionados para análisis computacional. . (B) Frecuencia de uso de “esclavitud”. La Guerra Civil (1861–1865) y el movimiento por los derechos civiles (1955–1968) se destacan en rojo. El número en la parte superior izquierda ($1e-4 = 10^{-4}$) es la unidad de frecuencia. (C) Frecuencia de uso a lo largo del tiempo para “la Gran Guerra” (azul), “Primera Guerra Mundial” (verde) y “Segunda Guerra Mundial” (rojo).

Las dos herramientas fundamentales

| | |
|---|---|
|  |  |
| <p>Google Books constituye una infraestructura fundamental para la culturomía, albergando una amplia colección de textos digitalizados que abarcan una multiplicidad de disciplinas, épocas y lenguas.</p> <p>Este proyecto, iniciado por Google, ha facilitado la transición de obras literarias, documentos históricos y tratados científicos al dominio digital, democratizando el acceso al conocimiento y ofreciendo una plataforma inigualable para la investigación académica.</p> <p>La contribución de Google Books a la culturomía radica en su capacidad para suministrar un corpus textual extenso, necesario para el análisis cuantitativo de patrones a gran escala.</p> | <p>a herramienta analítica de gran valor permitiendo el rastreo y la visualización de la frecuencia de n-gramas (secuencias de palabras) dentro del extenso corpus de Google Books.</p> <p>Esta herramienta posibilita investigaciones sobre la evolución del lenguaje y las fluctuaciones en la prevalencia de conceptos culturales, ofreciendo una perspectiva cuantitativa sobre las transformaciones en el discurso y las prácticas discursivas a lo largo del tiempo. Varios corpus. Varios idiomas</p> <p>La simplicidad de su interfaz, combinada con la profundidad de su base de datos, convierte a NGram Viewer en un instrumento esencial para el análisis empírico, además es gratuito</p> |
| <p>Google Books Now - Google Books Classics - El proyecto</p> | <p>Ngram Viewer</p> |

★ Ojo con los datos es antiguo **pero fijaros en 2011** tenían **5,195,769** libros escaneados con las siguientes palabras por idioma”: 500 mil millones de palabras.

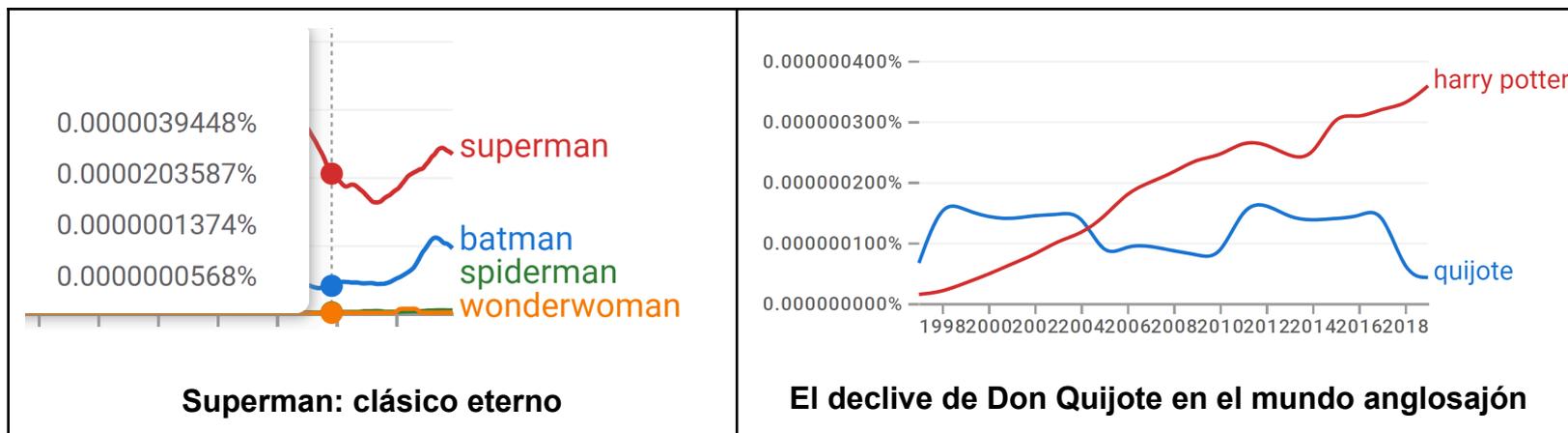
- inglés (361 mil millones)
- francés (45 mil millones)
- español (45 mil millones)
- alemán (37 mil millones)
- chino (13 mil millones)
- ruso (35 mil millones)
- hebreo (2 mil millones).



A lo largo del trabajo podemos ver algunos ejemplos como por ejemplo este de la popularidad de algunos personajes históricos de EEUU

[Veamos otros ejemplos directos de Michel et al.](#)

A continuación vemos otros ejemplos de todo tipo que he recopilado y he inventado 📍

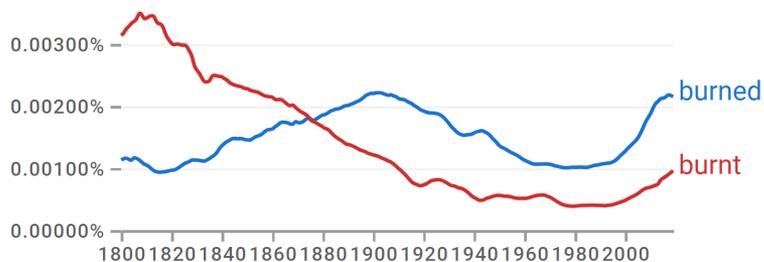


- 🦸 [Cultura popular: superhéroes famosos](#)
- 🧙 **El quijote Vs Harry Potter**
 - Aquí en el [resultado en el Corpus Inglés](#)
 - Aquí el [resultado en el Corpus Español](#)
- 🥤 [En la alimentación: lucha entre dos refrescos](#)



- ★ **En la censura y política:** [Marc Chagall](#) o [Herman Hesse](#) [Primo de Rivera en España](#)
- ★ **Tendencias académicas:** [¿Cuáles son las tendencias actuales en traducción?](#)
- ★ **Movimientos:** [tenemos un tipo que analiza el marxismo](#)
- ★ **Medios:** [Otro que analiza cuatro medios de comunicación](#)
- ★ **Ciudades:** [para ver las ciudades más literarias](#)
- ★ **Términos:** para ver el uso de un [término en un contexto político social](#)

- ★ Según los autores “*Culturomics has profound consequences for the study of language, lexicography, and grammar*” (Véase figura 2 del artículo)



Uso de Burnt or Burned

- Más ejemplos lingüísticos
 - [Backwards compatibility o backward compatibility](#)
 - [video game o computer game](#)

 - Otros experimentos y ejercicios similares
 - [GoogleWriting Tools: Ngram Viewer and Define](#)
 - La alternativa francesa 🇫🇷 [Gallicagram](#)
 - [Think Outside the box](#) 📦
- ★ [Aquí podemos encontrar más ejemplos](#) de todo lo que hemos visto
- ★ Aquí muy importante, errores y limitaciones: [Pitfalls](#)

2. Cultural analytics



Lev Manovich, el padre
Manovich.net

La Cultural Analytics, según Lev Manovich, es un enfoque interdisciplinario que utiliza métodos computacionales, matemáticos y de visualización de datos para analizar grandes volúmenes de información cultural. Su objetivo es identificar patrones, tendencias y cambios en artefactos culturales como imágenes, textos, vídeos y comportamientos digitales. Este método permite abordar la cultura contemporánea en su escala masiva y diversidad, superando las limitaciones de las metodologías tradicionales en humanidades

4 Rasgos de la cultural analytics

Uso de Big Data cultural: Analiza grandes conjuntos de datos culturales, como imágenes, textos, vídeos y comportamientos digitales, para identificar patrones y tendencias a gran escala.

Metodologías computacionales: Emplea herramientas avanzadas como algoritmos, visualización de datos y análisis estadístico para explorar fenómenos culturales.

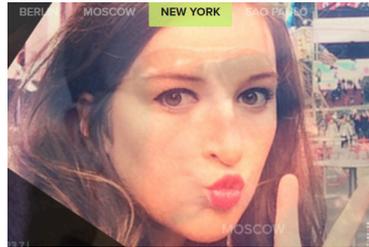
Escala masiva: Permite estudiar la cultura contemporánea en su complejidad y diversidad global, superando las limitaciones de los enfoques tradicionales centrados en casos individuales.

Interdisciplinariedad: Combina perspectivas de humanidades, ciencias sociales y ciencias computacionales para enriquecer el análisis cultural.

Proyectos singulares

...
Bajando la pelota

Selfiecity



Analiza **miles de selfies publicados en Instagram** desde cinco ciudades (Bangkok, Berlín, Moscú, Nueva York y São Paulo). Utiliza métodos computacionales para estudiar **demografía, poses y expresiones faciales, revelando patrones culturales como diferencias** en la inclinación de la cabeza y frecuencia de sonrisas entre ciudades

Ver: <https://selfiecity.net/selfiexploratory/>

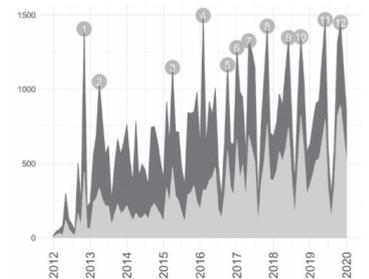
Big Bang Data



Una [exposición que explora el fenómeno de los datos masivos](#) desde perspectivas culturales, políticas y sociales. Este proyecto utiliza **visualizaciones y análisis de datos para reflexionar sobre el impacto del big data en la cultura contemporánea**

Ver vídeo 2: [Big Bang Data Datificación](#)

Ver vídeo 1: [La producción social de datos](#)



Desde una óptica del análisis de redes sociales, este artículo explora la evolución temporal de la actividad y la conformación de la red de actores que debaten en Twitter sobre la #fpdual durante el periodo comprendido entre los años 2011 y 2019. Los resultados alcanzados han permitido trazar los periodos relativos a la fase inicial, la rápida expansión y la consolidación de la actividad en el debate

[Descargar](#)

3. Cultumetría crítica

La cultumetría crítica surge en el marco del proyecto LETRAL IA (dirigido por Ana Gallego Cuiñas y Daniel Torres-Salinas) como respuesta a los nuevos desafíos del análisis literario en la era del big data y la información digital. Este concepto integra influencias metodológicas de diversas corrientes.

Informetría

Entendida como el estudio cuantitativo de la producción documental, informetría se adapta a las características de la producción literaria

Analítica Cultural

Desarrollada por Lev Manovich, que aplica métodos computacionales y visualización de datos al estudio de fenómenos

Culturomics

Análisis masivo de textos digitalizados para identificar tendencias culturales en grandes corpus.

A pesar de sus logros metodológicos carecen de una dimensión crítica profunda: suelen describir el qué (qué patrones o tendencias existen en los datos), pero no siempre exploran el por qué ni las implicaciones ideológicas de esos hallazgos



Es aquí donde se hace necesaria la perspectiva de la Teoría Crítica de la Escuela de Fráncfort, con su énfasis en estructuras de poder, alienación y crítica cultural. Pensadores como Theodor W. Adorno, Max Horkheimer o Walter Benjamin aportan marcos teóricos para analizar la cultura más allá de la apariencia cuantitativa, cuestionando la aparente neutralidad de los datos y de los objetos estéticos. **De hecho, la teoría crítica se centra en la evaluación reflexiva de la sociedad y la cultura con el fin de revelar y desafiar las estructuras de poder, sosteniendo que la ideología es uno de los principales obstáculos para la liberación humana.** **Conceptos básico:** Razón Instrumental (Horkheimer y Adorno), Industrias culturales (Adorno), Distant Reading (Moretti)

Aportación seminal

Daniel Torres Salinas & Ana Gallego Cuiñas

Big Borges: What Can Big Data Show About a Classic Writer on Social Networks?

[Enlace al artículo](#)

| Borges quotations mentioned on Twitter | Indicators of dissemination (interactions and reactions) | | | | |
|--|--|-------------|---------------|------------------------------|-----------------|
| | Nº Tweets | Nº Retweets | Nº Favourites | Nº Replies + r. with comment | Total Attention |
| 1. The pursuit of serenity seems to me to be a more reasonable ambition than the pursuit of happiness. | 1091 | 26991 | 89066 | 2113 | 118170 |

Herramientas para una medición de la cultura: n-gram y wikipedia



1. Inside Books Ngram Viewer

Antes de empezar con un corpus masivo recordemos Los Corpus tradicionales ¿cómo funcionan, cuáles son sus elementos y para qué sirven?, para ello tomaremos algunos materiales de la presentación: [Uso de corpus y diccionarios combinatorios en la traducción](#) de M. Orozco-Jutorán

| Corpus | Autor | Volumen | Fuentes | Periodo | Distribución | Características |
|--|--|----------------------------------|--|-------------------------------------|---------------------------|---|
| Corpus del Español (CdE) ➤ Abrir enlace | Mark Davies, Brigham Young University (EEUU) | Ca. 7,2 mil millones de palabras | Prensa digital | 2012-2019 (ampliaciones periódicas) | 78% América 22% España | Búsqueda y comparación por países y en el tiempo; el corpus más actual y amplio |
| Corpus del Español del Siglo XXI (CORPES) ➤ Abrir enlace | RAE | Ca. 286 millones de formas | Textos escritos y orales: ficción, no ficción, prensa, Internet | 2001-2016 (ampliaciones periódicas) | 70% América 30% España | |
| Corpus de Referencia del Español Actual (CREA) ➤ Abrir enlace | RAE | Ca. 160 millones de formas | Textos escritos y orales: ficción, no ficción, prensa y medios de comunicación | 1975-2000 | 50% América 50% España | Amplia variedad de géneros y temáticas; búsqueda por países, temáticas, medios y autores. |
| Corpus Diacrónico del Español (CORDE) ➤ Abrir enlace | RAE | Ca. 200 millones de formas | Textos escritos (prosa y verso): ficción, no ficción, prensa | Hasta 1974 | 26% América 74% España | |

¿Cuándo es útil consultar un corpus monolingüe en español?

1. Dudas sobre frecuencia de uso en general o en variedades regionales de un término.

Ejemplo: ¿Cómo traducir «*freelance writer*» al español?

Supongamos que, después de consultar en diccionarios bilingües, o sin haber consultado, se me ocurren tres opciones:

- escritor independiente
- escritor *freelance*
- escritor autónomo

Me pregunto:

- ✓ ¿Qué opción es más frecuente?
- ✓ ¿Qué opción es más habitual en el español peninsular?
- ✓ ¿Qué opción es más habitual en otras variedades del español?

2. Dudas sobre qué palabras se suelen combinar con otras (para conferir naturalidad a la traducción).

Ejemplo: ¿Cómo traducir «*it can be a powerful tool*» al español?

Lo primero que me viene a la cabeza es traducirlo por «puede ser una herramienta poderosa», pero «herramienta» me parece muy poco adecuado en el contexto, que en este ejemplo es educativo, y me decido por «recurso». Me pregunto:

- ✓ ¿«Recurso poderoso» es natural en español o es un anglicismo?
- ✓ ¿Cuáles son los adjetivos (con ese sentido) que suelen combinarse con «recurso»?

[Fuente Mariana Orozco](#)

Ya hemos visto que **NGram** es una aplicación lanzada por Google en 2010, en este caso nos encontramos ante un **corpus masivo de carácter multilingüe**, en el siguiente cuadro vemos las **diferencias esenciales entre un corpus tradicional y un corpus masivo**.

| | | |
|---------------------------|--|---|
| Corpus Tradicional | Corpus de tamaño moderado, utilizado comúnmente en estudios de lingüística aplicada y descriptiva. | <ul style="list-style-type: none"> - Tamaño limitado. - Diseñado para estudios específicos. - Puede ser monolingüe o multilingüe. |
| Corpus Masivo | Colecciones de textos de gran tamaño que abarcan una amplia gama de géneros y registros. | <ul style="list-style-type: none"> - Gran volumen de datos. - Representación de múltiples géneros. - Útil para estudios estadísticos y aprendizaje automático. |

La lingüística cuantitativa, una subdisciplina de la lingüística que utiliza métodos matemáticos y estadísticos para analizar y modelar fenómenos lingüísticos ha ganado prominencia gracias a la posibilidad de analizar grandes cantidades de datos con precisión. Asimismo, han facilitado la emergencia de subdisciplinas como la lingüística de corpus y la sociolingüística computacional

○

*“Al examinar los libros de forma colectiva, Google puede procesar el texto y proporcionar la **repetición de la aparición de palabras basada en datos** estadísticos. Con la herramienta de búsqueda Google **Ngram Viewer**, puede buscar a través de estos **enormes datos estadísticos de manera rápida** y efectiva. Al comparar la popularidad relativa de las palabras, **puede establecer cómo cambia el idioma y la cultura a lo largo del tiempo**. Ngram puede hacer mucho más que simplemente informar la frecuencia de las palabras”* ([Fuente](#))

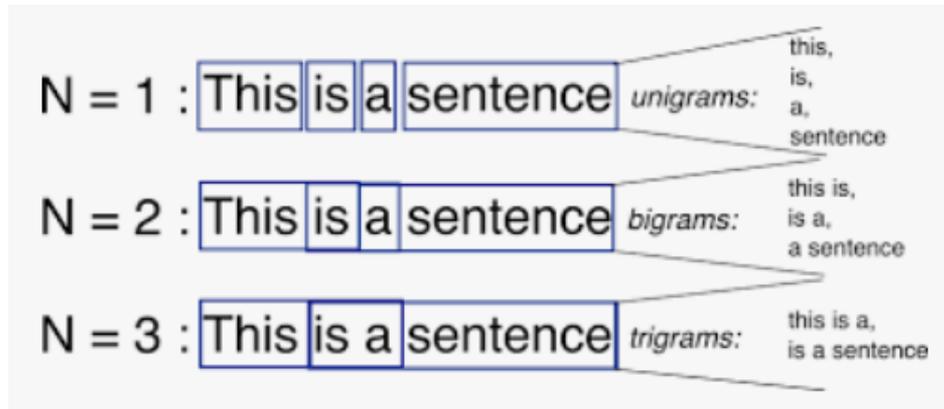
Ahora demos otro paso conceptual Pero ¿Qué es un **Ngram**?

Son elementos de una muestra de texto

Los N-Grams, siendo **secuencias de 'n' ítems** (que pueden ser **fonemas, sílabas, letras, palabras** o pares de palabras) extraídos de un texto o corpus de habla.

El N-Grama puede estar compuesto por **grandes bloques de palabras** o por una única palabra o **Token**.

Tenemos dos tipos de **tokens: words and nonwords**. Dependiendo del número de elementos se puede denominar **unigrama, bigrama o trigrama**



Unigrama (1-grama): Una sola palabra. Ejemplo: "biblioteca"

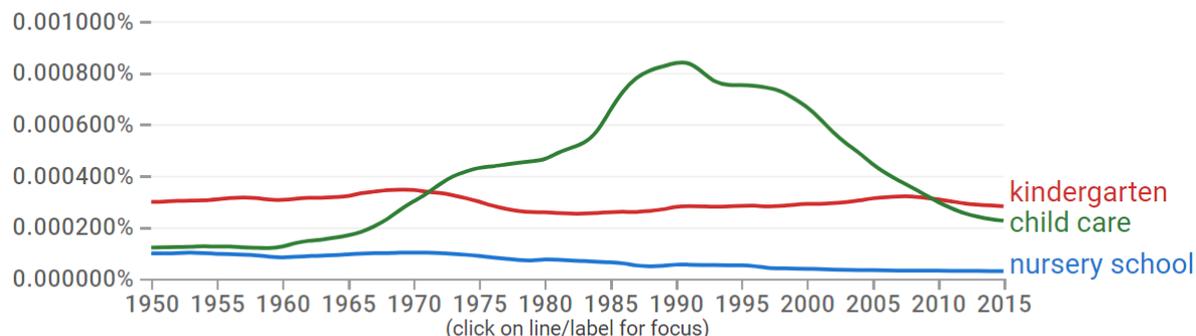
Bigrama (2-grama): Dos palabras consecutivas. Ejemplo: "biblioteca universitaria"

Trigrama (3-grama): Tres palabras consecutivas. Ejemplo: "acceso a información"

[Ejemplo básico del concepto de N-Gram y sus aplicaciones](#) (especialmente predicción de texto)

★ Comparado unigramas y bigramas.

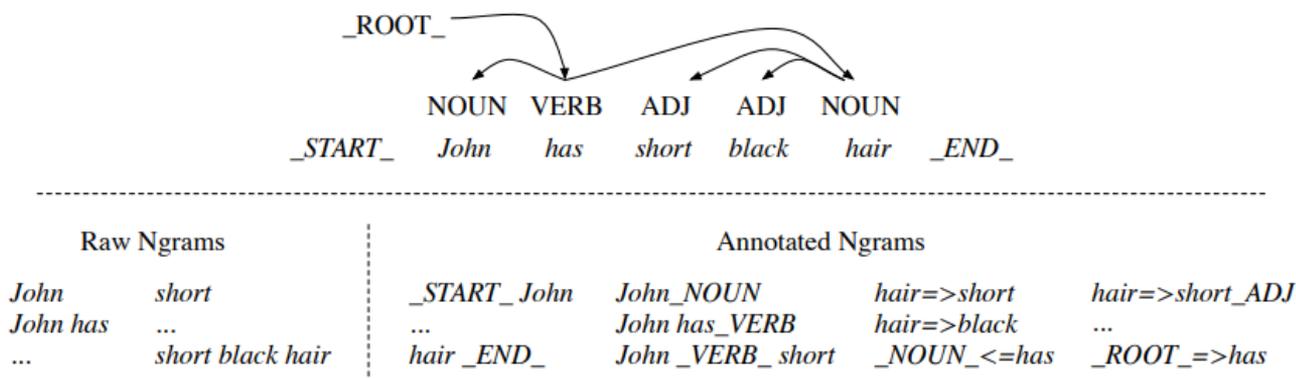
[En el siguiente Ejemplo](#) en el siguiente gráfica vemos las frecuencia de tres términos concretos (Kindergarten, Child Care y Nursery School) durante un período de tiempo (1950 y 2015). Claramente se ve el uso de los tres términos. En este caso Kindergarten es un unigram y Chil Care y Nursery School son Bigramas



NGram es una colección de Corpus de diferente naturaleza de varios idiomas y períodos

En inglés, chino (simplificado), francés, alemán, hebreo, italiano, ruso o español. Corpora especializados en inglés, tales como inglés americano, inglés británico, inglés de ficción e “inglés un millón. Todos los corpus se generaron en julio de 2009, julio de 2012 y febrero de 2020; actualizan estos corpus a medida que continúa el escaneo de libros

★ Representación de N-grams con en Google Ngram Viewer



[Fuente](#)

El diagrama contextualiza cómo herramientas como Google Ngram Viewer procesan y representan datos lingüísticos, ilustrando la relación entre palabras en una oración según su estructura gramatical. En la parte superior, un árbol de dependencias muestra cómo las palabras están interconectadas: "John" (sustantivo) es el sujeto del verbo "has", mientras que "short" y "black" califican al sustantivo "hair". Estas relaciones gramaticales son esenciales para enriquecer el análisis lingüístico, y las marcas *START* y *END* delimitan el inicio y final de la oración, facilitando el procesamiento computacional. La tabla inferior diferencia entre *n-grams* "crudos" y anotados, vinculando directamente con los datos utilizados en Ngram Viewer. Los *n-grams* crudos son secuencias consecutivas de palabras, como "John has" o "short black". Por otro lado, los anotados incorporan información gramatical (como "John_NOUN" o "hair=>short_ADJ"), permitiendo un análisis más profundo y estructurado. Esto no solo mejora la capacidad de búsqueda y visualización en herramientas como Ngram Viewer, sino que también es clave en tareas de traducción y análisis lingüístico avanzado.

★ Cobertura de Ngram Viewer

- Tiempo: El Google Ngram Viewer ofrece un análisis diacrónico del lenguaje, **abarcando desde el año 1500 hasta el 2022**, con datos en intervalos anuales. Aunque ha sido actualizado varias veces, incorporando más libros y mejorando las tecnologías de OCR, es esencial reconocer que **la representatividad de los datos puede fluctuar a lo largo del tiempo**. La base de datos, aunque extensa, **puede mostrar sesgos, especialmente en épocas más antiguas**, y **no necesariamente refleja el lenguaje oral o no publicado**. Se muestran intervalos.
- **Idiomas y Corpus**: El Google Ngram Viewer cuenta con corpus en diversas lenguas, entre las que se incluyen el inglés, chino (simplificado), francés, alemán, hebreo, italiano, ruso y español. Además, **en el caso del inglés, se dispone de corpora especializados** que reflejan variaciones geográficas y contextuales, tales como el **inglés americano, inglés británico, inglés de ficción** y el corpus denominado **"inglés un millón", este último reducido a un millón de tokens para ofrecer una representación estandarizada**. Es relevante mencionar que todos los corpus se generaron en tres fechas clave: **julio de 2009, julio de 2012 y febrero de 2020**. Con el compromiso de reflejar la continua expansión de su biblioteca digital, Google tiene previsto actualizar estos corpus. Cada nueva versión contará con identificadores persistentes distintos para garantizar su rastreabilidad y distinción respecto a las anteriores.

Google Ngram Viewer se actualizó en julio de 2024

Esta actualización incorporó un nuevo conjunto de datos que incluye palabras y frases más recientes. Además, se realizaron mejoras en el reconocimiento óptico de caracteres (OCR) y en la detección de idiomas, y se añadieron más libros al corpus. Aunque las versiones anteriores (2009, 2012, 2019) se eliminaron del menú desplegable del corpus, aún están disponibles y pueden accederse mediante operadores de búsqueda específicos, como :eng_2019 o :fre_2012. También se amplió el límite de palabras en las consultas de búsqueda, permitiendo ahora hasta siete palabras por consulta, en comparación con el límite anterior de cinco.

| | | | | | |
|------------------------------|------------------|----------------------------|--------------|---------------------|----------|
| American English 2019 | eng_us_2019 | English One Million | eng_1m_2009 | Spanish 2019 | spa_2019 |
| American English 2012 | eng_us_2012 | Chinese 2019 | chi_sim_2019 | Spanish 2012 | spa_2012 |
| American English 2009 | eng_us_2009 | Chinese 2012 | chi_sim_2012 | Spanish 2009 | spa_2009 |
| British English 2019 | eng_gb_2019 | Chinese 2009 | chi_sim_2009 | Russian 2019 | rus_2019 |
| British English 2012 | eng_gb_2012 | French 2019 | fre_2019 | Russian 2012 | rus_2012 |
| British English 2009 | eng_gb_2009 | French 2012 | fre_2012 | Russian 2009 | rus_2009 |
| English 2019 | eng_2019 | French 2009 | fre_2009 | Italian 2019 | ita_2019 |
| English 2012 | eng_2012 | German 2019 | ger_2019 | Italian 2012 | ita_2012 |
| English 2009 | eng_2009 | German 2012 | ger_2012 | | |
| English Fiction 2019 | eng_fiction_2019 | German 2009 | ger_2009 | | |
| English Fiction 2012 | eng_fiction_2012 | Hebrew 2019 | heb_2019 | | |
| English Fiction 2009 | eng_fiction_2009 | Hebrew 2012 | heb_2012 | | |
| | | Hebrew 2009 | heb_2009 | | |

★ Elementos del interfaz

- **Caja de Búsqueda:** Aquí es donde se introducen las palabras o frases (n-gramas). Se pueden introducir múltiples términos separados por comas
- **Selección de Corpora:** elegir entre diferentes corpora lingüísticos. Esto incluye distintos idiomas como inglés, español, francés, entre otros, ...
- **Intervalo Temporal:** Proporciona la capacidad de ajustar el rango de años que se desea analizar.
- **Opciones de Búsqueda:** Estas incluyen:
 - Case-insensitive: Ignora las diferencias entre mayúsculas y minúsculas.
 - Ver ejemplo de diferencias, con “La casa” y “la casa”
 - Permite seleccionar entre las diferentes versiones de la base de datos.
- **Gráfico:** Una vez realizada la búsqueda, se muestra un gráfico que representa la frecuencia relativa de los términos buscados a lo largo del tiempo. Las líneas del gráfico reflejan la popularidad de cada término en relación con el total de palabras del corpus seleccionado para cada año.
 - Eje de Frecuencias y Eje Temporal: El eje Y representa la frecuencia relativa de los términos, mientras que el eje X muestra el rango temporal.
 - Smoothing: Ajusta la suavidad de la línea en el gráfico, lo que es útil para identificar tendencias generales al promediar variaciones anuales.
- **Links a Google Books:** Al hacer clic en puntos específicos del gráfico, se proporcionan enlaces a ejemplos de los libros donde aparece el n-grama, lo que facilita la exploración contextual de los términos.

★ Opciones y técnicas de búsqueda en NGram 🔍

En esta sesión, abordaremos [opciones avanzadas y técnicas prácticas](#) para optimizar las búsquedas en NGram, tales como el uso de operadores, delimitación de rangos temporales y selección de corpus específicos. Este conocimiento no solo enriquecerá su competencia lingüística, sino que también proporcionará herramientas para fundamentar decisiones traductológicas basadas en datos empíricos.

○ **Opción de ortografía case-sensitive**



- Compara el uso exacto de las mayúsculas y minúsculas, el análisis se vuelve especialmente relevante para **nombres propios, acrónimos y términos cuya capitalización puede alterar su significado o contexto**. Por ejemplo, “Apple” puede referirse a la marca tecnológica, mientras que “apple” indica el sustantivo común “manzana”.
- Esta sensibilidad permite identificar variaciones significativas en el uso de términos y puede **proporcionar resultados totalmente diferentes dependiendo de la capitalización**, lo que es fundamental para un análisis lingüístico preciso.
- Aquí tenemos algunos ejemplos
 - Nombres propios solamente: [Fitzgerald vs Dupont](#)
 - Acrónimos vs. palabras comunes: [RAM vs ram](#)
 - Nombres propios vs. términos comunes: [Amazon vs Amazon](#)

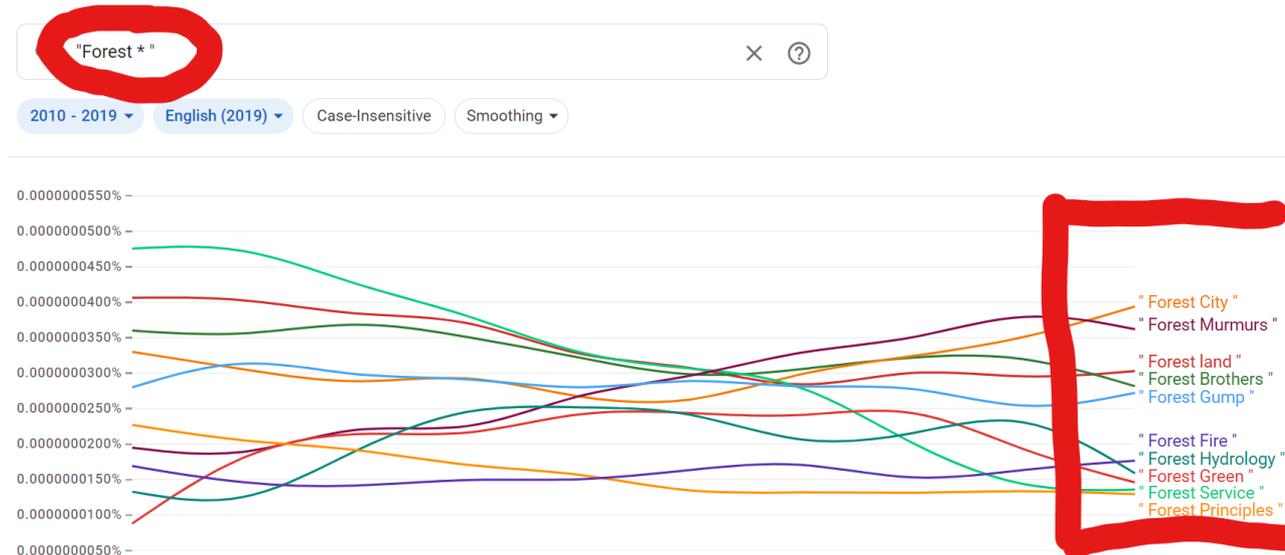
 Si nos situamos en sobre la línea de un N-Gram y pulsamos el botón derecho veremos solo ese Ngram

Wildcard

- El operador Wildcard (*) en Google NGram Viewer permite realizar búsquedas avanzadas al **sustituir una palabra específica por un comodín**, lo que resulta útil para explorar patrones o **identificar las combinaciones más comunes en un contexto determinado**. Este operador **devuelve un ranking con las 10 combinaciones más frecuentes** que aparecen en la posición donde se ha

colocado el comodín, ofreciendo un análisis enriquecido de las tendencias lingüísticas.

- Por ejemplo, al buscar **"the * cat"**, el visor podría mostrar combinaciones como "the black cat", "the white cat", o "the big cat". Esta funcionalidad es especialmente valiosa para estudiar **construcciones gramaticales**, identificar **términos relacionados en un campo semántico** o analizar la **evolución de frases hechas**



- Aquí tenemos algunos ejemplos
 - [Aquí por ejemplo con “Forest *”](#)
 - [Aquí por ejemplo con “Igualdad *”](#)
 - [Aquí por ejemplo con tecnología * innovación](#)

💡 En la zona inferior tenemos enlaces al lugar exacto donde se ha extraído la palabra, veremos la palabra con un enlace a diferentes períodos temporales que nos remiten a Google Books, una vez que vemos los resultados de un determinado N-Gram incluido podemos acceder al libro 💡 Si nos situamos en sobre la línea de un N-Gram y pulsamos el botón derecho veremos solo ese Ngram

- **Inflexión** _INF

- La inflexión es el proceso mediante el cual una palabra cambia su forma para reflejar diferentes categorías gramaticales, como aspecto, caso, género, modo, número, persona, tiempo y voz.
- En Google N-Gram Viewer, es posible analizar estas formas flexionadas agregando el sufijo **_INF** a una búsqueda. Por ejemplo, al buscar "run_INF", el visor incluye resultados como "run", "running", "ran" y otras formas derivadas.
- Esta funcionalidad es particularmente útil para estudios lingüísticos, ya que permite identificar tendencias en el uso de formas verbales, declinaciones y conjugaciones en diferentes períodos y contextos.
- Algunos ejemplos

- **En verbo:** Por ejemplo “[run_INF](#)” Permite estudiar las formas verbales asociadas al verbo "run
- **En verbo:** Por ejemplo “[be_INF](#)” permite estudiar el verbo irregular "be", que presenta múltiples formas dependiendo del tiempo y la persona.
- **En verbo:** Por ejemplo con “[Search_INF for Information](#)” me busca search y searching
- **En sustantivo:** Por ejemplo “[child_INF](#)” permite analizar cómo se utiliza la forma singular y plural.
- **En adjetivo:** Por ejemplo “[happy_INF](#)” permite investiga las formas singular y plural de un sustantivo irregular.



puedes combinar sin problemas Wildcard e Inflexiones

Operadores

En Google NGram Viewer, los operadores lingüísticos **permiten realizar búsquedas específicas** según la **categoría gramatical** de las palabras, como verbos, sustantivos, adverbios, artículos, entre otros. Además, **los operadores aritméticos ofrecen comparaciones cuantitativas entre términos**, mientras que los **operadores de dependencia permiten explorar relaciones sintácticas** entre palabras.

- **Categorías gramaticales**

- En Google NGram Viewer, los operadores de categorías gramaticales permiten **realizar búsquedas precisas basadas en la función gramatical de las palabras dentro de un corpus**. Estos operadores son especialmente útiles para lingüistas, traductores e investigadores interesados en analizar patrones y tendencias gramaticales.

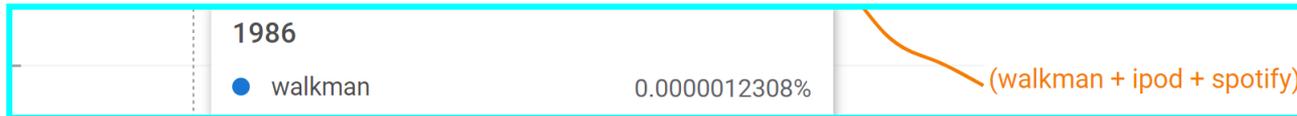
Google Books Ngram Viewer

🔍 light_NOUN,light_VERB,light_ADJ ✕ ?

- Estas categorías incluyen:
 - **Verbos (VERB)** explorar conjugaciones y tiempos verbales
 - **Sustantivos (NOUN)**, que abarcan singulares y plurales
 - **Adjetivos (ADJ)**, usados para describir sustantivos
 - **Adverbios (ADV)**, que modifican verbos o adjetivos
 - **Artículos (DET)**, como definidos e indefinidos.
- Un ejemplo:
 - [Play puede ser nombre \(play_NOUN\) y verbo \(play_VERB\)](#)
 - [Run como nombre y como verbo](#)
 - [Light como nombre, como verbo y como adjetivo](#)

○ Operadores Aritméticos

- Los operadores aritméticos **permiten combinar o comparar frecuencias de palabras o frases para análisis más detallados.**
- Por ejemplo, el operador de suma (+) combina las frecuencias de dos términos, como “dog + cat”, para observar su aparición conjunta a lo largo del tiempo.



- El operador de resta (-) permite analizar diferencias en frecuencia entre términos, destacando tendencias divergentes, como “war - peace”. Además, el operador de división / se pueden usar para calcular proporciones o correlaciones entre palabras, facilitando estudios más complejos de evolución lingüística o cultural.
- Por ejemplo:
 - [Comparativa entre walkman, ipod y spotify](#)
 - [guerra - paz](#): Aquí se analiza la frecuencia de la palabra "guerra" excluyendo las menciones de "paz". Destaca cómo ha variado el uso de "guerra" de manera independiente,
 - [literatura / ciencia](#)
 - Si la proporción literatura / ciencia es 100% en un punto del tiempo, significa que la palabra "literatura" y la palabra "ciencia" aparecen con la misma frecuencia
 - $\text{literatura} / \text{ciencia} = 0.001 / 0.001 = 1.0$ (100%).

- Si el valor supera el 100% significa que "literatura" se menciona más frecuentemente que "ciencia"
 - $\text{literatura} / \text{ciencia} = 0.002 / 0.001 = 2.0$ (200%)
- Si el valor es inferior al 100%, significa que "ciencia" se menciona más frecuentemente que "literatura".
 - $\text{literatura} / \text{ciencia} = 0.0005 / 0.001 = 0.5$ (50%).

○ Operadores Dependencia

- Los **operadores de dependencia** son herramientas avanzadas disponibles en Google Ngram Viewer para analizar la **relación sintáctica y semántica** entre palabras en el corpus. Estos operadores permiten buscar patrones más complejos que van más allá de las simples frecuencias, explorando cómo las palabras interactúan dentro de una estructura gramatical. Veamos un par de operadores.
 - **Operador =>** (dependencia directa). Este operador busca la relación directa entre una palabra y otra.
 - **Operador :** (**relación específica**). Permite buscar una relación gramatical específica entre palabras, definida por la etiqueta sintáctica.
- Algunos ejemplos:
 - [run=>quickly](#). Busca casos en los que el verbo "run" tiene una dependencia directa con el adverbio "quickly" (por ejemplo, "run quickly")
 - Por ejemplo todas las veces que aparece el término [house con charming](#)

- [Todas las veces que house aparece con charming, lovely o enchanting](#)
- [book:NOUN](#). Busca ocurrencias de la palabra "book" cuando funciona como un sustantivo (NOUN).

Google Books Ngram Viewer



Q book:NOUN X ?

- **Combinaciones avanzadas**
 - Todo se puede combinar quiero buscar [todos los nombres que aparecen dependiendo del término Pay](#), utilizamos tres de los operadores que hemos empleado. [Otros ejemplos combinación el tipo y la dependencia](#)

★ Otros experimentos y ejercicios similares

- [GoogleWriting Tools: Ngram Viewer and Define](#)
- [Gallicagram](#)
- [Una cuenta de twitter](#)

2. Herramientas Wikipedia

| | | |
|--|---|--|
| <p>Naturaleza colaborativa: Permite que cualquier usuario, desde cualquier parte del mundo, pueda editar casi cualquier artículo en cualquier momento. Esto fomenta un esfuerzo colectivo en el que el conocimiento es construido y mejorado por voluntarios.</p> <p>Datos de los Wikipedians Tipos de Usuarios en Wikipedia Ranking de usuarios ES</p> | <p>Proceso de revisión y discusión: Cada artículo en Wikipedia tiene asociada una página de "discusión" donde los editores pueden debatir mejoras, correcciones y contenido controvertido. Este proceso democrático asegura un consenso colectivo en torno al conocimiento compartido.</p> <p>Ver discusión en Durruti Ver historial en Golpe de Tejero</p> | <p>Ediciones en varios idiomas: Wikipedia está disponible en más de 300 idiomas, lo que facilita el acceso global al conocimiento y permite la inclusión de perspectivas culturales diversas.</p> <p>Número de Wikipedias y Usuarios activos en cada de ellas</p> |
| <p>Modelo de contenido libre: Wikipedia opera bajo una licencia de contenido libre, específicamente la Licencia de Documentación Libre de GNU y Creative Commons Attribution-ShareAlike, lo que permite que su contenido sea utilizado y distribuido libremente, siempre que se otorgue el debido crédito y se comparta de manera similar.</p> | <p>Sistema de referencias y citas: Los artículos en Wikipedia deben estar bien referenciados con fuentes fiables. Esto es crucial para asegurar la veracidad y fiabilidad de la información presentada.</p> <p>Ejemplos de citación de artículos científicos en Wikipedia (Torres-Salinas et al., 2019)</p> | <p>Funcionalidades complementarias: Wikipedia ofrece diversas herramientas y funcionalidades, como la posibilidad de crear libros a partir de artículos seleccionados, portales temáticos, resúmenes de noticias y acceso a bases de datos relacionadas como Wikimedia Commons</p> <p>Portales / Portal Al-Andalus Portales más famosos en ES Wikimedia</p> |

Los cinco pilares de Wikipedia («Wikipedia en español», 2024)

[Es una enciclopedia](#)

[Busca un «punto de vista neutral»](#)

[Es de contenido libre](#)

[Sigue unas reglas de etiqueta](#)

[No tiene normas fijas](#)

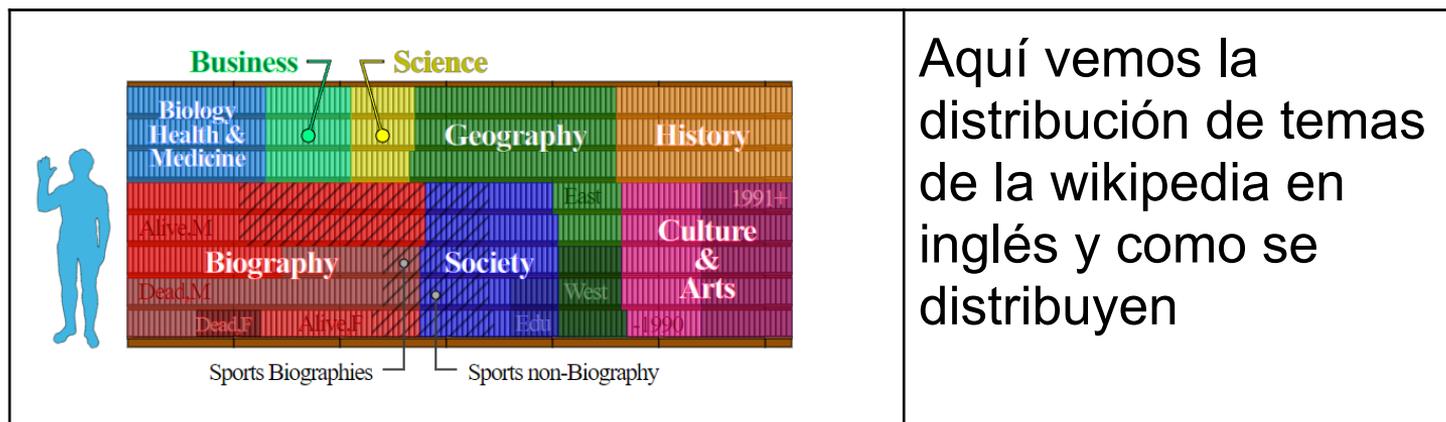
★ Algunas estadísticas globales de wikipedia

[En esta páginas podemos consultar diferentes aspectos](#)  /

- Tasa de crecimiento / Número de palabras / Número de páginas

[Wikipedia - Estadísticas](#) / [Estadística de Wikimedia](#)

★ Contenidos en Wikipedia



★ ¿Cómo podemos utilizar wikipedia en investigación?

| | | | |
|--|---|--|---|
| <p>Desarrollo del contenido a lo largo del tiempo: Observar cómo han crecido y evolucionado los artículos desde su creación. Esto puede dar luces sobre el interés en ciertos temas y cómo ha cambiado con el tiempo.</p> | <p>Popularidad de temas: Mediante el seguimiento de las vistas de página y la actividad de edición en los artículos, se pueden identificar tendencias en la popularidad de ciertos temas.</p> | <p>Cobertura geográfica y cultural: Analizar cómo la cobertura de temas varía por regiones y culturas, lo que podría indicar sesgos o deficiencias en la representación del conocimiento global.</p> | <p>Actividad de edición durante eventos: Examinar la actividad de edición durante eventos importantes (como elecciones, desastres naturales, lanzamientos de productos, etc.) podría mostrar cómo la información se expande y se corrige en tiempo real.</p> |
| <p>Polarización y consenso: Estudiar la discusión y la edición de temas controvertidos podría revelar información sobre polarización en la sociedad, así como sobre cómo se llega al consenso en la comunidad de Wikipedia.</p> | <p>Enlaces externos y referencias: El análisis de las referencias y los enlaces externos puede ofrecer una perspectiva sobre las fuentes de información más utilizadas y confiables en diferentes campos del conocimiento.</p> | <p>Demografía de los editores: Comprender quiénes son los editores de Wikipedia, su ubicación, y la diversidad de su demografía puede iluminar aspectos de inclusión y sesgo en la creación de contenido.</p> | <p>Cambios en las políticas de edición: Observar cómo las políticas de Wikipedia han cambiado y cómo esto afecta la creación y edición de contenido.</p> |

★ **Ejemplos de contenido popular**

[The top 50 Report - 2023](#)

[Wikipedia:Today's featured article/Most viewed](#)

[Buscar en Wikimedia en la preguntas artículos más vistos](#)

★ **Herramientas: [Pageviews](#)**

La página "Pageviews Analysis" de Wikimedia Toolforge proporciona herramientas analíticas para evaluar y comparar las vistas de página de los artículos de Wikipedia. Los usuarios pueden examinar datos de uso diario o mensual, filtrar por tipo de dispositivo y tipo de usuario (humano o bot), y obtener visualizaciones como gráficos de líneas o barras. Es una herramienta útil para entender la popularidad de diferentes páginas de Wikipedia y observar tendencias a lo largo del tiempo.

Opciones de consulta (I) Visitas (menú superior)

- Filtros que podemos hacer
 - Fechas
 - Tipo de Fecha
 - Proyecto
 - Plataforma Usuario
- Veamos un ejemplo de como funciona analizando la popularidad de los líderes políticos españoles actuales ([enlace a la consulta](#))
- A continuación veamos dos cuestiones
 - Tipos de gráficos que podemos emplear (línea, barra, radial, ...)
 - Exportación de los datos (CSV, PNG, JSON)

- Es una herramienta para ver las popularidad de eventos en diferentes contextos, por ejemplo a continuación vamos a ver la popularidad a través de diferentes proyectos de Wikipedia de una acontecimiento histórico como la Guerra Civil Español
 - [Guerra Civil Española en es.wikipedia](#) (7.085.887 visitas)
 - [Spanish Civil War en en.wikipedia](#) (12.872.615 visitas)
 - [Guerre d'Espagne en fr.wikipedia](#) (3.813.visitas)
 - [Spanische Bürgerkrieg en de.wikipedia](#) (1.957.041 visitas)

Opciones de consulta (II) Langviews (menú superior)

- Podemos buscar la información de una página en todos los idiomas

Hemos buscado Civil Española y [vemos que ocurre en todos los idiomas](#)

← Hacer otra consulta

Guerra civil española 2015-07-01 - 2024-02-29

Lista Gráfico

Enlace permanente Descargar

| # | Idioma | Título de página | Insignias | Visitas ↓ | Promedio diario |
|----------------|--------------------|---|-------------------|-------------------|-----------------|
| Totales | 112 idiomas | 103 títulos únicos | ★ × 5, 🏆 × 1, × 1 | 39.102.532 | 12.351 |
| 1 | en | Spanish Civil War | | 12.872.615 | 4066 |
| 2 | es | Guerra civil española | | 7.085.887 | 2238 |
| 3 | fr | Guerre d'Espagne | | 3.813.248 | 1204 |
| 4 | ru | Гражданская война в Испании | | 2.424.560 | 766 |
| 5 | it | Guerra civile spagnola | | 1.980.162 | 625 |

Opciones de consulta (III) Topviews(menú superior)

- Aquí podemos las páginas webs más populares pero es interesante ya que también podemos buscar webs específicas y conocer su posición específica en el ranking de Wikipedia
- En este caso vemos las guerras en España [más populares en wikipedia en Febrero de 2024](#), curiosamente la más consultada es la Primera Guerra Mundial (posición 30 del ranking) frente a la Guerra Civil Española (posición 325 del ranking)

Opciones de consulta (IV) Redirecciones (menú superior)

- Sí bien no es una herramienta de datos, nos permite conocer cual es el control de autoridades o cual es al entrada oficial sobre un tema concreto
- por ejemplo aquí podemos [cuál es la entrada exacta que deberíamos consultar para la 11-M y cuales son las entradas que redirigen a la misma](#)

★ Otras Herramientas: [xtools](#) - Page History

| | | | |
|----------------------|--|-------------------------|---|
| ID: | 41788272 | Minor edits: | 16 · (34.8%) |
| Wikidata ID: | Q2438463 · 16 sitelinks | IP edits: | 2 · (4.3%) |
| Page size: | 6,792 bytes | Bot edits: | 11 · (23.9%) |
| Total edits: | 46 | (Semi-)automated edits: | 8 |
| Editors: | 35 | Reverted edits: | 0 |
| Assessment: |  Stub | First edit: | 2014-01-30 11:47 · Ravave · +2,684 |
| Pageviews (30 days): | 452 | Latest edit: | 2024-03-07 12:39 · Create a template · +12 |
| | | Max. text added: | 2014-01-30 11:47 · Ravave · +2,684 |
| | | Max. text deleted: | 2014-07-30 15:43 · Mister Xip · -52 |

3. Ejercicios prácticos

1. Utilizando el Corpus español me gustaría conocer la evolución en el uso de los términos “ordenador” “computadora” y “pc”. Describe brevemente la evolución
2. ¿Qué término se utiliza más “escritor independiente” o “escritor freelance”
3. Describeme la popularidad de las siguientes escritoras españolas (rosa montero,almudena grandes,rosa chacel,carmen laforet) a lo largo de las décadas (1980-1990, 1990-2000, etc...)
4. Consulta en el Diccionario Panhispánico de dudas la palabra Yincana. A continuación cuatro voces incorrectas del término Yincana. Compara todos los términos en Ngram (Corpes español) y establece un ranking de la más empleada. ¿A partir de de qué fecha se empieza a emplear más el término Yincana? ¿Serías capaz de encontrar una explicación a este fenómeno?
5. En el siguiente artículo June Casagrande utiliza NGram Viewer para ver el uso de una palabra. ¿qué palabra analiza? ¿qué fenómeno o tendencia identifica mediante NGram? ¿a qué conclusión sobre su uso llega sobre la palabra estudiada?
6. Una rápida con el corpus español
 - a. ¿Qué se utiliza más Clinex, Klinex?
 - b. ¿Qué se utiliza más Clinex, Klinex, Kleenex?
 - c. ¿Qué se utiliza más Clinex, Klinex, Kleenex o pañuelo de papel?
7. Inhóspito e inhabitable ¿qué palabra es la que consideras que deberíamos utilizar y cuándo se produjo el gran cambio en su uso?
8. Según en n-Gram en español 2019 que es más importante para nosotros ¿el dinero, la salud o el amor?
 - a. ¿Dime tres nombres que aparezcan/dependan de dinero?
 - b. ¿Dime tres verbos que aparezcan/dependan de dinero?
 - c. ¿Dime tres adjetivos que aparezcan/dependan de amor?
9. Finalmente realiza el mismo pequeño estudio con el American-English 2019.

10. Explorando las aplicaciones de Ngram

- a. Considerando el tema de su tesis doctoral o su campo de especialización, reflexione sobre cómo los N-Grams de Google, que son secuencias de 'n' palabras extraídas de una vasta colección de textos digitalizados, podrían ser aplicados para enriquecer su investigación. Analice qué patrones, tendencias lingüísticas o evoluciones conceptuales podrían descubrirse mediante este enfoque. Diseñe un experimento que utilice N-Grams para explorar aspectos específicos de su temática, como la frecuencia y la evolución de términos clave a lo largo del tiempo. Después, compartiremos y discutiremos colectivamente los resultados y metodologías aplicadas, en un esfuerzo por ampliar nuestra comprensión y aplicabilidad de los N-Grams en el ámbito académico

11. Miniensayo Wikipedia

- a. A través de las dos herramientas que hemos visto, Visitas y Page History (incluida en Xtools), compara básicamente los datos de algunos de los temas que proponemos. Recaba la información que consideres oportuna y realiza un miniensayo con alguna gráfica comparativa. Te dejo algunas ideas: podrías analizar la evolución de su popularidad en un período largo de tiempo, la comunidad que existe en torno a ambas (por ejemplo, número de usuarios y usuarios destacados). Podrías asimismo determinar cuál es más discutido a través de las ediciones y otros indicadores, o podrías analizar cuál es más popular en el exterior
- b. **Temas a escoger una opción**
 - i. Ciencias políticas: Pablo Iglesias VS Santiago Abascal
 - ii. Patrimonio: Alhambra VS Sagrada Familia
 - iii. Literatura: Javier Cercas VS Javier Marías
 - iv. Cultura Popular: Patrulla X (Xmen) VS Avengers
 - v. Cine: El Padrino VS Centauros del Desierto
 - vi. Tema libre, si lo consideras prueba con los temas que te interesen

Bibliografía básica

- Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022, septiembre 7). Where is the science in Wikipedia? Identification and characterization of scientifically supported contents. Zenodo. <https://doi.org/10.5281/zenodo.6967465>
- Barroso-Hurtado, D., Arroyo-Machado, W., & Torres-Salinas, D. (2021). Formación Profesional Dual: Evolución de red de actores en Twitter. *Educación XX1*, 24(2). <https://doi.org/10.5944/educxx1.28136>
- Bohannon, J. (2011). Google Books, Wikipedia, and the Future of Culturomics. *Science*, 331(6014), 135-135. <https://doi.org/10.1126/science.331.6014.135>
- EL PAÍS: el periódico global. (s. f.). Recuperado 20 de marzo de 2024, de <https://elpais.com/>
- Hai-Jew, S. (2014). Exploring the Google books Ngram viewer for «Big Data» text corpus visualizations.
- Lin, Y., Michel, J.-B., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. July, 169-174.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science (New York, N.Y.)*, 331(6014), 176-182. <https://doi.org/10.1126/science.1199644>
- Orozco-Jutorán, M. (s. f.). Ejemplos de uso de corpus monolingües generales para traducir al español.
- Orozco-Jutorán, M. (2022). Ejemplos de uso de corpus monolingües generales para traducir al español. <https://www.mdx.cat/handle/10503/129716>
- Torres Salinas, D., & Gallego Cuiñas, A. (2023). Big Borges: What Can Big Data Show About a Classic Writer on Social Networks? En A. Gallego Cuiñas & D. Torres-Salinas (Eds.), *Humanities and Big Data in Ibero-America* (pp. 207-234). De Gruyter. <https://doi.org/10.1515/9783110753523-013>
- Torres-Salinas, D., Romero-Frías, E., & Arroyo-Machado, W. (2019). Mapping the backbone of the Humanities through the eyes of Wikipedia. *Journal of Informetrics*, 13(3), 793-803. <https://doi.org/10.1016/j.joi.2019.07.002>
- Wikipedia en español. (2024). En Wikipedia, la enciclopedia libre. https://es.wikipedia.org/w/index.php?title=Wikipedia_en_espa%C3%B1ol&oldid=158129897
- Wikipedia:Content assessment. (2024a). En Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=1210317568
- Wikipedia:Estadísticas. (2024b). En Wikipedia, la enciclopedia libre. <https://es.wikipedia.org/w/index.php?title=Wikipedia:Estad%C3%ADsticas&oldid=157914952>
- Zhang, S. (s. f.). The Pitfalls of Using Google Ngram to Study Language. *Wired*. Recuperado 20 de marzo de 2024, de <https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>