

UNA HERRAMIENTA PARA USAR INTERNET EN INVESTIGACIÓN LEXICOGRÁFICA

Luis Delboy

Resumen:

El objetivo de este artículo es doble: mostrar los recursos disponibles para utilizar Internet como fuente de información lexicográfica y presentar una herramienta desarrollada para la web de la Academia que facilita enormemente la investigación utilizando los sorprendentes recursos de Google. Esta publicación será de especial utilidad para los lexicógrafos sincrónicos.

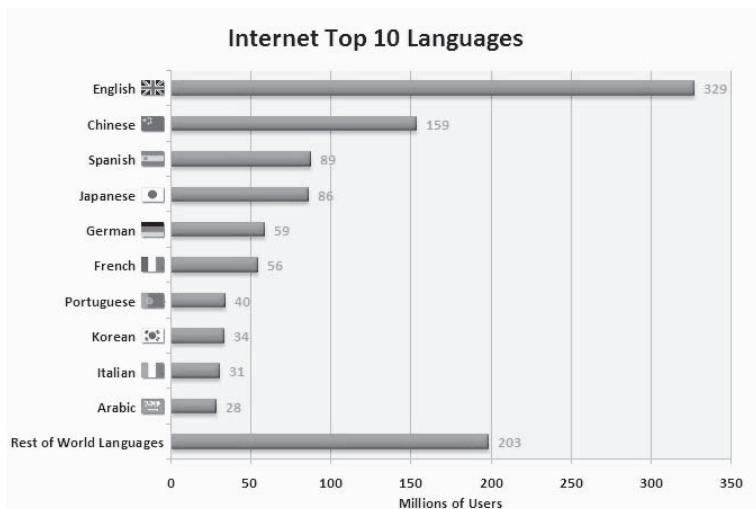
Palabras clave:

Internet; web de la Academia; investigación lexicográfica

El propósito de este documento es presentar a los lectores el agregador de búsquedas desarrollado para la página web de la Academia Peruana de la Lengua, disponible en <http://www.academia-peruanadelalengua.org/lexico>. Esta herramienta permite realizar búsquedas múltiples de términos en los corpus de páginas web indexadas por Google en los países hispanohablantes así como en algunos diccionarios generales y especializados.

INTRODUCCIÓN

Aunque sólo el 20 % ciento de los hispanohablantes acceden a Internet, nuestro idioma es el tercero en cantidad de usuarios. No es posible determinar cuántas páginas web existen en ningún idioma, pero haciendo una proyección, hoy debe haber unos 15,000 millones de páginas publicadas, indexadas por Google. Las páginas en español corresponden aproximadamente al 5% de ese total.



750 millones de páginas, el equivalente a un par de millones de libros es un corpus considerable, que por sí solo puede justificar el interés de lexicólogos y lexicógrafos en utilizarlo como materia prima para su trabajo.

Pero la importancia de Internet como fuente de investigación, no reside exclusivamente en su tamaño. Lo que la hace realmente interesante es que por su propia naturaleza recoge un espectro de normas y dialectos enormemente representativo de la riqueza y diversidad del lenguaje.

¿QUÉ SE PUEDE ENCONTRAR EN INTERNET?

Lo más difícil de encontrar hoy en la web en español, es afortunadamente lo que los lexicógrafos y lexicólogos saben encontrar en el mundo real: textos literarios, libros, especialmente libros raros y curiosos, documentación anterior al mundo digital, básicamente anterior a 1990, publicaciones orientadas a un público ajeno a la informática.

Inversamente, Internet es especialmente rica en textos no literarios contemporáneos. Desde documentación administrativa a textos técnicos, páginas periodísticas y, sobre todo, textos de usuarios extraordinariamente cercanos a la oralidad.

UNAS PALABRAS SOBRE INDEXACIÓN

A diferencia de los textos impresos que sólo son indexados en ediciones de alto nivel que cuentan con índices analíticos, los textos de internet están indexados en su totalidad. Los buscadores como Google, Yahoo o MSN rastrean permanentemente las páginas de Internet e indexan cada una de las palabras que aparecen no sólo en los textos visibles, sino en los textos adjuntos, como los documentos Word o documentos PDF. Esto es lo que permite, por ejemplo, que al buscar en Google la frase *Compilaciones de Peruanismos Anteriores a Juan de Arona* aparezca como resultado el discurso de Incorporación la academia de Enrique Carrión.

Este proceso masivo de indexación es completamente automatizado. Esto quiere decir que no está sujeto a intervención humana para casos particulares. Una vez puesto a andar el programa, el «robot» que recorre las páginas de internet saltando de enlace en enlace lo único que hace es identificar los datos de cada página: dirección

(incluyendo país de origen cuando aparece en la misma), idioma (declarado o implícito) y texto completo.

Al robot no le importa si la ortografía está de acuerdo con alguna norma o si es caprichosa, si es una palabra común o un neologismo absoluto. Simplemente cumple con recoger todas y cada una de las palabras para depositarlas en un servidor. Allí, ocurren unos cuantos procesos:

- En general se eliminan los acentos, de manera que la fruta «tumbo» y el verbo «tumbó» pasan a ser una sola cosa.
- Se analiza los errores de tipeo más frecuentes (si le pedimos a Google «cevche», pregunta atentamente si no queríamos poner «ceviche», pero igual da los 13 resultados donde la palabra estaba escrita sin i.
- Si hay dos grafías comunes (cebiche, ceviche) el servidor asume que la más frecuente es la correcta, y cuando se pide la menos frecuente, también sugiere la primera.

El proceso más importante es el que determina la posición de los resultados. Es decir, cuáles son los resultados que aparecen primero y cuáles son los que aparecen al final.

Una búsqueda exótica como «malarrabia» tiene 330 resultados. Las primeras 50 apariciones de Malarrabia se refieren al plato cuaresmal canónico de los piuranos, lo cual es enteramente comprensible. Más adelante aparece el atole malarrabia mexicano y el postre Malarrabia de camote, guayaba y plátano de los dominicanos.

Un término más frecuente como «mazamorra» genera 85,000 resultados, y curiosamente para los peruanos, el primero no es nuestro típico postre, sino un baile chileno. ¿A qué se debe esa preferencia? A la manera como funciona el algoritmo de Google. La página sobre el baile que aparece en primer lugar tiene más «puntos» porque está escrita con un código sencillo, más fácil de interpretar por los robots que la página sobre el postre y porque hay por lo menos un enlace hacia esa página, mientras que nadie ha puesto enlaces sobre la de la mazamorra morada.

INDEXACIÓN POR PAÍSES

La mayor parte de páginas en español corresponden a dominios .com. .org y .net. Esos dominios son internacionales, y no señalan el país de origen. Sin embargo, aproximadamente **el 10% de todas las páginas están en dominios de país**. .pe , .ar, .mx, etc. Google permite hacer búsquedas estrictamente dentro de dominios nacionales. Esto es de enorme interés para fines lexicológicos y dialectológicos, porque las búsquedas dentro de cada país permiten explorar mejor sus normas.

Por ejemplo en el **Perú** hay **13,300** menciones a **cebiche**, contra **773** a **ceviche**. Es claro que la norma peruana es con **b**. Pero en **México** la relación es inversa: **15,200** **ceviches** contra **2,200** **cebiches**

Cuando se toma en cuenta a la web en su conjunto, la grafía más popular es largamente ceviche con 1,040,000 registros contra 187,000 de cebiche, a la peruana. Dato curioso, los dos únicos países donde cebiche con b existe en proporción significativa son Argentina y Cuba.

TIPOS DE DOCUMENTOS INDEXADOS

Previsiblemente, los buscadores indexan las páginas web convencionales, cuyos contenidos, en general, tienden a ser representativos de las normas cultas oficiales. Pero también indexan sectores donde la aproximación a la norma se busca con menos o ninguna intensidad.

Un caso típico son los blogs, bitácoras virtuales que pueden ser escritas y publicadas por cualquiera que tenga acceso a Internet. Son un fenómeno de expresión individual, donde es fácil encontrar un lenguaje muy cercano a la oralidad. En numerosos casos, esa oralidad se expresa a través de una grafía que procura representar deliberadamente el hablar. Ejemplo: «¿Como estai?» «Orale guey», «Nos juimonos» En otros casos, se puede encontrar atisbos de la grafía paralela que se está desarrollando más en el mundo de los mensajes electrónicos, incluyendo sustitución de c- qu- y k por k («no kiero

karne k-ra»). Muchas veces las mayores incursiones a la oralidad no se presentan en los textos principales de los blogs, sino en los comentarios que los lectores pueden poner libremente en ellos.

Esta inclinación a la oralidad también puede encontrarse en los grupos de interés, lo que se solía llamar «usenet», que es previa a la internet actual y continúa viviendo con gran intensidad.

HERRAMIENTAS PARA ACCEDER A LA INFORMACIÓN

La principal herramienta para acceder a lo que se ha publicado en internet son los buscadores, y entre los buscadores, el más interesante en este momento para el mundo hispano es Google, porque es el que tiene más cobertura en la web en español y porque es por ahora el único que proporciona resultados por países,

Sin embargo, para fines de investigación, usar Google directamente puede ser un ejercicio muy laborioso: para buscar una palabra en distintos países, por ejemplo, es necesario abrir una por una las distintas versiones de Google y hacer búsquedas individuales, una por una.

Adicionalmente existen una cantidad de diccionarios en línea que también son herramientas de consulta. El más importante es el DRAE, pero existen otros que lo complementan,

El Agregador de la Academia Peruana de la Lengua

Para simplificar la búsqueda y comparación de términos, hemos desarrollado para la Academia Peruana de la Lengua un sencillo agregador que permite hacer búsquedas múltiples en distintas fuentes

Perú General Países (a-ch) Países (e-v) Diccionarios Traducciones

Castellano Quechua

Google

La Web [Imágenes](#) [Grupos](#) [Noticias](#) [Más »](#)

palangana [Búsqueda avanzada](#)
[Preferencias](#)

Búsqueda: la Web páginas en español páginas de Perú

Google Peru: páginas con dominio **.pe** en quechua y en español. La sección quechua todavía es débil, y la de español tiene amplia cobertura

Perú	General	Países (a-ch)	Países (e-v)	Diccionarios	Traducciones
google	todos los idiomas	google español	imágenes	grupos	libros

Google General, con cuatro opciones:

- **General** (todas las páginas de la web, en cualquier idioma)
- **Español** (todas las páginas en español sin considerar país)
- **Imágenes** (fotos y dibujos del término buscado)
- **Grupos:** búsqueda en grupos de discusión
- **Libros:** búsqueda en el contenido de libros. Este es un tipo reciente de búsqueda, especialmente interesante para los investigadores. Los resultados pueden mostrar la reproducción del párrafo del resultados y proporcionan adecuada referencia bibliográfica. En algunos casos es posible leer los libros mismos.

Perú	General	Países (a-ch)	Países (e-v)	Diccionarios	Traducciones
Argentina	Bolivia	Colombia	Costa Rica	Cuba	Chile

Países. Cada uno de los países que Google distingue. Están divididos en dos series: de Argentina a Chile y de Ecuador a Venezuela. Perú no está en esta clasificación.

Perú	General	Países (a-ch)	Países (e-v)	Diccionarios	Traducciones
Diccionario de la Real Academia	Diccionario World Com	Definiciones en la Web:	Jergas del Habla Hispana		

Diccionarios. Hemos seleccionado los siguientes diccionarios que pueden ser agregados a esta herramienta de búsqueda. Existen otros diccionarios que no pueden ser agregados.

- **DRAE** El Diccionario de la Real Academia Española
- **World Com:** complementa al diccionario de la Academia, con traducciones, y acceso a foros de discusión donde se debate acerca de palabras.
- **Definiciones en la Web:** servicio de Google que recoge definiciones encontradas en Internet, no necesariamente dentro de diccionarios

- **Jergas del Habla Hispana** Esfuerzo individual que contextualiza expresiones jergales y las presenta comparadas entre países

Perú	General	Países (a-ch)	Países (e-v)	Diccionarios	Traducciones
Inglés: dictionary.com		Inglés: urbandictionary		Español-Inglés	Inglés-Español

Traducciones Esta sección incluye tres diccionarios que incluyen castellano, y la muestra de un diccionario de jerga en inglés generado por los propios usuarios.

- **Dictionary.com** proporciona traducciones a treinta idiomas.
- **Urban Dictionary** es un diccionario de jerga en inglés, alimentado por los propios usuarios, con las más recientes expresiones.
- **Español-Inglés** de Collins es un diccionario de al inglés que incluye expresiones no registradas en otros diccionarios.

El uso de la herramienta es sencillo: basta con colocar la palabra o frase buscada en el área ubicada en la parte superior izquierdo de la página y seleccionar el tipo de búsqueda que se quiere. Para continuar buscando, basta con escoger un nuevo tipo de búsqueda.

En <http://www.academiaperuanadelalengua.org/lexico> está publicada una guía audiovisual que demuestra el uso de la herramienta.