Recurso: Indicadores sobre la Presencia de las Lenguas en la Internet

Daniel Pimienta
Observatorio de la Diversidad Lingüística y Cultural en la Internet
http://funredes.org/lc
Enlace de los recursos: http://funredes.org/lc2022

Resumen

Se requieren indicadores confiables y reproducibles del espacio de lenguas en la Internet para apoyar políticas públicas apropiadas y estudios lingüísticos bien informados. Las fuentes actuales son escasas y, a menudo, fuertemente sesgadas. El modelo para producir indicadores sobre la presencia de lenguas en la Internet, lanzado por el Observatorio en 2017, ha alcanzado un nivel de madurez apreciable y sus productos de datos se comparten en licencia CC-BY-SA 4.0. Alcanza ahora 329 lenguas (locutores de L1 > un millón) y todos los sesgos asociados al modelo han sido controlados a un umbral aceptable, dando crédito a los datos, dentro de un intervalo de confianza estimado de +-20%. Algunos de los indicadores (principalmente el porcentaje de locutores de L1+L2 conectados a la Internet por lengua y derivados) se basan en Ethnologue Global Dataset #24 para datos demo-lingüísticos y ITU, completado por el Banco Mundial, para el porcentaje de personas conectadas a la Internet por país. El resto de indicadores se basa en las fuentes anteriores más una gran combinación de cientos de fuentes diferentes de datos relacionados con los contenidos web por lengua. Este póster de investigación se centra en la descripción de los nuevos recursos lingüísticos creados. Las consideraciones metodológicas sólo se exponen brevemente y se desarrollarán en otro artículo.

Palabras clave: Recurso Lingüístico, Lenguas, Internet, Indicadores, Multilingüismo

1. Introducción

El Observatorio de la Diversidad Lingüística y Cultural en la Internet[1]trabaja con métodos alternativos para medir indicadores de presencia de lenguas en la Internet desde 1996. El método estándar para calcular el porcentaje de contenidos web por lengua es, lógicamente, aplicar un algoritmo de reconocimiento de lengua a todas las páginas web existentes y contar. La gran extensión de la Web hace que este enfoque sea poco práctico, excepto para apuntar a subconjuntos más pequeños, como lo hizo de manera eficiente el Language Observatory Project, antes de que el proyecto se desvaneciera (Mikami, 2005). Los intentos de utilizar ese enfoque aplicándolo a un objetivo con un número limitado de páginas web, que se supone que representan fielmente toda la web, son propensos a grandes sesgos, como se muestra en el método definido por Alis Technologies en 1997[2] y reutilizado en 1999 (Lavoie, 1999) y 2003 (O'Neil, 2003) por OCLC. Ocho mil sitios web fueron seleccionados al azar por números de IP y las conclusiones se derivaron de una medición única, en lugar de una serie repetitiva tratada estadísticamente como una variable aleatoria.

Desde 2011, W3Techs[3], de hecho un excelente y confiable proveedor de estadísticas para tecnologías Web, proporciona resultados actualizados diariamente para los contenidos Web por lengua, aplicando un algoritmo de reconocimiento de lengua a la página de inicio de los 10 millones de sitios web clasificados como los más visitados por Alexa.com[4]. El método es análogo al utilizado para las otras 25 tecnologías Web que son encuestadas por esta empresa, arrojando resultados sumamente interesantes. Sin embargo, las lenguas son un tipo de tecnología web bastante diferente de las bibliotecas de Java Script o los servidores web y el procesamiento de las lenguas de los contenidos web de la misma manera puede generar serios errores. El problema comienza con el enfoque hacia las páginas de inicio de la selección de sitios web: si planea computar los contenidos web, debe enfocar las páginas web para evitar darle el mismo peso a un sitio web de 10 páginas en comparación con un sitio web de diez mil páginas. Además, las páginas de inicio de los sitios web que no están en inglés a menudo incluyen palabras en inglés (ya sea por la voluntad de presentar el sitio en inglés, ya sea porque hay pocas palabras en inglés, como *copyright*, *abstract* o botones de navegación en esa lengua). Eso es una causa de error para el algoritmo. De todos modos, la mayor parte del error está en otra parte: se debe a **la falta de consideración del multilingüismo**, lo que hace que el algoritmo cuente solamente como sitios web en inglés, sitios que ofrecen unas decenas de opciones de lengua en sus interfaces. Muy a menudo, el sitio web configura la opción de lengua automáticamente, de acuerdo con la preferencia del usuario, una práctica cada vez más común, especialmente para los sitios más visitados en el mercado global (Facebook.com es solo un ejemplo) y el algoritmo que cuenta una lengua por página de inicio tomara el inglés en esos casos. No es de extrañar entonces por qué, desde 2011, el porcentaje de

[1] http://funredes.org/lc
[2] https://web.archive.org/web/20010730164601/http://alis.isoc.org/palmares.es.html
[3] http://W3Techs.com
[4] Sitios de recopilación y análisis de tráfico web pertenecientes a la corporación Amazon, a punto de ser retirados del mercado.

inglés en la web se mantiene estable e incluso crece por parte de W3Techs, a pesar de que las evidencias indican que la Internet ha cambiado drásticamente en la última década, con el chino convirtiéndose en la primera lengua en términos de usuarios, y la mayoría de las lenguas asiáticas y el árabe están en auge. La Web es hoy probablemente más multilingüe que la humanidad. Según los últimos datos de Ethnologue, la proporción de locutores L1+L2 sobre locutores L1 es 10 361 716 756 / 7 231 699 136 = 1,43. Nadie se sorprenderá entonces de que más del 50% de los sitios web muestren páginas en más de una única lengua. El no prestar la debida atención al multilingüismo es un sesgo inaceptable para tales estudios. W3Techs podría, sin cambiar su selección actual de sitios web y su programa principal, corregir sus sesgos con algunos cambios tales como:

- Analizar las opciones de lengua que se ofrecen en la página de inicio y contabilizar cada opción, no solo la versión en inglés.
- Encontrar un método para obtener una estimación aproximada del número de páginas del sitio web y multiplique cada versión lingüística por ese número para contar las páginas web en lugar de los sitios web.
- Cuando el algoritmo informa más de una lengua en la página de inicio, como precaución, no contabilizar el sitio web como inglés, sino como su segunda lengua.

Los nuevos resultados serán drásticamente diferentes...

El problema preocupante es que, por la unicidad de la fuente, la calidad comprobada del resto de sus encuestas, su larga trayectoria y su eficiente mercadeo, un gran porcentaje de la comunidad de investigación lingüística (y hacedores de políticas públicas) está tomando los datos de W3Techs como insumos confiables. Desafortunadamente, las buenas teorías alimentadas por números incorrectos difícilmente pueden proporcionar resultados correctos.

El ejemplo más sintomático de la situación lo da el agregador de estadísticas Statista[5]que titula su anuncio de 2022 sobre lenguas en la Internet[6] con una declaración que suena como un hecho: *el inglés es la lengua universal de la Internet*, respaldado por datos de W3techs, donde los contenidos web en inglés representan el 63,7% del total, mientras que los en chino solo el 1,3%.

Al mismo tiempo, el Observatorio de la Diversidad Lingüística y Cultural en la Internet computa el inglés y el chino, con el mismo porcentaje, en torno al 20%, mientras que el hindi, con sus 224 millones de internautas, alcanza el 3,8% (frente al solo 0,1% medido por W3Techs) y concluye su último anuncio con esa frase: *la transición de Internet entre el dominio de las lenguas europeas, el inglés a la cabeza, hacia las lenguas asiáticas y el árabe, el chino a la cabeza, está muy avanzada y el ganador es el multilingüismo, pero las lenguas africanas tardan en ocupar su lugar.*

Una, al menos, de las dos fuentes será extremadamente incorrecta y los investigadores deben tener cuidado y verificar los sesgos de un método antes de sacar conclusiones de los datos producidos...

2. Los métodos alternativos

Durante el periodo 1998-2007, el método alternativo del Observatorio que proporcionó series coherentes se limitaba al inglés, alemán y las 5 lenguas latinas (francés, italiano, español, portugués y rumano). Usó motores de búsqueda para contar un vocabulario comparable[7] para cada lengua (Pimienta, 2009). Después de 2007, la "evolución hacia el marketing" de los motores de búsqueda hizo que el método quedara obsoleto, ya que sus reportes del número de ocurrencias de una palabra buscada se volvieron poco confiables.

Hoy en día, se computan 329 lenguas, aquellos con más de un millón locutores L1, según Ethnologue, una limitación adoptada para evitar sesgos demasiado fuertes por consecuencia de la hipótesis de trabajo del enfoque: *todas las lenguas en un mismo país se computan con el mismo porcentaje de locutores conectados a la Internet, la cifra nacional proporcionada por la UIT/Banco Mundial.* Esta hipótesis prohíbe comparar lenguas dentro de un país, es difícilmente aplicable a lenguas con pocos locutores y tiende a ofrecer un sesgar positivo a las lenguas de inmigración en los países en desarrollo (que pueden estar menos conectados que el promedio) y negativo a las lenguas europeas en los países en desarrollo. (que suelen estar mejor conectados que la media).

El método actual es **una aproximación indirecta a los contenidos**, basada en la observación experimental de que

---

[55] http://statista.com De paso, no perderé la oportunidad de cuestionar la ética de dos fenómenos emergentes que podrían estar correlacionados. 1) Demasiados investigadores perezosos citan a Statista como fuente de datos en lugar de la verdadera fuente. 2) Statista ofrece algunos datos en acceso gratuito, pero la identificación de la fuente de esos datos solo es accesible para clientes pagos. Hagámoslo más simple entonces y citemos a Google como la madre de todas las fuentes o, aún más simple, ¡citemos a Internet como la matriz de todas las fuentes! 😊

[6] https://es.statista.com/chart/26884/idiomas-en-internet/

[7]Un conjunto de palabras para cada idioma, seleccionadas con muchas precauciones lingüísticas, cuyas ocurrencias eran reportadas por los motores de búsqueda y permitían, por conteo, los resultados.

la relación entre el porcentaje mundial de contenidos y el porcentaje mundial de locutores conectados se ha mantenido siempre entre 0,5 y 1,5 (para lenguas con plena existencia digital).

Se sugiere algún tipo de **ley económica** natural, que vincularía, para cada lengua, **la oferta** (contenidos web y aplicaciones) a **la demanda** (locutores conectados a la Internet). Cuando aumenta el número de personas conectadas, el número de páginas web lógicamente aumenta al mismo tiempo, más o menos en la misma proporción. Esto sucede porque gobiernos, empresas, instituciones educativas, etc., y algunas personas crean contenidos para responder a esa demanda.

Además, las encuestas y los estudios han informado constantemente que los usuarios de la Internet prefieren usar su lengua materna y también aprovechan la oportunidad de usar, como segunda opción, su(s) segundo(s) lengua(s)[8].

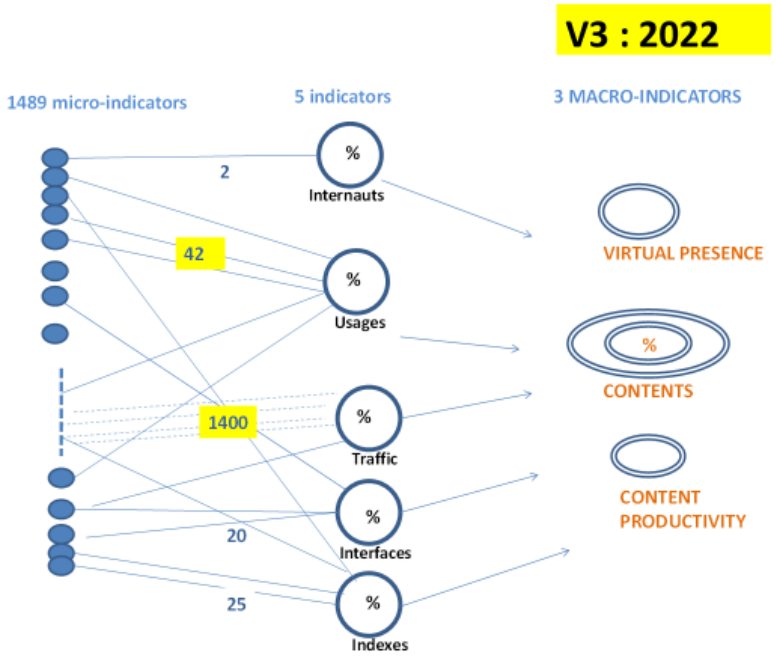
Así, dependiendo de cada lengua, existe algún tipo de modulación de dicho ratio, para hacerlo por encima o por debajo de uno. Esto significaría que algunas lenguas tienen más propensión a la producción de contenido que otras, dependiendo de un conjunto de factores relacionados con las lenguas dentro del contexto de su país, tales como:

- Obviamente, la cantidad relativa de **locutores L2**, ya que algunas personas producen, por ejemplo por razones económicas, contenidos en una lengua diferente al de su lengua materna,

Pero también:

- La proporción del **tráfico** de Internet según la tarifa del país, el contexto cultural o educativo.
- El número de **suscripciones** a redes sociales y otras aplicaciones de Internet.
- El **soporte tecnológico digital** de la lengua y su presencia en las interfaces de las aplicaciones y programas de traducción, lo que facilitaría o no la producción.
- El nivel de inmersión del país donde vive el locutor en cuanto a los progresos de la **Sociedad de la Información** (comercio electrónico, solicitudes gubernamentales para pagar impuestos, etc.).

Entonces, si fuera posible recopilar varios indicadores sobre cada una de las características mencionadas, se aproximaría la fluctuación de la modulación de los contenidos web en torno a uno y deduciría de alguna manera la proporción de contenidos. Este es el núcleo del método y se sintetiza en el siguiente diagrama que muestra todos los indicadores que se procesan para cada lengua y la cantidad correspondiente de fuentes que está utilizando el modelo. La primera y la segunda versión de la metodología están totalmente documentadas, incluido el análisis de todos los sesgos; consulte la síntesis en (Pimienta, 2019). La descripción detallada de la versión 3 está en camino.



[8] Véase, por ejemplo, el informe de la encuesta de la Unión europea en https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556o, para el desafiante caso de la India, este informe:https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

Figura1: Diagrama sobre indicadores

Este diagrama ha evolucionado, de la versión 1 a la versión 3, a lo largo de la ardua tarea de perseguir los sesgos, en términos de número de fuentes y también en términos de indicadores. El cálculo del modelo establecido, bastante complejo, se basa en gran medida en una variedad de **operaciones de ponderación**, con, la mayoría de las veces, **el vector de porcentaje de personas conectadas por país**, que es el núcleo matemático del proceso. Las fuentes de indicadores por lengua disponibles son escasas; la mayoría de los indicadores se obtienen por país y, raros son los que cubren más que un subconjunto de países. La fuente de datos se extrapola a todos los países, ponderando con el vector mencionado, y la transformación de los datos por país en datos por lengua se obtiene ponderándolos con los datos demo-lingüísticos (cantidad de locutores de cada lengua en cada país).

3. Indicadores que arroja el modelo

Para cada una de las 329 lenguas procesadas, el modelo produce los siguientes indicadores por lengua (tenga en cuenta que todos los porcentajes mundiales se basan en cifras L1+L2 y representan la proporción correspondiente para cada lengua).

Indicadores intermedios:
Internautas: locutores conectados a la internet
Usos
Tráfico
Interfaces y programas de traducción: en términos de porcentaje mundial de los números correspondientes de aplicaciones y programas de traducción soportados
Índices: en términos de porcentaje mundial de la calificación de los países en los parámetros de la Sociedad de la Información

Resultados del modelo (también llamados macroindicadores):
Locutores conectados: porcentaje dentro del total mundial de locutores L1+L2 conectados a Internet
Contenidos: porcentaje de contenidos Web (computado como el promedio de los 5 indicadores intermedios)
Productividad de contenidos: ratio contenidos/internautas
Coeficiente de presencia virtual: ratio contenidos/cuota mundial de locutores

Indicadores más avanzados
Ciber-geografía de las lenguas: una distribución de los resultados del modelo cumulados por familias de lenguas (europeas, asiáticas, árabe, americanas, africanas)

Indicador de globalización cibernética

CGI (L) = (L1 + L2)/L1 (L) x S (L) x C (L)

Donde:

L1+L2/L1(L) es la proporción de multilingüismo de la lengua L
S(L) es el porcentaje de países del mundo que tienen locutores de la lengua L
C(L) es el % de locutores de la lengua L conectados a la Internet.

Este es un indicador de las ventajas estratégicas de una lengua en el ciberespacio.

Además, para algunas lenguas, se ha mostrado la lista de países que tienen los mayores porcentajes de locutores conectados.

Los archivos de Excel con los resultados finales se pueden descargar desde http://funredes.org/lc2022.

Se encuentra en proyecto una base de datos de acceso a los resultados, con posibilidad de consulta por nombre de lengua o código iso.

4. Ejemplos de indicadores producidos

A continuación se presentan algunos ejemplos de datos, limitados a los mejores resultados, para la mayoría de los casos. Los mismos datos están disponibles para cualquiera de los 329 lenguajes procesados.

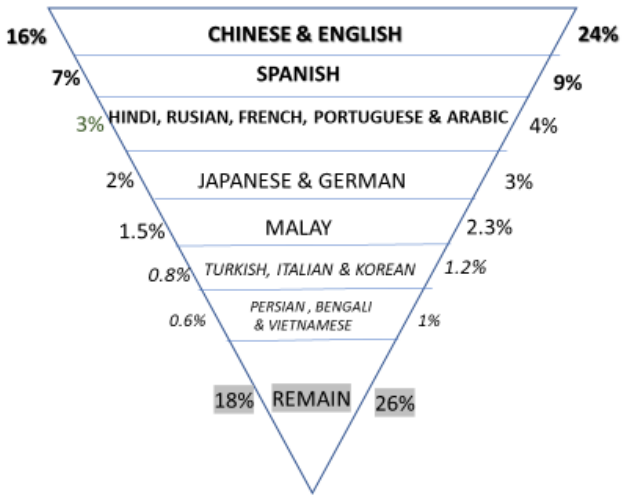


Figura 2: Porcentaje de ventanas de contenido para las principales lenguas

La pirámide invertida debe leerse como una expresión del intervalo de confianza: el porcentaje de contenidos web en chino (o inglés) está entre el 16% y el 24%, el resto de lenguas juntas representan entre el 18% y el 26% del total.

LENGUA	LOCUTORES CONECTADOS
noruego	**96,89%**
danés	**96,42%**
sueco	**93,94%**
catalán	**92,88%**
japonés	**92,63%**
finlandés	**92,07%**
alemán. suizo	**91,55%**
limburgués	**91,42%**
flamenco occidental	**91,30%**
holandés	**91,14%**
gallego	**91,07%**
sajón. Superior	**89,81%**
estonio	**89,26%**
alemán estándar	**89,17%**
letona	**89,04%**
bávaro	**88,24%**

Tabla 1: Lenguas principales en locutores conectados

RANGO Contenidos L1+L2	ISO	LENGUAS	Internautas L1+L2	Población Mundial. L1+L2	Locutores Conectados L1+L2	Contenidos L1+L2	Presencia Virtual L1+L2	Productividad Contenidos L1+L2
1	**zho**	***chino***	18,46%	14,72%	71,38%	21,60%	1,47	1,17
2	**eng**	**inglés**	14,83%	13,01%	64,86%	19,60%	1,51	1,32
3	**spa**	**español**	6,79%	5,24%	73,72%	7,85%	1,50	1,16
4	**hin**	**hindi**	4,19%	5,80%	41,16%	3,76%	0,65	0,90
5	**rus**	**ruso**	3,51%	2,49%	80,32%	3,76%	1,51	1,07
6	**fra**	**francés**	2,98%	2,58%	65,80%	3,33%	1,29	1,12
7	**por**	**portugués**	2,99%	2,49%	68,43%	3,13%	1,26	1,05
8	**ara**	***árabe***	3,97%	3,53%	63,99%	3,09%	0,87	0,78
9	**jpn**	**japonés**	1,99%	1,22%	92,63%	2,66%	2,18	1,34
10	**deu**	**alemán**	2,04%	1,30%	89,17%	2,37%	1,82	1,16
11	**msa**	***malayo***	2,36%	2,36%	56,93%	1,96%	0,83	0,83
12	**tur**	**turco**	1,17%	0,85%	78,05%	1,14%	1,35	0,98
13	**ita**	**italiano**	0,87%	0,66%	75,83%	1,00%	1,53	1,14
14	**kor**	**coreano**	0,90%	0,79%	65,16%	0,98%	1,24	1,09
15	**fas**	***persa***	1,08%	0,81%	75,91%	0,88%	1,09	0,82
16	**ben**	**bengalí**	1,11%	2,58%	24,55%	0,88%	0,34	0,79
17	**vie**	**vietnamita**	0,92%	0,74%	70,96%	0,85%	1,15	0,92
18	**urd**	**urdu**	0,95%	2,22%	24,38%	0,66%	0,30	0,70
19	**tha**	**tailandés**	0,80%	0,59%	77,95%	0,65%	1,12	0,82
20	**pol**	**polaco**	0,60%	0,39%	87,09%	0,63%	1,59	1,04
21	**mar**	**marathi**	0,69%	0,96%	41,06%	0,58%	0,60	0,83
22	**tel**	**telugu**	0,68%	0,92%	41,69%	0,56%	0,60	0,82
23	**tam**	**tamil**	0,61%	0,82%	42,15%	0,51%	0,62	0,83
24	**jav**	**javanés**	0,62%	0,66%	53,76%	0,44%	0,66	0,70
25	**nld**	**holandés**	0,38%	0,24%	91,14%	0,41%	1,73	1,08
26	**guj**	**guyaratí**	0,44%	0,60%	41,47%	0,36%	0,61	0,83
27	**ukr**	**ucranio**	0,40%	0,32%	71,02%	0,35%	1,09	0,88
28	**kan**	**kannada**	0,41%	0,57%	41,11%	0,33%	0,59	0,82
29	**ron**	**rumano**	0,32%	0,23%	79,57%	0,30%	1,29	0,93
30	**aze**	***azerbaiyano***	0,33%	0,23%	81,54%	0,28%	1,21	0,85
		RESTO	**22,60%**	**30,10%**		**15,13%**		
		TOTAL	**100,00%**	**100,00%**		**100,00%**		

Tabla2: Indicadores principales de las 30 principales lenguas en porcentaje de contenidos

Debe leerse así: el inglés representa el 13% de la población mundial L1+L2 y el 14,8% de la población conectada a la Internet; el 64,7% de los locutores de inglés L1+L2 están conectados a la Internet; el 19,6% de los contenidos de la Web están en inglés; el coeficiente de presencia virtual del inglés es de 1,5, lo que significa que los contenidos en inglés están sobrerrepresentados en un factor superior al 50%; la productividad de contenido del inglés es de 1,32, la más alta después del japonés.

Las macro lenguas están escritas en itálico.

LENGUA	PRESENCIA VIRTUAL
japonés	**2,18**
noruego	**1,88**
alemán estándar	**1,82**
sueco	**1,82**
danés	**1,78**
holandés	**1,73**
finlandés	**1,69**
catalán	**1,68**
alemán suizo	**1,63**
polaco	**1,59**
italiano	**1,53**
estonio	**1,51**
ruso	**1,51**
inglés	**1,51**
hebreo	**1,50**
griego	**1,50**
español	**1,50**
chino	**1,47**
letona	**1,46**
gallego	**1,46**

Tabla 3: Lenguas principales en presencia virtual

LENGUA	PRODUCT. CONTENIDOS
japonés	**1,34**
inglés	**1,32**
chino	**1,17**
alemán estándar	**1,16**
español	**1,16**
italiano	**1,14**
francés	**1,12**
noruego	**1,10**
sueco	**1,10**
coreano	**1,09**
holandés	**1,08**
ruso	**1,07**
griego	**1,07**
caboverdiano	**1,05**
danés	**1,05**
portugués	**1,05**
finlandés	**1,04**
polaco	**1,04**
catalán	**1,03**
alemán, suizo	**1,02**
hebreo	**1,00**

Tabla 4: Lenguas principales en productividad de contenidos

LENGUAS DE (*)	ÁFRICA	AMÉRICAS	MUNDO ÁRABE	ASIA	EUROPA	PACÍFICO (**)
Internautas	**29,8%**	**56,7%**	**64,0%**	**49,3%**	**82,6%**	
Contenidos	**2,89%**	**0,22%**	**3,09%**	**44,77%**	**45,39%**	
Presencia Virtual	**0,28**	**0,68**	**0,87**	**0,65**	**1,39**	
Productividad contenidos	**0,51**	**0,68**	**0,78**	**0,72**	**0,95**	
Población L1+L2	**9,15%**	**0,31%**	**3,53%**	**48,21%**	**30,91%**	
Porcentaje de conectados	**5,18%**	**0,32%**	**3,89%**	**44,60%**	**39,51%**	
NÚMERO DE LENGUAS	**138**	**8**	**1**	**135**	**47**	**0**

Tabla 5: Ciber-geografía de las lenguas

(*) Tiene que entenderse como lenguas nativas. Por ejemplo, las 8 lenguas indígenas de las Américas con más de un millón de locutores de L1 incluidas en el modelo son: aimara, guaraní, criollo haitiano, húnsrik, criollo jamaiquino, q'eqchi', kiche y quechua.
(**) No se incluyen lenguas del Pacífico ya que ninguna tiene más de 1 millón de locutores.

Se lee de esa manera: 29.8 % de los locutores de lenguas africanas están conectados en la Internet, juntos representan el 5 % de la población mundial conectada cuando representan el 9 % de la población mundial; los contenidos en lenguas africanas representan el 2.9 % del total con una productividad de 0.5 y una presencia virtual de 0,3.

LENGUA	CGI	CGI%
inglés	**1,61**	**14,24%**
francés	**1,09**	**9,66%**
alemán	**0,42**	**3,75%**
ruso	**0,31**	**2,76%**
español	**0,27**	**2,40%**
Arábica	**0,18**	**1,56%**
malayo	**0,17**	**1,51%**
italiano	**0,17**	**1,50%**
chino	**0,16**	**1,46%**
portugués	**0,15**	**1,37%**
tailandés	**0,15**	**1,37%**
romaní	**0,15**	**1,35%**
turco	**0,15**	**1,34%**
griego	**0,15**	**1,31%**
ucranio	**0,15**	**1,31%**
polaco	**0,13**	**1,15%**
persa	**0,12**	**1,10%**
rumano	**0,12**	**1,06%**
hindi	**0,12**	**1,04%**

Tabla 6: Indicador de Globalización Cibernética

La segunda columna se calcula dividiendo el valor CGI por el total de CGI para todas las lenguas procesados. Se menciona como una forma de medir, por ejemplo, el peso relativo de las dos primeras posiciones, cercano al 25% del total.

CHINO	L1+L2	% Conectados	CONECTADOS	% DE LOS CONECTADOS.
TOTAL	**1 525 335 340**	**71,38%**	**1 088 735 519**	**100%**
China	**1 448 870 000**	**70,64%**	**1 023 512 815**	**94,01%**
China-Taiwán	**37 320 000**	**88,82%**	**33 148 541**	**3,04%**
China–Hong Kong	**10 942 800**	**92,41%**	**10 112 585**	**0,93%**
Malasia	**7 838 700**	**89,56%**	**7 019 949**	**0,64%**
Singapur	**4 026 000**	**75,88%**	**3 054 766**	**0,28%**
Estados Unidos	**2 894 390**	**88,50%**	**2 561 503**	**0,24%**
Vietnam	**2 500 000**	**70,64%**	**1 766 054**	**0,16%**
Indonesia	**2 054 000**	**53,73%**	**1 103 542**	**0,10%**
Tailandia	**1 729 000**	**77,84%**	**1 345 918**	**0,12%**
Canadá	**1 212 600**	**97,00%**	**1 176 222**	**0.11%**
Filipinas	**1 010 280**	**43,03%**	**434 689**	**0,04%**
RESTO	**4 937 570**	**71,04%**	**3 507 738**	**0,32%**

Tabla 7: Reparto de locutores de chino conectados por países principales

HINDI	L1+L2	%CON.	CONECTADO	% DE CONEX.
TOTAL	**600 800 970**	**41,15%**	**247 258 401**	**100%**
India	**596 000 000**	**41,00%**	**244 360 000**	**98,87%**
Kuwait	**700 000**	**98,60%**	**690 200**	**0,28%**
Estados Unidos	**643 000**	**88,50%**	**569 048**	**0,23%**
Nepal	**1 307 600**	**25,00%**	**326 900**	**0,13%**
Sudáfrica	**463 000**	**68,00%**	**314 840**	**0,13%**
Arabia Saudita	**171 000**	**97,86%**	**167 345**	**0,07%**
Australia	**160 000**	**86,54%**	**138 472**	**0,06%**
Canadá	**111 000**	**97,00%**	**107 670**	**0,04%**
Yemen	**316 000**	**30,00%**	**94 800**	**0,04%**
RESTO	**929 370**	**52,63%**	**489 127**	**0,20%**

Tabla 8: Distribución de locutores de hindi conectados por países principales

5. Referencias bibliográficas

Ethnologue Global Dataset (2022). https://www.ethnologue.com/product/ethnologue-global-dataset-0

Lavoie B.F., O'Neill E. T. (1999). How "World Wide" is the Web? *Annual review of OCLC Research*, https://web.archive.org/web/20031006155123/http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003496

Mikami Y., et al. (2005). The Language Observatory Project (LOP), In *Poster Proceedings of the Fourteenth International World Wide Web Conference*, pp. 990-991, May 2005, Japan

O'Neill E.T., Lavoie B.F., Bennett R. (2003). Trend in the Evolution of the Public Web: 1998 – 2002. *D-Lib Magazine*, 9.4
http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

OIF (2022). Le français dans le monde, Gallimard, ISBN : 9782072976865.
Synthèse en ligne: https://francophonie.org/sites/default/files/2022-03/Synthèse_La_langue_française_dans_le_monde_2022.pdf

Pimienta, D., Prado D., Blanco A. (2009). Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, in *UNESCO Publications for the World Summit on the Information Society*, CI.2009/WS/1
http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

Pimienta D. (2019). Indicators of Languages in the Internet, in Proceedings of International Conference Language Technologies for All (LT4All), 4-6 December 2019, UNESCO, Paris; PP 315-319
https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.79.pdf

6. Agradecimientos

Los estudios de la versión 3 fueron financiados por la Organización Internacional de la Francofonía y los resultados alimentaron el Capítulo de Internet de (OIF, 2022).

La idea de utilizar varias fuentes de datos por país y transformarlas en datos por lengua fue concebida por primera vez por Daniel Prado en 2012.