**Una historia breve de la observación de las lenguas en la Internet**

**Daniel Pimienta, Observatorio de la Diversidad Lingüística y Cultural en Internet, agosto 2022**

La medición del espacio de representación de las lenguas en la Internet[1] no fascina a las multitudes y, sin embargo, lo que está en juego, a nivel lingüístico, cultural, socioeconómico e incluso geopolítico, está lejos de ser neutral. Muchas lenguas se encuentran amenazadas o simplemente en declive y la intensidad de su presencia en la Internet es un indicador determinante de su futuro. En 2020, el comercio electrónico representa el 20% de las ventas minoristas globales totales y las plataformas deben hablar la lengua de sus clientes para competir.

Una leyenda recorre la Internet desde sus orígenes sobre una exclusividad norte-americana que se traduciría también en la creencia de una dominación estable y perenne, que haría del inglés, para siempre, la llamada "*lengua franca*" del ciberespacio.

Este recorrido histórico pretende desbrozar la desinformación sobre este tema, fuente de lamentables renuncias a un objetivo que todos los actores del desarrollo reconocen como fundamental: la importancia de la creación de contenidos en lengua local y de políticas públicas dirigidas a promover las requeridas condiciones (lucha contra la factura digital, acompañada de programas de alfabetización al mundo digital).

Este recorrido en el tiempo seguirá el camino particular de una organización pionera en este campo y que, a pesar de muchos azares, se ha mantenido, hasta el día de hoy, como un actor destacado: el Observatorio de la diversidad lingüística y cultural en la Internet.

Este observatorio nació en 1995, como un proyecto de una ONG (organización no gubernamental) de investigación-acción en el campo, ubicada en Santo Domingo (República Dominicana) y cuyo nombre expresa la esencia de la visión: Fundación Redes y Desarrollo (FUNREDES)[2]. Este proyecto luego se convirtió en un programa[3], por la continuidad de su accionar, entre 1996 y 2017, fecha en que se disolvió esta ONG. Posteriormente, el Observatorio tomó una forma asociativa independiente en Francia, al tiempo que reivindicó y mantuvo su patrimonio histórico[4].

FUNREDES se formó en 1993 a partir de una matriz original creada dentro de la Unión Latina, luego de que el autor lograra convencer, en 1987, a su visionario secretario general, Philippe Rossillon, preocupado por la estrecha asociación entre el inglés y las redes informáticas, de que las redes iban a vivir un destino deslumbrante y que era importante defender, desde adentro, la diversidad lingüística y cultural, antes que oponerse a un surgimiento incontenible.

Entre 1988 y 1993, cuando este programa de la Unión Latina se independizó, transformándose en FUNREDES, manteniendo vínculos armoniosos y activos con la Unión Latina, se desarrolló una importante actividad, en parte con el apoyo de la Unión Europea y la colaboración de la

[1]Los dos indicadores básicos son, para cada lengua, el porcentaje global de locutores conectados y el porcentaje de contenidos en la Web.
[2]Fundación Redes y Desarrollo (http://funredes.org)
[3] http://funredes.org/lc
[4] http://funredes.org/lc/JO-OBDILCI.pdf

UNESCO. Se han creado tres redes nacionales (Perú, República Dominicana, Haití) y el primer software de interfaz de red multilingüe desde una computadora personal (MULBRI[5]), acciones piloto de un gran proyecto de red latinoamericana (REDALC[6]), con, como originalidad, una visión centrada en el usuario, su cultura y su lengua y la importancia de la alfabetización digital e informacional, así como la integración de los profesionales de la información; todo ello en contraste con un contexto donde la visión tecnológica era muy predominante y transversal.

FUNREDES tomó entonces el relevo, en 1993, y se resistió a una visión simplista de la brecha digital, que consistiría sólo en una cuestión de acceso tecnológico; impulsó los conceptos de *brecha de contenidos* y *brecha lingüística*, partiendo del principio de una necesidad natural de navegar el ciberespacio en la lengua materna y evitar la trampa de la aculturación. También promovió la importancia de asociar la *alfabetización informacional* a cualquier política de lucha contra la brecha digital (Pimienta, 2007).

Fue la expresión del presidente Chirac, durante la cumbre de la Francofonía en Cotonou, en 1995, quien vio en la naciente Internet una entidad enteramente angloparlante y estadounidense, que provocó el deseo de confrontar este prejuicio con la realidad, tratando de medir el lugar de las lenguas y culturas en la Internet. Rápidamente, la Unión Latina, a través de su Director de Terminología e Industrias de las Lenguas, el argentino Daniel Prado, se sumó al proceso y comenzó una larga y fructífera colaboración.

En ese entonces, los motores de búsqueda reportaban fielmente el número de ocurrencias de una palabra o frase en todas las páginas web indexadas, las cuales representaban una proporción de páginas existentes superior al 80%. Así que fue una gran herramienta para tales estudios.

Para lenguas latinas[7], inglés y alemán, se constituyó una muestra de palabras, en cada lengua, cuidadosamente seleccionadas para representar un conjunto semántica y sintácticamente equivalente, tarea más fácil de decir que de hacer. Para la cultura, se compuso una lista de un vasto conjunto de personajes en una serie de categorías (literatura, ciencia, música, cine…). Las medidas de ocurrencias de los buscadores permitieron, con herramientas estadísticas tradicionales, obtener resultados en ambos campos, lengua y cultura[8].

La proporción del inglés en la Web se midió entre 75% en 1997 y 52% en 2001. Las personalidades latinas estuvieron bien representadas en los sectores culturales donde la separación entre comercio y cultura fue más marcada; en cambio, en los sectores regidos por las leyes del mercado, la cultura de los Estados Unidos se impuso claramente. Los resultados sobre cultura se estabilizaron, destacando los personajes más "globalizados" y el estudio se interrumpió tras la tercera campaña, en 2001[9].

En cuanto a los resultados sobre lenguas, hubo que interrumpirlos en 2007 porque los motores de búsqueda, en primer lugar Google, habían evolucionado de forma incompatible con el método utilizado: los retornos de búsquedas sobre el número de ocurrencias perdieron credibilidad, la cobertura del índice fue reduciéndose enormemente, pudiendo caer por debajo del 5% de todas las páginas de la Web, y el juego publicitario comenzaba a distorsionar los resultados de los buscadores que se volvían diferentes para cada usuario, según sus búsquedas

[5] https://funredes.org/gopher/b/6/6.4/6.4.3/6.4.3.2/6.4.3.2.2/lb.html
[6] https://funredes.org/gopher/b/6/6.1/6.1.1/6.1.1.1/lg.html
[7] español, francés, italiano, portugués y rumano.
[8] Ver https://funredes.org/lc2005/espanol/index.html. Para un resumen consultar (Pimienta, 2001).
[9] https://funredes.org/lc2005/espanol/index.html

pasadas y otros factores relacionados con la información personal recopilada por ellos. Esto convirtió a este proyecto en uno de los primeros testigos de los efectos de la evolución tóxica de los motores de búsqueda y la actual inversión donde quien pensaba buscar es realmente el sujeto de la búsqueda inversa que Google procesa, marca del naciente *capitalismo de vigilancia*, que se impondría y cambiaría la trayectoria de la Internet, desde la utopía inicial de la democracia participativa (Pimienta, 2005) hasta la situación actual de amenaza contra las democracias (Pimienta, Rodríguez, 2020).

## Prehistoria: hasta 1997

Es útil comprender el contexto de este período en lo que respecta a las redes. La Web nació en 1992, cuando las diversas redes de investigación (como BITNET/EARN o HP-Net) y libertarias (como Usenet o Fidonet) convergieron hacia el protocolo de Internet.

Internet nació con un sistema de codificación de caracteres de 7 bits (ASCII[10]), que permite que la lengua inglesa, sin marcas diacríticas, se codifique sin obstáculos, pero perjudica a la mayoría de las otras lenguas que deben codificar más caracteres diferentes de los 128 permitidos. Tomará algunos años, con la creación del protocolo MIME[11], en 1997, para ir superando progresivamente este límite en correos electrónicos y páginas Web, hasta el éxito del estándar UNICODE[12] que acomoda los alfabetos de diferentes lenguas, en constante evolución para localizar más lenguas.

El antecesor de la Web, Gopher, un sistema simple de menús estructurados en árbol, habría permitido medir fácilmente el lugar de las lenguas, pero aparentemente nadie tuvo esta idea y es razonable pensar que en el nacimiento de la Web, en 1992, más del 80% de los "sitios Gopher", generalmente universidades, estaban en inglés; del mismo modo es razonable situar la cuota inicial del inglés en la Web, al momento de su nacimiento, en 1992, en el 100%.

Los primeros intentos de medir el lugar de las lenguas en la Web datan del período 1997-2000 y se describen, con los del período posterior, 2000-2005, en (Pimienta et al., 2009).

## La efervescencia inicial: 1997-2007

Una docena de actores se presentaron durante el período, algunos del mundo de la investigación, otros impulsados por consideraciones de marketing. Si nos centramos en los elementos más serios desde el punto de vista metodológico[13], la idea de una evolución de la presencia del inglés en la Web para el período, pasando del 80% al 50%, tiene sentido. Pero los elementos más influyentes no son necesariamente los más serios y tres estudios específicos de EEUU (en 1997, 2000 y 2003), que comparten una metodología inválida[14], así como una operación de marketing

[10] https://fr.wikipedia.org/wiki/American_Standard_Code_for_Information_Interchange
[11] https://fr.wikipedia.org/wiki/Multipurpose_Internet_Mail_Extensions
[12] https://en.wikipedia.org/wiki/Unicode
[13] Además de los estudios del Observatorio ya mencionados (1998-2007), el de Xerox (2000), los del Proyecto Observatorio de la Lengua Japonesa (LOP), el del proyecto catalán del IDESCAT y las mediciones indirectas realizadas con Google con la técnica "del complemento del conjunto vacío". Ver (Pimienta et al., 2009) para más detalles y fuentes.
[14] El método común consistía en seleccionar al azar, a partir de números de IP, 3000 sitios y aplicarles una vez un algoritmo de reconocimiento de lengua. Para validar este enfoque habría sido necesario repetirlo varias veces y tratar estadísticamente los múltiples resultados como una variable aleatoria, estudiando su distribución (media, varianza, etc.). Lanzar el dardo una sola vez al centro del blanco no prueba la habilidad del tirador...

de uno de los buscadores de la época, Inktomi, conteniendo un grosero error, anclan en los medios la idea de una presencia estable del inglés en torno al 80%, en la década 1997-2007.

¡Sin embargo, la estabilidad es realmente la última característica creíble para un campo en crecimiento exponencial y geográfico como la Internet! Hará falta la publicación por parte de la UNESCO de dos textos sobre el tema (UNESCO, 2006) y (Pimienta et al., 2009), para que los medios proyecten finalmente un valor más realista cercano al 50% para la presencia del inglés en la Web. Durante el mismo período, una empresa, GlobalReach[15], ha producido datos creíbles sobre la distribución de usuarios de la Internet por lengua, desde el año 2000.

El nacimiento del proyecto académico Language Observatory Project (LOP), coordinado por Yoshiki Mikami, de la Universidad de Nagasaki en Japón (Mikami 2005), en forma de un consorcio mundial de universidades, utilizando una técnica basada en algoritmos de reconocimiento de lenguaje y métodos poderosos de exploración web[16], da esperanza en la profesionalización académica de esta materia. Rápidamente se emprenden colaboraciones entre la LOP, FUNREDES y la Unión Latina, tres miembros activos de la Red Mundial por la Diversidad Lingüística, MAAYA[17], nacida en 2006, bajo el impulso de Adama Samassekou, al margen de la Cumbre Mundial para la Sociedad de la Información, y aglutinadora de acciones significativas en este campo[18].

Esta ligera efervescencia sobre el tema se calmará sin embargo en el próximo período 2007-017

## Cruzando el desierto: 2007-2017

El proyecto UPC/IDESCAT finalizó en 2006; GlobalReach dejó de producir sus datos en 2007; y en 2011, la LOP desapareció, arrasada por el tsunami que afectó a Japón. Por su parte, el Observatorio, en el marco de FUNREDES, y bajo el paraguas de MAAYA, propone, entre 2010 y 2013, un gran proyecto europeo de investigación sobre el tema y gestiona, con el apoyo conjunto de la Unión Latina, la OIF[19]y la UNESCO, para crear un poderoso consorcio europeo de investigación[20] que responde a dos convocatorias del programa europeo de investigación[21]. Pero la prioridad no parece fundamental para la Unión Europea y el esfuerzo queda en vano, a pesar de un primer intento que se sitúa a medio punto de evaluación del umbral exigido y una importante inversión humana y económica en el periodo 2010-2012.

Por lo tanto, es necesario decidir ceñirse al enfoque artesanal: en el marco de MAAYA, y con el apoyo de la OIF, continúan las colaboraciones para estudios puntuales centrados en el francés en la Internet, que nutren el trabajo del Observatorio de la OIF acerca de la lengua y la obra “Le

---

[15] https://web.archive.org/web/20000412001030/http://www.glreach.com/globstats/index.php3

[16]La LOP no busca explorar todo el universo sino que se concentra en espacios geográficos más limitados que hacen posible esta exploración sistemática.

[17]https://web.archive.org/web/20150704174747/http://www.maaya.org/?lang=es

[18]MAAYA organizó cuatro Simposios Internacionales sobre Multilingüismo en el Ciberespacio, en 2009, 2011, 2012 (ver https://web.archive.org/web/20150704174747/http://www.maaya.org/?lang=es) y, en 2019, bajo el impulso del brasilero Claudio Menezes (ver https://doity.com.br/iv-simc). Por iniciativa de Daniel Prado, MAAYA reunió a varios autores en una obra que sigue siendo referencia en el tema (MAAYA, 2012). Para la historia de MAAYA, ver (Pimienta y Prado, 2016).

[19]Organización internacional de la Francofonía: https://francofonía.org

[20]Ver https://web.archive.org/web/20180831105048/http://dilinet.org/mod/resource/view.php?id=105

[21]Ver https://funredes.org/lc/dilinet/es/.

français dans le monde" (OIF, 2014) y (OIF, 2019), o sobre el español (Pimienta y Prado, 2016); sin embargo, ya no es posible la producción sistemática de indicadores para varias lenguas.

En 2012, la Unión Latina suspendió sus actividades; en 2017 FUNREDES cesa sus actividades; hacia el final del período, MAAYA encuentra dificultades para mantener sus actividades y la antorcha del tema del plurilingüismo en el ciberespacio es retomada por el sector IFAP de la UNESCO[22] y su dinámica rama rusa, dirigida por el carismático Evgeny Kuzmin, que reúne regularmente a actores sin fines de lucro en torno al tema de la diversidad lingüística y cultural en el ciberespacio, entre 2008 y 2019[23], los primeros en coordinación con MAAYA. La UNESCO sigue siendo la entidad de las Naciones Unidas que se ocupa formalmente de este tema[24] (UNESCO, 2015).

Durante este período, dos actores comerciales se vuelven esenciales, ya que son los únicos que producen datos y logran mantener sus actividades hasta el día de hoy:

- InternetWorldStats, una empresa de marketing en Colombia, produce datos sobre la Internet a nivel mundial desde 2002, incluyendo, desde 2004, su ranking de las 10 lenguas más utilizados en la Internet, en términos de usuarios[25].
- W3Techs, empresa que produce datos en torno a tecnologías Web, y que incluye, desde 2011, en su lista con fuerte color tecnológico, una clasificación de lenguas muy populares en la Web, que actualiza diariamente[26] y cuya historia mantiene[27].

El número de teorías o elaboraciones lingüísticas construidas sobre el edificio de estas dos fuentes es impresionante; sin embargo, la experiencia adquirida por el Observatorio y el análisis de los múltiples sesgos del método W3Techs le permiten estimar que los datos producidos exageran el lugar del inglés en proporciones muy significativas pero, hasta 2017, no le es posible oponer otras figuras.

## Nacimiento y maduración de una alternativa: desde 2017

El más fiel apoyo de este proyecto desde sus inicios, la OIF, permitió en 2017, a través de un proyecto de MAAYA, generar un nuevo enfoque que permita al Observatorio volver a producir indicadores (Pimienta, 2017). El modelo establecido parte de la idea inicial de Daniel Prado, que guio el trabajo entre 2012 y 2017, con el apoyo del español Álvaro Blanco[28]: multiplicar las más diversas fuentes cuantitativas sobre la presencia de lenguas en Internet y, para compensar la evidente escasez de estas fuentes, completar con las fuentes por país, mucho más frecuentes, transformándolas, a partir de datos demolinguísticos, en fuentes por lengua.

En 2017, el Observatorio consiguió dar coherencia matemática a este enfoque indirecto, estructurarlo y generalizarlo para extraer resultados válidos para un gran número de lenguas[29]. El método permite una estimación directa del porcentaje de personas conectadas por lengua y

[22] https://en.unesco.org/programme/ifap
[23] http://www.ifapcom.ru/en/722/
[24] Ver https://www.unesco.org/en/communication-information/multilingualism-cyberspace
[25] https://www.internetworldstats.com/stats7.htm
[26] https://w3techs.com/technologies/overview/content_language
[27] https://w3techs.com/technologies/history_overview/content_language/ms/y
[28] Quien seguirá de asistente principal del responsable de proyecto hacia el día de hoy.
[29] Primero, para limitar los sesgos causados por las suposiciones simplificadoras requeridas para que el modelo funcione, el límite se establece en 149 lenguas con más de 5 millones de locutores como primera lengua.

una estimación indirecta del porcentaje de contenidos por lengua en la Web, así como otros indicadores útiles, y marca un importante punto de inflexión en esta historia de medición. El modelo establecido se basa en 3 tipos de fuentes:

1. Datos demolinguísticos: número de locutores de cada lengua en cada país, diferenciando entre primera (L1) y segunda lengua (L2).
2. El porcentaje de personas conectadas a la Internet por país, datos actualizados anualmente por la UIT[30], que juega un papel esencial en los cálculos.
3. El mayor número posible de fuentes cuantitativas sobre lenguas o países, en relación con elementos relacionados directamente (por ejemplo, número de suscriptores a redes sociales) o indirectamente (por ejemplo, número de teléfonos móviles por habitante) a la Internet, clasificadas entre *tráfico*, *usos*, *contenido*s, *índice*[31], *interfaces*.

A partir de estos 3 pilares y partiendo de una serie de cálculos con los datos de entrada (extrapolación de datos incompletos, ponderaciones demo-lingüísticas, ponderaciones por porcentaje de personas conectadas, medias, medias reducidas, método de cuartiles, etc.) el modelo generó indicadores de salida[32].

Como era de esperar, los resultados contradicen los datos de W3Techs, que están sujetos a sesgos muy significativos[33] y el porcentaje de inglés se encuentra, en 2017, en el nivel esperado por la extrapolación de curvas preexistentes: el 30% de los sitios estarían en inglés y el francés es la cuarta lengua en cuanto a contenido, por detrás del chino y el español, con una ventaja cómoda sobre los siguientes: ruso, alemán, portugués y árabe.

El método es complejo, su implementación requiere mucho tiempo, dada la cantidad de fuentes a encontrar, evaluar y luego utilizar, entonces el énfasis, desde el principio, ha estado en el análisis completo de los sesgos inducidos por el método, por su hipótesis de trabajo[34] y por las muchas fuentes procesadas.

Entre 2017 y 2022 se dedica, por tanto, un gran esfuerzo a la eliminación de sesgos, con notables avances en 2021, gracias al apoyo del Ministerio de Relaciones Exteriores de Brasil, a través del Instituto Internacional de la Lengua Portuguesa (https://iilp.cplp.org), bajo la coordinación y apoyo lingüístico de la Cátedra UNESCO de Políticas Lingüísticas para el Multilingüismo (https://www.unescochairlpm.org), en la persona de su gerente, el lingüista brasilero Gilvan Müller de Oliveira, que permite utilizar la mejor fuente demolinguística existente[35] y extender los resultados a las 329 lenguas con más de un millón de locutores L1, a partir de las cuales se producen interesantes resultados sobre la ciber-geografía de las lenguas (Pimienta, 2021).

---

[30]Unión Internacional de Telecomunicaciones: http://itu.int

[31]Este elemento hace referencia a rankings de países en su avance hacia la sociedad de la información (gobierno electrónico, datos abiertos, etc.).

[32]Ver (Pimienta, 2017) para detalles del método y su sustrato teórico.

[33]Para comprender las razones de estos sesgos, siendo el principal la falta de consideración del multilingüismo en la Web, ver (OIF, 2022) o (Pimienta, 2022).

[34]Una suposición de trabajo necesaria para el modelo es que los locutores de diferentes lenguas en un país comparten el mismo porcentaje de conexión a la Internet (la tasa promedio nacional proporcionada por la UIT). Este supuesto prohíbe comparar lenguas dentro de un mismo país, es difícil de aplicar a lenguas con un bajo número de locutores y tiende a dar un sesgo positivo para las lenguas de inmigración en los países en desarrollo (que pueden estar menos conectados que la media) y, por el contrario, un sesgo negativo para las lenguas europeas en los países en desarrollo (que tienden a estar mejor conectadas que la media). Por lo que respecta a las lenguas de Francia, se utilizó otro enfoque (Pimienta y Prado, 2014) y se puso en línea una base de datos: http://baseldf.fr.

[35]Ethnologue (https://www.ethnologue.com)

| | Lenguas africanas | Lenguas americanas | Árabe como macro-lengua | Lenguas asiáticas | Lenguas europeas | Resto de lenguas | TOTAL |
|---|---|---|---|---|---|---|---|
| **Internautas %** | **29,8%** | **56,7%** | **64,0%** | **49,3%** | **82,6%** | **47,06%** | **56,91%** |
| **Contenidos** | **2,89%** | **0,22%** | **3,09%** | **44,77%** | **45,39%** | **3,64%** | **100%** |
| **Presencia virtual** | **0.31** | **0.71** | **0.88** | **0.93** | **1.47** | **0.47** | **1** |
| **Productividad de contenidos** | **0,56** | **0,69** | **0.79** | **1.00** | **1.15** | **0,57** | **1** |
| **Locutores L1+L2** | **9,21%** | **0,31%** | **3,53%** | **48,24%** | **30,91%** | **7,81%** | **100%** |
| **Población conectada** | **5,21%** | **0,32%** | **3,89%** | **44,63%** | **39,51%** | **6,36%** | **100%** |
| **Lenguas con L1>1M** | **138** | **8** | **1** | **135** | **47** | | **329** |

En marzo de 2022, el método alcanza la madurez y se controlan todos los sesgos, a costa de redefinir algunos indicadores[36] y de una reformulación de la visión metodológica: se trata de una aproximación indirecta de los contenidos, a partir de la observación experimental de que la relación entre el porcentaje global de contenidos y el porcentaje global de locutores conectados se ha mantenido siempre entre 0,5 y 1,5 (para lenguas con existencia numérica completa).

Esto sugiere la existencia de una especie de ley económica natural, que vincularía, para cada lengua, la oferta (contenidos web y aplicaciones) con la demanda (locutores conectados a la Internet). Cuando aumenta el número de personas conectadas, el número de páginas web aumenta rápidamente, más o menos en la misma proporción. Esto es así porque gobiernos, empresas, instituciones educativas, etc., y parte de los nuevos usuarios están creando contenidos para atender esta demanda y/o aumentar la oferta.

Es importante tener en cuenta que las encuestas y los estudios han informado consistentemente que los internautas prefieren usar su lengua materna y también aprovechan la oportunidad de usar su(s) segundo(s) lengua(s) como segunda(s) opción(es)[37].

Así, dependiendo de cada lengua, se produce una especie de modulación de la ratio mencionada (% de contenidos / % de conectados), para hacerlo más o menos superior o inferior a uno, teniendo unas lenguas mejor productividad de contenidos que otras, en función de un conjunto de factores que les afectan, en cada uno de sus contextos nacionales, tales como:

- Evidentemente, el correspondiente número de locutores de L2, ya que algunas personas producen, por ejemplo por motivos económicos, contenidos en una lengua distinta a su lengua materna.

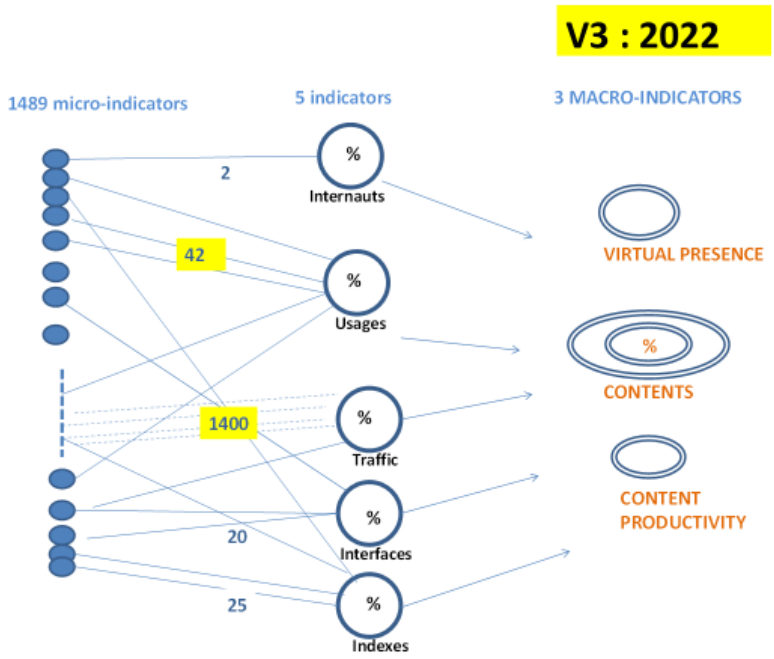
[36]Las estadísticas lingüísticas de Wikimedia, además de la aplicación de Internet con mayor diversidad lingüística, de una calidad y extensión raras en este contexto, alimentaron un indicador de contenido en la entrada del modelo. En la versión 2021 se desarrolla un esfuerzo muy importante para compensar los sesgos naturalmente occidentales de Wikipedia y otros elementos de la galaxia Wikimedia, estableciendo fórmulas que penalizan las versiones lingüísticas basadas en copias de otras lenguas y débilmente actualizadas, e integrando todas las existentes. enciclopedias en línea en todas las lenguas. El resultado es decepcionante y es necesaria una conclusión ineludible: las enciclopedias no son un fiel reflejo de la realidad de los contenidos por lengua. El indicador se elimina y se encuentra en la salida del modelo, dando lugar a una redefinición, que finalmente aporta más claridad a los conceptos utilizados. Prescindir de la mejor estadística existente sobre lenguas en Internet, la de Wikimedia, es una frustración, pero esta decisión impuesta por la realidad ha permitido obtener un modelo donde se dominan todos los sesgos y la lógica matemática, establecida sobre la ponderación. operaciones, muchas veces con la distribución de personas conectadas por país (e indirectamente por lengua), terminaron imponiendo su consistencia al modelo.

[37]Véase, por ejemplo, el informe de investigación de la Unión Europea: https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556 y para el difícil caso de la India, este informe: https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

Pero también:

- La proporción del tráfico Internet, según la tarifa, el contexto cultural o educativo del país.
- El número de suscripciones a redes sociales y otras aplicaciones de la Internet.
- El soporte tecnológico digital de la lengua y su presencia en interfaces de aplicaciones y programas de traducción, que pueden o no facilitar la producción de contenidos.
- El nivel de inmersión del país donde vive el hablante en cuanto a la manifestación de la sociedad de la información (comercio electrónico, gobierno electrónico, etc.).

Así, si fuera posible recoger diferentes indicadores sobre cada una de las características mencionadas, podríamos medir las modulaciones de este indicador en torno al valor uno y deducir la proporción de contenidos. Esto es exactamente lo que logra el modelo establecido al utilizar cerca de 1500 fuentes (microindicadores) para calcular 5 indicadores que permiten producir los resultados del modelo (macroindicadores), tal como se presenta en el siguiente diagrama.




Y esto permite obtener los siguientes resultados para las 30 lenguas con mayores porcentajes de contenidos[38].

| Rango | | | | Población | Altavoces | | Presencia | Productividad |
|---|---|---|---|---|---|---|---|---|
| Rango | | | % Internautas | % Población Mundial | % conectados | % Contenidos | Presencia Virtual | Productividad de Contenidos |
| L1+L2 | ISO | LENGUAS | L1+L2 | L1+L2 | L1+L2 | L1+L2 | L1+L2 | 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 |
| 4 | **rus** | **ruso** | 3,51% | 2,49% | 80,32% | **3,76%** | 1.51 | 1.07 |
| 4 | **fra** | **francés** | 2,98% | 2,58% | 65,80% | **3,33%** | 1.29 | 1.12 |

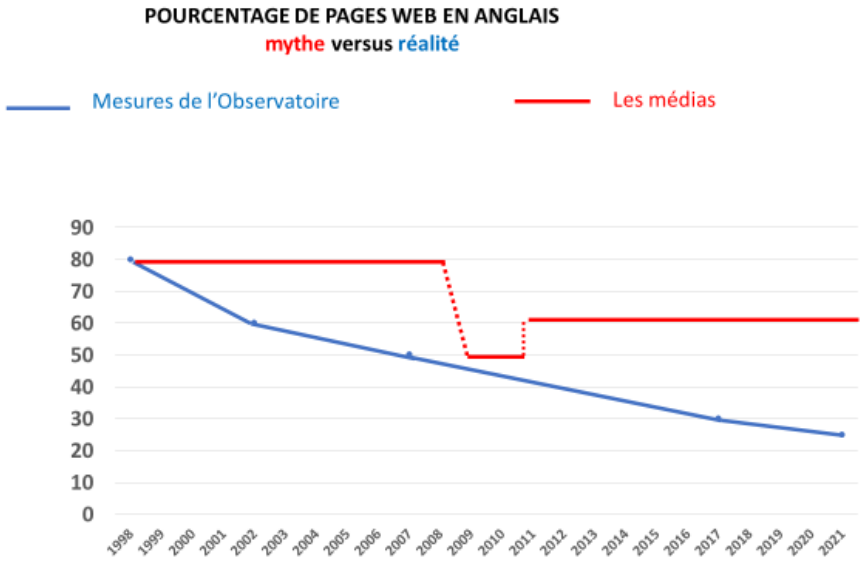
[38]Los porcentajes se expresan en relación con la población mundial L1+L2. Según la fuente Ethnologue, la población mundial (L1) es de 7 231 699 136 personas, mientras que la población mundial de locutores de L1 o L2 es de 10 361 716 756, es decir que más del 43% de la población mundial sería al menos bilingüe.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | **por** | **portugués** | 2,99% | 2,49% | 68,43% | **3,13%** | 1.26 | 1.05 |
| 4 | **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 |
| 9 | **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 |
| 12 | **ita** | **italiano** | 0,87% | 0,66% | 75,83% | **1,00%** | 1.53 | 1.14 |
| 12 | **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 |
| 15 | **ben** | **bengalí** | 1,11% | 2,58% | 24,55% | **0,88%** | 0.34 | 0.79 |
| 15 | **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 |
| 18 | **tha** | **tailandés** | 0,80% | 0,59% | 77,95% | **0,65%** | 1.12 | 0.82 |
| 18 | **pol** | **polaco** | 0,60% | 0,39% | 87,09% | **0,63%** | 1.59 | 1.04 |
| 18 | **mar** | **marathi** | 0,69% | 0.96% | 41,06% | **0,58%** | 0,60 | 0.83 |
| 18 | **tel** | **telugu** | 0,68% | 0,92% | 41,69% | **0,56%** | 0,60 | 0.82 |
| 18 | **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 |
| 24 | **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 |
| 26 | **ukr** | **ucranio** | 0,40% | 0,32% | 71,02% | **0,35%** | 1.09 | 0.88 |
| 26 | **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 |
| 29 | **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%** | | |

El intervalo de confianza en los resultados es del orden del +-20%, por lo que se debe considerar que las lenguas marcadas con el mismo color tienen un porcentaje idéntico de contenidos. Así, en 2022, el francés es la cuarta lengua de Internet en cuanto a contenidos, junto con el hindi, el ruso, el portugués y el árabe, representando cada una de estas lenguas entre el 3 y el 4% del total de contenidos de la Internet, mientras que el inglés y el chino representan cada uno entre el 16 y el 24 % del contenido.

Todos los resultados se dejan en libre acceso (CC-BY-SA 4.0) en la página http://funredes.org/lc2022.

Queda el arduo camino por recorrer para romper con los (malos) hábitos adquiridos durante más de 10 años de utilizar, sin precaución, fuentes seriamente sesgadas que sugieren erróneamente que el inglés se ha mantenido estable entre 2011 y 2022 por encima del 50% del contenido Web… mientras que en realidad su lugar ronda ahora el 20%, a par con el chino! La historia se repite y las siguientes dos curvas resumen esta historia de desinformación.




## El futuro de las lenguas en la Internet

A largo plazo, llegará el momento en que los locutores de todas las lenguas serán usuarios con índices de conectividad superiores al 90%, como ocurre hoy con el noruego, el danés, el sueco, el catalán, el japonés y el finlandés, por citar a los campeones. Pero, para los contenidos, probablemente permanecerán brechas significativas entre las respectivas representaciones de las lenguas en el mundo y en el ciberespacio, algunos sobrerrepresentados (presencia virtual mayor a 1) mientras que otros estarán menos favorecidos. Uno de los indicadores elaborados, el grado de ciber-globalización de una lengua:

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

Dónde:

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

proporciona información sobre las ventajas estratégicas de una lengua en el ciberespacio. Este indicador muestra que la ventaja del inglés continuará y que su presencia virtual seguirá estando entre las más altas. El francés se sitúa luego, con una notable diferencia con respecto a los siguientes (alemán, ruso y español). La demografía sigue siendo el factor clave, asociado con la capacidad de las lenguas para atraer a aprendedores de una segunda lengua. Las lenguas africanas, que hoy siguen siendo las menos presentes en el ciberespacio, podrán tomar su revancha, hacia 2050, cuando la población de África podría haberse duplicado, a condición que allí se resuelve la brecha digital. Esta perspectiva también podría beneficiar a las lenguas europeas más presentes en este continente: el inglés y el francés en primer lugar.

En cuanto al futuro de la observación de lenguas, sólo podemos esperar que este campo atraiga nuevas voluntades y conozca una mayor diversidad de enfoques de los que todos puedan

beneficiarse. Aunque por fin el sector se está profesionalizando y el uso de herramientas algorítmicas de reconocimiento de lenguaje va acompañado de la seriedad metodológica requerida[39], enfoques artesanales como el del Observatorio conservarán su espacio.

De este recorrido histórico acerca de las lenguas en el ciberespacio se puede destacar que donde ha sido el mas dinámico ha sido dentro del espacio de cruce entre las lenguas latinas, con una fuerte presencia de actores francófonos, hispanófonos y lusófonos, tal vez porque existe ahí una sensibilidad particular al tema de la **diversidad**, tema muy presente dentro de las culturas asociadas a esas lenguas.

## BIBLIOGRAFÍA


- Pimienta D., Lamey B. (2001). « El español en la sociedad de la información: Internet en español », Congreso de la lengua española, Valladolid
https://congresosdelalengua.es/valladolid/paneles-ponencias/sociedad-informacion/pimienta-d.htm

- Pimienta D. (2005). "At the Boundaries of Ethics and Cultures: Virtual Communities as an Open-Ended Process Carrying the Will for Social Change (the "MISTICA" experience)" in *Localizing the Internet. Ethical Issues in Intercultural Perspective*., Capurro, R. & al. (Eds.). Schriftenreihe des ICIE Bd. 4, München: Fink Verlag, 2005
https://funredes.org/mistica/english/cyberlibrary/thematic/icie/

- Mikami Y. et al. (2005). "The language observatory project (LOP)", *Proceedings of the 14th international conference on World Wide Web, WWW 2005*, Chiba, Japan
https://www.researchgate.net/publication/221022705_The_language_observatory_project_LOP

- UNESCO. (2006). « Mesurer la diversité linguistique dans l'Internet », CI.2005/WS/06
https://unesdoc.unesco.org/ark:/48223/pf0000142186_fre

- Pimienta D. (2007). «Brecha digital, brecha social, brecha paradigmática», en *Brecha digital y nuevas alfabetizaciones. El papel de las bibliotec*as, Biblioteca de la Universidad Complutense de Madrid, ISBN: 978-84-691-3466-5 – 2008 – p. 11
https://eprints.ucm.es/id/eprint/8224/3/Brecha_digital_y_nuevas_alfabetizaciones.pdf

- Pimienta D., Prado D., Blanco A. (2009). « Douze années de mesure de la diversité linguistique sur l'Internet: bilan et perspectives » UNESCO CI-2009/WS/1.
https://unesdoc.unesco.org/ark:/48223/pf0000187016_fre

- Maaya. (2012). « Net.lang. Réussir le cyberespace multilingue », Coordonné par Vannini L., Le Crosnier H., C&F Éditions, ISBN 978-2-915825-08-4, 2012 - https://cfeditions.com/NetlangFR/ (téléchargeable en français, anglais ou russe).

- OIF. (2014). « Le français dans l'Internet », *Rapport 2014 "La langue française dans le monde"*, pp. 501-541, Nathan, 2014 -http://www.francophonie.org/Rapports-Publications.html

- Pimienta D., Prado D. (2014). « Étude sur la place des langues de France dans l'Internet », *Langue & Recherche*, Délégation générale à la langue française et aux langues de France, 2014.
http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et-la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet


[39]En este sentido, se hacen sugerencias a W3Techs en (Pimienta, 2021) para permitir que su algoritmo supere los sesgos existentes, a un coste moderado, y ofrecer la observación creíble que todo el mundo, el Observatorio primero, quiere.

- UNESCO. (2015). « Una Década de promoción del plurilingüismo en el ciberespacio», CI-2015/WS/5 https://unesdoc.unesco.org/ark:/48223/pf0000232743_spa
- Pimienta D., Prado D. (2016). "Medición de la presencia de la lengua española en la Internet: métodos y resultados", en *Revista Española de Documentación Científica 39(3)*, e141. ISSN-L:0210-0614. - http://dx.doi.org/10.3989/redc.2016.3.1328

- Pimienta D., Prado D. (2016) « Un milliard de Latins… dans l'Internet ? », in *Hermès, La Revue, 2016/2 (n° 75) Langues romanes : un milliard de locuteurs*. http://www.cairn.info/revue-hermes-la-revue-2016-2.htm

- Prado D. (2016) « Les langues romanes minoritaires et l'Internet », *in Hermès, La Revue, 2016/2 (n° 75) Langues romanes : un milliard de locuteurs.* http://www.cairn.info/revue-hermes-la-revue-2016-2.htm

- Pimienta D., Prado D. (2016) "Ten Years of MAAYA, the World Network for Linguistic Diversity: Time for Balance and Perspectives", in Proc. of *Multilingualism in Cyberspace*, IFAP/UNESCO – P184. http://www.ifapcom.ru/files/2016/UGRA_ENGL_BLOK_WEB.pdf

- Pimienta D. (2017). « Uma abordagem alternativa para a produção de indicadores da presença de línguas na Internet", IV Simpósio Internacional sobre Multilinguismo no Ciberespaço, Brasília, 2019 https://doity.com.br/iv-simc
http://funredes.org/lc2017/Alternativa%20Lingua%20Internet.docx (em português)
http://funredes.org/lc2017/Alternativa%20Lengua%20Internet.docx (en español)
El texto original en inglés ha sido producido en 2017 pero las traducciones se elaboraron en 2019.

- OIF. (2019). « La présence de la langue française dans le cyberespace (synthèse)", *Rapport 2019 "La langue française dans le monde"*, pp. 337-341, OIF, Gallimard. https://www.francophonie.org/sites/default/files/2021-04/LFDM-20Edition-2019-La-langue-française-dans-le-monde.pdf
Étude complète accessible à http://observatoire.francophonie.org/2018/Place-francais-sur-Internet-D-Pimienta.pdf

- Pimienta D., Rodríguez Leal LG., (2020), - "¡Va de retro Internet! Una visión crítica de la evolución de la Internet desde la sociedad civil", *Revista Ibero-Americana de Ciência da Informação*, V13 N3, pp. 979-1000 https://periodicos.unb.br/index.php/RICI/article/view/33041/27497

- Pimienta D. (2021) «Internet y Diversidad lingüística: ciber-geografía de las lenguas con mayor número de locutores», *LinguaPax Review 2021, Language Technologies and Language Diversity*, p. 29 https://www.linguapax.org/wp-content/uploads/2022/02/LinguapaxReview9-2021-low.pdf

- OIF. (2022). « La présence de la langue française dans le cyberespace », dans *"La langue française dans le monde 2019-2022"*, Gallimard/OIF- ISBN : 9782072976865
https://gallimard.fr/Catalogue/GALLIMARD/Hors-serie-Connaissance/La-langue-francaise-dans-le-monde
Synthèse accessible en ligne : https://www.francophonie.org/sites/default/files/2022-03/Synthèse_La_langue_française_dans_le_monde_2022.pdf

- Pimienta D. (2022). «Recurso: Indicadores de la presencia de lenguas en Internet», traducción del artículo en ingles presentado en SIGUL2022/LREC2022, Marsella http://funredes.org/lc2022/Res.Ind.lang.Internet.es.pdf