**Un enfoque alternativo para producir indicadores de la presencia de las lenguas en la Internet**

**Daniel Pimienta**

**pimienta@funredes.org**

**Observatorio de lenguas y culturas en la Internet**

**http://funredes.org/lc**

**Red Mundial para la Diversidad Lingüística**

**http://maaya.org**

**Octubre 2019**

**RESUMEN**

Dada la dificultad para obtener datos fiables sobre la presencia de las lenguas en la Internet, un enfoque alternativo se presenta para producir una serie de indicadores para las 140 lenguas con más de 5 millones de hablantes. El método se basa en la recopilación de una serie de micro-indicadores que miden presencia de lenguas o países en distintos espacios o aplicaciones de la Internet. Un cálculo de ponderación transforma porcentajes mundiales por país en porcentaje mundiales por lengua. Se definen cinco indicadores relacionados con la presencia de las lenguas en la Internet: *usuarios, tráfico, uso, contenidos, índices societales e interfaces*. A partir de ahí 4 macro-indicadores se deducen, para cada lengua tratada: *poder, capacidad, gradiente y productividad de contenidos*. Los valores de los indicadores se calculan procesando un conjunto de 369 micro-indicadores y el de los macro-indicadores se obtienen con operaciones simples a partir de los indicadores. El artículo recuerda que importantes sesgos han marcado la historia de la cibermetría lingüística. Todos los posibles sesgos derivados del método, de sus hipótesis explicitas o implícitas y de sus fuentes se analizan y finalmente se propone una estimación que tiene en cuenta esos sesgos. Comparaciones críticas y análisis de los respectivos sesgos se hacen con los escasos indicadores existentes (W3Techs y InternetWorldStats). Se concluye con recomendaciones para que la cibermetría lingüística regrese donde debe permanecer: en el mundo académico y de investigación. Créditos pertenecen a la Organización Internacional de la Francofonía (OIF), que financió una serie de estudios sobre el lugar del francés en la Internet, lo que permitió el desarrollo de este método y a Daniel Prado, el primero que concibió la idea de recoger diversas fuentes para la medición de lenguas en la Internet, así como para transformar datos por país en datos por lengua.

Palabras claves: lenguas, Internet, indicadores, cibermetría lingüística, sesgos

Contenido

[1. Introducción 5](#_Toc58576964)

[2. Metodología 6](#_Toc58576965)

[2.1 Indicadores de lenguas en la Internet 6](#_Toc58576966)

[2.2 Micro-indicadores por lengua o país 8](#_Toc58576967)

[2.2.1 Por lengua 8](#_Toc58576968)

[2.2.2 Por país 9](#_Toc58576969)

[2.3 Fuentes de información 9](#_Toc58576970)

[2.4.1 El caso de la lengua materna (L1) 10](#_Toc58576971)

[2.4.2 El caso de la segunda lengua (L2) 11](#_Toc58576972)

[2.5 Proceso 13](#_Toc58576973)

[2.5.1 La extrapolación 13](#_Toc58576974)

[2.5.2 Proceso para los indicadores 14](#_Toc58576975)

[2.5.3 Proceso para lograr resultados diferenciados por temáticas 17](#_Toc58576976)

[2.5.4 Tipos de ponderaciones utilizadas 17](#_Toc58576977)

[3. Resultados 18](#_Toc58576978)

[4. Análisis de los resultados 20](#_Toc58576979)

[4.1 Acerca de Wikipedia 20](#_Toc58576980)

[4.2 Comparación de los resultados con los de InternetWorldStats (IWS) 21](#_Toc58576981)

[4.3 Sensibilidad de los factores 23](#_Toc58576982)

[5. Limitaciones metodológicas, análisis de sesgo y controles realizados 24](#_Toc58576983)

[5.1 ¡Ojo con los sesgos! 24](#_Toc58576984)

[5.1 Limitaciones y sesgos propios del método 25](#_Toc58576985)

[5.2 Lenguas 26](#_Toc58576986)

[5.2.1 Selección de la fuente para el cálculo de L1 26](#_Toc58576987)

[5.2.2 El caso de L2 26](#_Toc58576988)

[5.2.3 La reducción del número de lenguas 28](#_Toc58576989)

[5.2.4 Comprobación de la invariancia de los resultados con el número de lenguas 28](#_Toc58576990)

[5.3 Países 28](#_Toc58576991)

[5.4 Fuentes 29](#_Toc58576992)

[5.4.1 Conceptos básicos 29](#_Toc58576993)

[5.4.2 Excepciones a los principios básicos 30](#_Toc58576994)

[5.4.3 La cuestión de las fechas 30](#_Toc58576995)

[5.4.4 La cuestión del significado de la transformación país > lengua 30](#_Toc58576996)

[5.4.5 Las limitaciones debidas a las fuentes 31](#_Toc58576997)

[5.4.6 Sesgo potencial de Alexa.com y W3Techs 32](#_Toc58576998)

[5.4.7 La corrección de los sesgos de W3Techs 36](#_Toc58576999)

[5.4.8 Limitaciones/sesgos relacionados con el grado de localidad de las fuentes 38](#_Toc58577000)

[5.4.9 Sobre el principio de ponderación 39](#_Toc58577001)

[6. Conclusiones y perspectivas 41](#_Toc58577002)

[7. Bibliografía 42](#_Toc58577003)

[Anexo I. Lista de micro-indicadores 43](#_Toc58577004)

[Anexo II. Fuentes seleccionadas 45](#_Toc58577005)

[Anexo III: Valores elegidos para L2 46](#_Toc58577006)

[Anexo IV: Sitios totalmente locales 48](#_Toc58577007)

**LISTA DE TABLAS Y FIGURAS**

**TABLAS**

[TABLA 1: Descripción de los indicadores 15](#_Toc58577008)

[TABLA 2: Descripción de los macro-indicadores 16](#_Toc58577009)

[TABLA 3 : Los 3 tipos de ponderaciones utilizados 18](#_Toc58577010)

[TABLA 4 : Indicadores para las 15 primeras lenguas en términos de poder 18](#_Toc58577011)

[TABLA 5 Lenguas ordenadas por capacidad 19](#_Toc58577012)

[TABLA 6 Lenguas ordenadas por porcentaje de personas conectadas 19](#_Toc58577013)

[TABLA 7 Lenguas ordenadas por gradiente 20](#_Toc58577014)

[TABLA 8 : Artículos Wikipedia vs. presencia en la Internet 20](#_Toc58577015)

[TABLA 9 : Diferencias de hipótesis con IWS para L1 + L2 21](#_Toc58577016)

[TABLA 10 : Datos InternetWorldStats (junio de 2016) 22](#_Toc58577017)

[TABLA 11: Comparación con IWS después de reducción a 100% 22](#_Toc58577018)

[TABLA 12 : Simulación con datos de IWS 23](#_Toc58577019)

[TABLA 13 Distribución de lenguas en la Web según Inktomi (2000): 24](#_Toc58577020)

[TABLA 14 Países: afiliación de territorios 29](#_Toc58577021)

[TABLA 15 : Relaciones de tráfico / suscriptores 34](#_Toc58577022)

[TABLA 16 : Distribución de sitios web con fuerte localidad 39](#_Toc58577023)

[TABLA 17 : Simulación para interfaces 40](#_Toc58577024)

[TABLA 18 : Simulación para índice 41](#_Toc58577025)

**ESQUEMAS**

[FIGURA 1 : Indicadores de la presencia de las lenguas en la Internet 8](#_Toc22209582)

# Introducción

Durante el período 1998-2007, el autor y Daniel Prado colaboraron, desde sus respectivas instituciones, Fundación Redes y Desarrollo (FUNREDES) y la Unión Latina, para el diseño de métodos para la medición de las lenguas en la Internet que podrían proporcionar indicadores fiables y reproducibles (véase [6]). Al mismo tiempo, otras iniciativas[[1]](#footnote-1) existían con los mismos objetivos. Desde 2007, cambios profundos en los comportamientos de los motores de búsqueda en la Web, así como en el tamaño del espacio web indexado, han vuelto obsoletos los métodos y eso ha creado un vacío en la producción de indicadores de lenguas en la Internet[[2]](#footnote-2). Entre 2010 y 2012, bajo la dirección del autor, FUNREDES y MAAYA propusieron lanzar un ambicioso proyecto de investigación con el objetivo de llenar ese vacío. La UNESCO, la OIF y la Unión Latina se han unido a este proyecto y un consorcio de prestigiosas instituciones de investigación europeas se formó para tratar de obtener la financiación del Programa Marco de Investigación de la Unión Europea (proyecto DILINET[[3]](#footnote-3)). Sin embargo, este esfuerzo no logro el éxito para asegurar los fondos, a pesar de la persistencia y la calidad de los equipos de investigación participantes (dos intentos se hicieron en 2012 y 2013, en respuesta a llamados a propuesta europeos[[4]](#footnote-4) y un último intento con socios de Qatar aborto en 2014).

Para llenar ese vacío, un nuevo método menos ambicioso, más pragmático y mucho menos costoso, fue propuesto por Daniel Prado, en 2012, y ha abierto una nueva colaboración con el autor, bajo el sombrero institucional de MAAYA y con el apoyo de la OIF. El método se elaboró en base en la observación del comportamiento del francés en una amplia variedad de espacios y aplicaciones de la Internet. Dos de los primeros estudios ayudaron a proporcionar resultados en términos de clasificación del francés en la Internet. El segundo, realizado en 2013, alimentó el capítulo sobre la Internet del informe "*Le français dans le monde*" en 2014 (véase [3]) y fue seguido por un estudio similar del español en la Internet (véase [10]). El último estudio para la OIF, más ambicioso, que inspira este artículo, ha logrado la aplicación de un enfoque estadístico, autorizado por el aumento del número de fuentes, el cual permite obtener resultados en términos de indicadores para un largo conjunto de lenguas, más allá del francés.

El método se basa en la recopilación de información cuantitativa sobre el uso de las lenguas en muchas *aplicaciones y espacios de la Internet*, constituyendo una serie de **micro-indicadores**. La compilación y organización de los datos de las fuentes permite la medición de la presencia de las lenguas en la Internet y poner los resultados en perspectiva para la construcción de una serie de **indicadores** que miden la cuota correspondiente de cada lengua en la Internet, en términos de *usuarios* de la Internet, del *tráfico* generada, del *uso* de los servicios, de *contenidos*, de *interfaces* de software y servicios de traducción en la lengua considerada y en base a una serie de *índices* que evalúan criterios relacionados con la sociedad de la información. Una síntesis de la presencia de las lenguas en la Internet se produce en forma de una serie de **macro-indicadores**, que hacen la síntesis de todos los indicadores lingüísticos:

* *poder* de la lengua en la Internet*,*
* *capacidad* de sus locutores,
* dinamismo de sus locutores conectados, llamado *gradiente*
* *productividad* de sus locutores conectados en términos de producción de contenidos.

Las *agrupaciones temáticas* de micro-indicadores podrían diferenciar aún más el potencial de una lengua dada en lo que respecta a estos temas y así dar algo de luz sobre sus fortalezas y debilidades en la Internet[[5]](#footnote-5).

El marco metodológico pautado es de utilizar la mayor cantidad de fuentes disponibles para cuantificar el papel de las lenguas en la Internet, ya sea directamente cuando cifras relativas a las lenguas estén disponibles, lo que por desgracia es poco frecuente, o indirectamente, usando cifras por país y transformándolas en cifras por lengua con un método que es un pilar del modelo establecido[[6]](#footnote-6).

Esta transformación de cifras relacionadas con los países en cifras relacionadas con lenguas es un elemento original que no tiene precedentes y que abre la posibilidad de manejar el tema de las lenguas en la Internet, en un contexto donde datos relativos a lenguas en la Internet son sumamente escasos y de fiabilidad dudosa (Wikimedia siendo una admirable excepción).

Esta aproximación se apoya en hipótesis implícitas que deben ser explicitadas y evaluadas, de ahí unas precauciones deben tomarse para asegurar la consistencia y fiabilidad de los resultados. La discusión sobre los límites del método y los controles que se han realizados para asegurar la fiabilidad de un método que implica cierta complejidad, tanto en los cálculos realizados como en la comprensión de los conceptos que resultan, por lo tanto, ocupa una parte significativa de este artículo.

# Metodología

## 2.1 Indicadores de lenguas en la Internet

Se van a explorar espacios (por ejemplo, teléfonos inteligentes o e-gobierno) y aplicaciones (por ejemplo, motor de búsqueda de Google o la red social Facebook) de la Internet en búsqueda de fuentes de datos. Cuando fuentes cuantitativas fiables pueden ser identificadas se definen los micro-indicadores correspondientes. Métodos estadísticos (diferentes tipos de promedios ponderados, promedios simples o truncados) son entonces desplegados para construir *indicadores* con la mejor credibilidad. Seis indicadores se identifican que miden la cuota mundial**[[7]](#footnote-7)** de cada lengua de acuerdo con varios elementos característicos de la Internet:

* **Usuarios:** (personas con acceso a la Internet) que se relaciona con la cuota de hablantes de cada lengua que tienen la posibilidad de conectarse a la Internet. Un único micro-indicador (ofrecido por la UIT) aporta las respuestas que se necesitan y servirá como una fuente fundamental para el método establecido.
* **Usos**: Se refiere a las suscripciones a aplicaciones a través de la conexión a la Internet. Once micro-indicadores están involucrados en la construcción de este indicador.
* **Tráfico**: Indicación del tráfico generado por parte de los usuarios hacia las aplicaciones. Trescientos dieciséis micro-indicadores se utilizan para construir este indicador.
* **Índices**: Se refiere a la clasificación de los países en diversos aspectos de la sociedad de la información. Cinco micro-indicadores se utilizan actualmente para construir este indicador.
* **Contenidos**: Se refiere a los contenidos en la Web para cada lengua (por el momento, recoge sobre todo datos de la galaxia Wikimedia). Trece micro-indicadores proporcionan datos para este indicador.
* **Interfaces y traducción**: se refiere a la presencia de las lenguas en las interfaces de aplicaciones o como lengua de traducción. Veintitrés micro-indicadores nutren este indicador.

Cuatro **macro-indicadores** de la presencia de las lenguas en la Internet expresan la síntesis de los indicadores precedentes:

* un macro-indicador del **poder** de la lengua en la Internet, que mide la cuota mundial de la lengua en la Internet, promedio de los seis indicadores anteriores;
* un macro-indicador de la **capacidad** de la lengua en la Internet que se calcula como la división de *poder* con el porcentaje de la población mundial en esa lengua;
* un macro-indicador del **gradiente** de la lengua en la Internet que se calcula como la división del *poder* por el porcentaje de personas conectadas a la Internet en esa lengua.
* un macro-indicador de la **productividad** en términos de creación de **contenidos** que se calcula como la división del porcentaje de contenidos en esa lengua por el porcentaje de usuarios de la Internet en esa misma lengua.

El indicador de *poder* se expresa como cuota mundial. Los otros tres indicadores son sin dimensión y normalizados a 1. El concepto descrito es tan alto como el valor es mayor que 1 y recíprocamente.

El procesamiento estadístico se basará principalmente en la única fuente considerada como a la vez fiable y esencial, que proporciona la UIT: el porcentaje de personas conectadas a la Internet por país. La lectura del artículo mostrará que esta fuente está involucrada en un gran número de operaciones que se realizarán, en particular, la ponderación.

El siguiente diagrama muestra todos los indicadores que se procesan para cada lengua.

FIGURA 1 : Indicadores de la presencia de las lenguas en la Internet



Todos los *micro-indicadores* se presentan en el *Anexo 1: Lista de micro-indicadores*. Los **indicadores** y **macro-indicadores** se detallan en una tabla en la *sección 4.5.3 Tratamientos para resultados*.

## 2.2 Micro-indicadores por lengua o país

Un total de 369 micro-indicadores representan la fuente de datos para esta edición del estudio. Algunos (36) informan directamente sobre las lenguas en la Internet, los otros (333) ofrecen datos por país.

### 2.2.1 Por lengua

Los micro-indicadores por lengua se refieren a los *contenidos* (13) e *interfaces* (23) y son 36 en total para esta edición.

Las fuentes de *contenidos* se expresan en términos de unidades por lengua (por ejemplo, número de artículos de Wikipedia por lengua) o porcentaje mundial (por ejemplo, el porcentaje mundial de los libros vendidos por parte de Amazon[[8]](#footnote-8) por lengua). En el primer caso, el valor se convierte en porcentaje mundial dividiendo por el total. El indicador de *contenidos* se calcula como **la media truncada en 20%**[[9]](#footnote-9) de los 13 micro-indicadores de *contenidos*: el número de libros por lengua en Amazon, los datos de W3Techs[[10]](#footnote-10) y 11 micro-indicadores producidos por Wikimedia).

Los micro-indicadores **de interfaces y traducción** se miden mediante un número binario que expresa la ausencia (0) o existencia (1) de la lengua en la interfaz de la aplicación o la presencia o ausencia de la lengua en la aplicación de traducción. Hay 23 aplicaciones analizadas. El indicador *interfaz* muestra inicialmente el porcentaje de presencia de la lengua en todas las aplicaciones, luego este valor se transforma en porcentaje mundial por el método de ponderación con el porcentaje de usuarios de la Internet en cada lengua.

### 2.2.2 Por país

Más del 90% de los micro-indicadores de esta edición del estudio fue obtenido de fuentes que proporcionan información por país, las cuales pueden ofrecer datos en distintas modalidades:

* en cantidades (por ejemplo, el número de usuarios de la Internet por cada país);
* en porcentajes nacionales (por ejemplo, el porcentaje de consultas a Facebook por país, o el porcentaje de tráfico de la Internet móvil realizado en cada país);
* en porcentajes mundiales (por ejemplo, la distribución del tráfico mundial al sitio Facebook.com por país);
* en forma de notación en una escala predeterminada (en el caso de índices que proporcionan clasificaciones para los países de acuerdo con criterios, tales como WebIndex[[11]](#footnote-11) proporcionando índice de país sobre la sociedad de la información tomando valores de 0 a 1 o de 0 a 100, según el caso).

A fin de lograr un proceso homogéneo, todos los micro-indicadores de diversas fuentes, están transformados en **porcentaje mundial** dentro del modelo de trabajo.

La transformación de un dato por país en dato por lengua se realiza ponderando los datos por país con el número de hablantes de cada lengua en cada país.

Las fuentes pueden proporcionar valores para el presente (2016), o más a menudo para una fecha anterior, a veces incluso proyecciones para el futuro.

Las fuentes de datos disponibles rara vez cubren todos los países del mundo[[12]](#footnote-12) y muy a menudo sólo una minoría. Por lo tanto, la transformación "datos por país" a "datos por lengua" puede requerir algunas técnicas de extrapolación para complementar razonablemente a partir de los datos existentes en la fuente, los datos de los países no informados (véase el *capítulo 4.5.1 Extrapolación*).

## 2.3 Fuentes de información

Para reunir la cantidad mayor de datos se han identificado docenas de fuentes diferentes de información (estadísticas, encuestas, evaluaciones, índices, bases de datos, etc.) y luego evaluado para cada fuente su relevancia, su confianza y su coherencia con el método. Algunas fuentes que proporcionan resultados para demasiado pocos países han sido rechazadas y otras donde los porcentajes por país pierden sentido cuando se expresan por lengua fueron también excluidas.

Es relativamente fácil encontrar estadísticas por país en relación con el acceso a la Internet y las nuevas tecnologías (gracias a la UIT, las Naciones Unidas o el Banco Mundial); sin embargo, es mucho más difícil encontrar fuentes que proporcionan datos por lengua, fuera del universo Wikimedia y de la esfera del mundo abierto[[13]](#footnote-13) que se caracterizan los dos por una preocupación genuina por la transparencia, tanto para los datos como para las metodologías utilizadas para su producción.

Las compañías en línea, los propietarios de las redes sociales, de los motores de búsqueda, o de las aplicaciones de la Internet más reconocidas, son generalmente bastante secretos sobre sus estadísticas. Los únicos que son capaces de proporcionar cifras sobre las aplicaciones (con un fuerte enfoque en las redes sociales y todo lo que tiene que ver con el marketing en el mundial digital) son empresas de marketing en la Internet. La mayoría de las veces esas empresas piden un precio importante para acceder a sus datos estadísticos. Estas empresas a menudo tienen relaciones especiales con las principales aplicaciones para los cuales pueden ser amplificadores de audiencia o tienen los medios financieros para desarrollar métodos poderosos aunque a menudo aproximados (como es el caso de Alexa[[14]](#footnote-14), ahora una subsidiaria de Amazon, o de W3Techs).

Para encontrar fuentes, hay dos opciones:

1) recoger algo de información libre, aunque muy parcial que queda filtrada por una de estas empresas como una forma de atraer a los clientes a pagar la mayor parte de la información;

2) pagar para obtener los datos.

El primer enfoque se aplicó sistemáticamente con el despliegue de una intensa actividad de búsqueda en la Internet. Sin embargo, hay que tener en cuenta que las metodologías generalmente no son transparentes y que estas cifras son raramente sostenidas, lo que es un inconveniente para quienes quieren mantener una observación perenne. Los recursos limitados del estudio condujeron a la selección de Alexa y Statista[[15]](#footnote-15) como fuentes pagadas, en razón de su mejor relación calidad/precio, y esas dos fuentes alimentaron las estadísticas para la mayoría de las aplicaciones o espacios deseados. La elección de Alexa permitió cubrir ampliamente[[16]](#footnote-16) pero es necesario saber que las cifras de tráfico por país a cada aplicación, que facilita Alexa, presentan un sesgo significativo (véase el *capítulo 7.4.6 Sesgo potencial de W3Techs y Alexa.com*). La lista completa de las fuentes seleccionadas se detalla en el *Anexo II. Fuentes seleccionadas*.

### 2.4.1 El caso de la lengua materna (L1)

Para L1, este trabajo se basa en una contribución significativa de Daniel Prado, que se hizo cargo de los datos de la fuente del Proyecto Joshua[[17]](#footnote-17) (el cual proporciona la distribución de más de 7.500 lenguas por país) y ha requerido, por un lado, de ajustar las figuras, siempre que sea necesario, usando otras fuentes diversas, y, por el otro lado, adaptar las cifras a la división por país elegida. Ethnologue hubiese sido una fuente de mejor calidad a usar sin modificar, pero su precio estaba por encima del presupuesto previsto.

La matriz LOC1 cumple con la definición de locutores, para todas las lenguas seleccionadas y todos los países seleccionados:

LOC1 (i, j) = Número de hablantes L1 de la lengua i en el país j.

La fuente proporciona cifras de 7500 lenguas; quedaba por hacer una elección en la lista de lenguas que serán procesadas para el estudio. Hoy en día, el número estimado de lenguas que tienen una presencia en la Internet es alrededor de 500. Una posibilidad sería de acercarse lo más posible a esas lenguas. Otra posibilidad sería de seleccionar las lenguas para las cuales Wikipedia ofrece estadísticas (cerca de 300). Después de mucha reflexión y varias pruebas la elección final fue la lista de las 140 lenguas con más de 5 millones de hablantes (L1). Las cifras para las otras lenguas se acumularon en una categoría denominada " lenguas restantes" que sigue contabilizada en la última línea de la matriz LOC1.

La razón de esta elección es la **prudencia** frente a los sesgos potenciales ya que la hipótesis implícita del método de ponderación para transformar los datos por país en datos por lengua no procura la suficiente confianza en los resultados cuando el número de hablantes es bajo (para más detalles véase *5.1 Limitaciones específicas con el método*).

### 2.4.2 El caso de la segunda lengua (L2)

La primera prioridad es la de tener en cuenta de manera coherente lo que implica el multilingüismo. De hecho, cuando se menciona una segundo lengua, entonces se introduce implícitamente el factor siguiente: las personas contabilizadas en L2 también tienen una primera lengua (L1) y por lo tanto L1+L2 incluye las mismas personas más de una vez[[18]](#footnote-18). La evidencia dice entonces que, si uno trata de medir el porcentaje mundial de personas hablantes de una lengua, ya sea como lengua materna o como segunda lengua, esta cifra debe necesariamente basarse en el total de hablantes por lengua en el mundo cifra mucho más alta que la población mundial[[19]](#footnote-19). Esta evidencia es ignorada a menudo en muchas estadísticas y eso representa una fuente potencial de error serio. En el escenario que se adoptó, con base en cifras L2 de Ethnologue[[20]](#footnote-20), y con un suplemento promedio global L2 de 18% otorgado por precaución (y ajustado para crear cifras redondeadas) a las lenguas que faltan de la lista de 140 lenguas seleccionadas, las cuotas (ya no son estrictamente porcentajes) se calculan sobre la base de una suma que sería igual al 125% de la población mundial (el 25% representando el grado mundial de multilingüismo).

También es importante entender que esta noción es igualmente aplicable a todos los conceptos procesados: *usuarios, tráfico, usos, contenidos, interfaces e índices.* De hecho, el multilingüismo se aplicará a todos los elementos (los sitios web se pueden hacer en varias lenguas, lo mismo para el flujo de mensajes de correo electrónico o chats e intercambios en las redes sociales) y entonces si añadimos las cifras para todas las lenguas del mundo, los totales calculados en relación con la población mundial serán superior al 100%. En realidad, es probable que el grado de multilingüismo de los sitios web sea superior al de sus visitantes, sin embargo, la aproximación de considerarlos equivalentes es aceptable.

El método ideal para tratar el caso de L2 en el modelo, obviamente, sería de hacer lo mismo que para la L1 y producir una matriz como:

LOC2 (i, j) = número de locutores L1+L2 de la lengua i en el país j.

Entonces sería suficiente para aplicar sin ningún cambio el resto del método (teniendo en cuenta el multilingüismo para los totales) para obtener los resultados esperados con una exactitud muy apreciable. Desafortunadamente, los datos para construir tal matriz no existen y la existencia de fuertes diferencias entre las fuentes demo-lingüísticas para L2 hace que la realización de dicha matriz se hace muy difícil en el estado actual de los conocimientos sobre el tema. Por ello se propone otra manera de tratar el problema, simplificándolo.

El principio del método elegido es simple y consiste, para cada lengua, de obtener un número que representa el incremento que se aplicará a las cantidades de L1 para obtener el valor L1+L2. Este número será igual a 1 para la lengua i si L2 (i) = 0. Para esta edición, se tomará como fuente Ethnologue para L2, manteniendo los valores existentes para L1. Los valores de L2 por lengua distintos de cero se presentan en el *Anexo III : Valores elegidos para L2* .

La tasa global de aumento resultante para los indicadores (1,25 con las hipótesis de trabajo[[21]](#footnote-21)) es el resultado de la siguiente operación de ponderación:

j=L

Tg = ∑ L1(j) x T12(j)

j=1

donde:

L es el número total de lenguas,

L1 (j) el número de hablantes L1 para la lengua j

T12(j) la tasa de aumento de L1 a L1+L2 para la lengua j.

La siguiente ecuación muestra cómo se calcula el valor de los micro-indicadores para L1+L2 a partir del valor de los mismos para L1:

j= L

ML1+L2 (i) = Tg x ML1(i) / ∑ ML1(j) x T12(j)

j=1

La división opera la ponderación cuadrado a 100% y el producto por Tg es allí sólo para normalizar los resultados en un porcentaje igual a Tg.

Este método es obviamente menos preciso que una solución que pudiera trabajar a nivel de cada país con la matriz ideal LOC2; sus limitaciones y sesgos potenciales, y una variante potencial simple del método para que sea más preciso, se discuten en *7.2.2 El caso de L2*.

El método L1+L2 se aplica a todos los indicadores excepto índices, contenidos e interfaces y traducción los cuales por naturaleza tienen ya significado directamente para L1 + L2.

## 2.5 Proceso

Los indicadores expresados ​​directamente en términos de lengua (contenidos e interfaces) no implican ninguna dificultad para el proceso; se prestará entonces atención a los indicadores expresadas por país los cuales merecen un tratamiento particular. El principio del cálculo para convertir cifras expresadas en porcentaje mundial por país en cifras expresadas en porcentaje mundial por lengua, es el *producto matricial*. La primera matriz es LOC que tiene en el *axis x* las lenguas seleccionadas y en el *axis y* los países seleccionados. La segunda matriz es, el vector MPn que contiene los valores fuentes por país del micro-indicador n. El resultado del producto ofrece el valor del micro-indicador expresado en porcentaje por lengua (MLn).

LOC (i, j) = El número de hablantes de lengua i en el país j.

MPn (j) = El valor medido para el micro-indicador n en el país j.

j = P

MLn (i) = Σ LOC (i, j) x MPn (j) donde P es el número total de países.

j = 1

El producto matriz MLn = LOC +. x MPn en notación APL[[22]](#footnote-22) o = SUMAPRODUCTO (LOC; MPn) en notación Excel, es una operación de ponderación que produce un nuevo vector, esta vez del tamaño del número de lenguas, y para el cual MLn (i) = el valor del micro-indicador n deducido para la lengua i, a partir de los datos de demo-lingüística de lenguas por país (LOC) y de los valores medidos por país para el micro-indicador n (MPn). El corazón del método es en realidad la ponderación de los valores medidos para este micro-indicador en cada país con la presencia de la lengua en cada país. La validez de esta ponderación y sus posibles sesgos se discuten en el capítulo *7.1Limitaciones específicas para el método*.

Los totales del MLn son los mismos que los de MPn pero esta vez la distribución se hace por lengua en lugar de por país.

El principal inconveniente de este método de cálculo es que no proporcionara resultados fiables si la fuente de datos no informa todos los países debido a que el producto de matriz transformara esta ausencia por un valor cero que conlleva a resultados inaceptables para las lenguas con fuerte presencia en países en los que la fuente no informa.

### 2.5.1 La extrapolación

La manera de superar esta situación es el de estimar, de la mejor manera posible, los valores ausentes para los países no informados. Es factible y de una manera razonable, en la mayoría de los casos, la extrapolación de los datos que faltan a partir de la información existente y otra información disponible en otros lugares. No se requiere una precisión absoluta, sino un método sencillo donde las desviaciones a la realidad son reales, pero de impacto limitado en los resultados del procesamiento estadístico.

Se adoptaron dos métodos diferentes para resolver todos los casos:

**a) La extrapolación en proporción al porcentaje de personas conectadas por país.**

Este método se aplica cuando es razonable considerar que los valores son naturalmente proporcionales al porcentaje mundial de personas conectadas (este es el caso, por ejemplo, con el tráfico de sitios web). Si la fuente de datos se expresa en cantidades en vez de porcentaje se debe primero extrapolar el total mundial; si se expresa en porcentaje mundial, no será necesario. El porcentaje total o restante se distribuye entre los países no informados en proporción de sus respectivos pesos en términos de conexión a la Internet.

**b) Método de cuartiles**

En este método, los países no documentados tienen los valores complementados a partir de un cuartil de las figuras de origen en función de su porcentaje de personas conectadas. Después de varias pruebas, se determinó como sigue la asignación de los cuartiles:

* Si el país tiene menos de 15% de su población conectada: la nota más baja
* Si el país tiene más de 15% pero inferior al 35% de su población conectada: primer cuartil
* Si el país tiene más del 35% pero inferior al 65% de la población conectada: la mediana
* Si el país tiene más del 65% pero inferior al 85% de su población conectada: tercer cuartil
* Si el país tiene más del 85% de su población conectada: la nota más alta.

Los micro-indicadores para los cuales no existe un método de extrapolación que parezca obvio son los mismos para los cuales no aparece claramente el significado de la transformación de las cifras por país en cifras por lengua y por lo tanto están excluidos (véase la cuestión del significado de la transformación de país a lengua en la sección *5.4.4 La cuestión del significado de la transformación país > lengua*).

### 2.5.2 Proceso para los indicadores

Aparte el indicador de la UIT que proporciona porcentajes de personas conectadas a la Internet, por país, tal vez otros indicadores de tecnología, y al límite algunas estadísticas de Wikimedia[[23]](#footnote-23), ningún indicador individual refleja una verdad válida para todas las lenguas. De hecho, el ideal de un grado de globalización absoluto, que permita considerar que el comportamiento de los usos son los mismos para un indicador dado, independientemente del país o de la lengua, difícilmente será verificado.

Sólo un análisis estadístico de un conjunto importante de micro-indicadores, seleccionados con criterios los más cuidadosos posibles, diversos y globalizados permite, en ausencia de una medida fiable del contenido por lengua[[24]](#footnote-24), de cuantificar la presencia de las lenguas en la Internet, bajo la condición de analizar y comprender los posibles sesgos en los resultados que pueden derivarse de las hipótesis simplificadoras, de las fuentes y del mismo procesamiento estadístico.

El modelo también puede ayudar a entender, por diferenciación temática, donde una lengua tiene mayor o menor presencia; en la versión actual, sólo el indicador de tráfico que comprende 316 micro-indicadores permite aventurarse al análisis temático.

Todos los indicadores de la presencia de las lenguas en la Internet se presentan en las siguientes dos tablas. La primera tabla muestra los *indicadores* y cómo se calculan. La segunda tabla muestra los *macro-indicadores* y cómo se calculan. Todos los indicadores se expresan en términos de *cuota +mundia*l, basado en la población total de hablantes (población mundial multiplicado por el factor de multilingüismo, en la versión actual, el 125% de la población mundial).

TABLA 1: Descripción de los indicadores

|  |  |  |  |
| --- | --- | --- | --- |
| **INDICADOR** | **DEFINICIÓN** | **TÉCNICA** | **FIABILIDAD** |
| **A: USUARIOS** | Mono indicador derivado de las cifras de la UIT de % mundial de personas conectadas por país. | ponderación  país -> lengua  sin extrapolación | Muy fuerte  Sesgo muy marginal |
| **B: USOS** | Incluye 11 indicadores micro  - Las líneas telefónicas  - Mercado e.Commerce  - Descargas OpenOffice  - Uso de redes sociales + proyección 2021  - Uso de Netflix + proyección 2020  - Suscriptores Facebook, Twitter, YouTube, LinkedIn | ponderación  país-> lengua  extrapolado proporcionalmente  promedio truncado en 20% | Fuerte.  Bajo sesgo.  Pero tendría que ser ampliado para dar más sentido a la media del número de micro-indicadores. |
| **C: TRÁFICO** | Alexa mide el tráfico de una selección de 316 sitios web (ver detalle en *Anexo I*). | ponderación  país-> lengua  extrapolado proporcionalmente  promedio truncado a 20% | Relativamente buena  Pero fuerte sesgo pro occidental de Alexa confirmado por comparación tráfico vs. suscriptos |
| **D: INDICES**  **(sociedad de la información)** | Incluye 5 índices de WebIndex de los siguientes criterios:  - E.gobierno  - Acceso universal  - E.participación  - Infraestructura general  - Índice general | ponderación  país-> lengua  extrapolado por el método de cuartil.  Transformación en datos de porcentaje mundial por ponderación con datos UIT  Promedio | Bueno (son datos subjetivos cuantificados por un organismo competente).  Esta categoría debe ser extendido. |
| **E: CONTENIDOS**  **(Wikimedia y libros)** | Incluye 13 indicadores micro  - Número de libros en Amazon  - Las medidas contenidas por W3Techs lengua  - 11 indicadores lingüísticos de Wikimedia: artículos, usuarios o editores de Wikipedia, wikibook, Wikibooks, wikisource, Wikiversity, wiktionnnaire. | El uso directo de las cifras por lengua.  Promedio truncado a 20% de los micro-indicadores | Muy fuerte para Wikimedia y Amazon.  Pero muy baja presencia de algunas de las principales lenguas asiáticos.  tendría que ser extendido el número de micro-indicadores para dar más fuerza a la media |
| **F: INTERFACES**  **(e lenguas de traducción)** | Incluye 23 micro-indicadores binarios  - 12 lenguas de interfaz de Cortana, DuckDuckGo, Facebook, GoogleSearch, GoogleScholar, Skype, Telegrama, Wikibook, Wikipedia, Wikisource, Wikiquote, Wikiversity  - 1 lengua de contenido para Dmoz  - 10 lenguas de traducción de Bing, Dictionnary, Duolingo, FreeTranslator, GoogleTranslate, mensajería instantánea, en línea, Diccionario, SDL, Systran | % presencia en todos los 23 micro- indicadores.  % mundial ponderando con datosUIT. | Perfecto. |

TABLA 2: Descripción de los macro-indicadores

|  |  |  |
| --- | --- | --- |
|  | **DEFINICIÓN** | **TÉCNICA** |
| ***PODER*** | Mide la participación global de la lengua en la Internet | Promedio de los 6 indicadores |
| ***CAPACIDAD*** | Mide la fuerza de la lengua en la Internet, independientemente de su número de hablantes. | Relación entre poder y % mundial de hablantes |
| ***GRADIENTE*** | Mide la fuerza de los locutores conectados independientemente de su número. | Relación entre poder y % mundial de locutores conectados |
| ***PRODUCTIVIDAD*** | Mide la propensión de los locutores conectados a producir contenidos en su lengua.[[25]](#footnote-25) | Relación entre % contenidos y.% de locutores conectados |

Los primeros 3 macro-indicadores están unidas por las siguientes ecuaciones:

**Poder (**lengua**)** = capacidad (lengua) x % población mundial de locutores (lengua)

**Poder** (lengua) = gradiente (lengua) x % población mundial conectada a la Internet (lengua)

Hay que notar que:

* A- Los datos de la UIT sobre usuarios (personas conectadas a la Internet) son un indicador clave del modelo y "pesan" muy fuerte en los resultados, directamente en el promedio para calcular el *poder* que le da un peso igual de importante que los 300 micro-indicadores de tráfico de Alexa, e indirectamente en las técnicas de extrapolación y de ponderación, que afectan transversalmente todos los cálculos. Toda la estructura de procesamiento se basa en estos datos considerados muy fiables.
* C- Para las cifras de tráfico obtenidos de Alexa, la validez radica en la gran cantidad de sitios medidos, aunque los resultados tiene una fuerte desviación típica provocada por la selección de sitios muy marcados en una población local (ver sección 7*.4.7 Limitaciones / sesgo relacionado con el grado de localidad de las fuentes*). La técnica de promedio truncado permite eliminar valores extremos y dar más confianza en el resultado. Además, estos datos muestran un fuerte sesgo a favor de lenguas occidentales (principalmente inglés y francés) y sobre todo a expensas de las lenguas asiáticos, que se discute en la sección *7.4.6 Sesgo potencial de W3Techs y Alexa.com*.
* D- Los índices abren una perspectiva muy interesante sobre el alcance de elementos contextuales creados por organizaciones internacionales u organizaciones no gubernamentales. Por el momento, WebIndex[[26]](#footnote-26) ha sido incluido parcialmente; es un proyecto de la Fundación World Wide Web que mantiene una serie de índices que se basan en el trabajo serio de los servicios de las Naciones Unidas, el Banco Mundial y otras instituciones. Los índices utilizados en este estudio incluyen los servicios de gobierno electrónico, la participación ciudadana a través de la Internet, el grado de consecución del acceso universal y la infraestructura para la sociedad de la información, así como un índice general que incorpora todos los factores. Es importante saber que WebIndex ofrece otros indicadores más allá de los cinco que se han utilizado y que otras organizaciones ya ofrecen o van a ofrecer en el futuro índices que sería útil incluir en una versión futura[[27]](#footnote-27).
* E: Este indicador se denomina **contenidos**, pero en este momento está muy centrado en la galaxia Wikimedia la cual ofrece la enorme ventaja de proporcionar estadísticas detalladas y creíbles para la presencia de las lenguas en sus diferentes actividades. Esta complementado por los datos de W3Techs (poco creíbles) y los libros en Amazon, lo que amplía apenas la perspectiva. Este indicador, que abastece el secreto mejor guardado de la Internet, la distribución de contenidos por lengua, merecería claramente una ampliación del número de fuentes... siempre que las fuentes sean creíbles, lo que se ha convertido en un terrible reto. Hay que tomar en cuenta, sin embargo, que las principales lenguas asiáticas están muy poco representadas en Wikimedia[[28]](#footnote-28). También es importante entender que por naturaleza (indicador directo de lenguas) este indicador penalizará por completo (con notas nulas) lenguas fuera del universo Wikimedia, el cual, aunque es el grupo más preocupado de muy lejos por el asunto lingüístico, se limita a algunas 300 lenguas.
* F: La penalización de lenguas débilmente conectadas (pero también lenguas bien conectadas, pero no tomadas en cuenta en traducciones e interfaces) será aún más fuerte que para E. Este indicador va hacer retroceder aún más en la clasificación las lenguas con menor puntuación, aunque refleja una realidad que no puede ser ignorada.

### 2.5.3 Proceso para lograr resultados diferenciados por temáticas

El único indicador con un número suficiente de elementos para permitir un análisis más detallado por tema es el *tráfico* con sus 316 elementos. A cada sitio web se asocia un tema que lo representa y con la media truncada a 20% se calcula la diferencia positiva o negativa de cada tema en comparación con el promedio general. Sin embargo, estos resultados deben tomarse con precaución, especialmente si el número de sitios para un tema determinado es bajo. Hasta ahora, este análisis se ha realizado sólo por el francés que muestra su fuerza en contenidos de investigación científica, en MOOCs[[29]](#footnote-29), en contenidos profesionales, en redes sociales de música, en juegos y en motores de búsqueda.

### 2.5.4 Tipos de ponderaciones utilizadas

El método estadístico utilizado se basa principalmente en los métodos de ponderación y extrapolación; es útil para identificar y sintetizar los diferentes tipos de ponderación utilizados en el proceso y explicitar las hipótesis simplificadoras que afectan la validez de los resultados obtenidos (la discusión detallada de las implicaciones y posibles sesgos resultantes de las hipótesis se exponen en el capítulo 7*.1 Limitaciones específicas para el método*).

La ponderación principal es la que transforma los porcentajes por país en porcentajes por lengua: los valores de un micro-indicador expresado en términos de países se ponderan en relación con la distribución de las lenguas en los países con el fin de proporcionar la distribución por lengua. La suposición implícita que apoya la validez de este proceso es que *el concepto medido por el micro-indicador se expresa de la misma manera para todas las lenguas del país*.

La ponderación que permite convertir figuras L1 en figuras L1+L2 se hace en base a la tasa de aumento de L1 a L1+L2 para cada lengua. El supuesto simplificador de este método es *que los porcentajes de conexión a la Internet son los mismos para los hablantes L2 como para los hablantes L1, independientemente del país de residencia.*

La ponderación que permite convertir los valores expresados de un porcentaje de un criterio dado (interfaces, índices) en un porcentaje mundial se hace con respecto a los porcentajes de personas conectadas por lengua. El principio detrás de este método es que los porcentajes mundial obtenidos para el criterio son una desviación del porcentaje mundial de locutores conectados en función de la distribución por lengua de este micro-indicador. Una aproximación más intuitiva del significado de esta ponderación se proporciona en el capítulo *7.4.8 Sobre el principio de ponderación*.

TABLA 3 : Los 3 tipos de ponderaciones utilizados

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Demo-lingüística** | **Segundas lenguas** | **Internautas por lengua** |
| TIPO | País ---> Lengua | L1 ---> L1 + L2 | % criterio ->% mundial |
| INSUMO | Datos por país | Resultados L1 | Datos en% criterio |
| RESULTADO | Datos por lengua | Resultados L1 + L2 | Datos en% mundial |
| DATOS DE  PONDERACIÓN | Matriz LOC | Vector L1 + L2 / L1 por lengua | Datos UIT |
| ALCANCE | Todas las fuentes por país | Indicadores de usuarios, tráfico y uso | Índices e interfaces. |
| HIPOTESIS  IMPLICITA | Independencia en relación con lenguas en el país | Independencia de las tasas de conexión a la Internet en función de los países | Modulación de las tasas de conexión a la Internet de acuerdo con el criterio |

# Resultados

La siguiente tabla muestra todos los resultados acerca de las 15 lenguas más "*poderosas*" en la Internet. Los resultados, que son expresados en términos de cuota mundial, se muestran separados entre primera lengua (L1) y primera más segunda lengua (L1+L2), para cada indicador (usuarios, tráfico, uso, contenidos, índices e interfaces ) y para el macro-indicador "poder".

TABLA 4 : Indicadores para las 15 primeras lenguas en términos de poder

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **USUARIOS** | | **TRÁFICO** | | **USOS** | | **CONTENIDOS** | **INTERFAZ** | **ÍNDICES** | **PODER** |
|  | **L1** | **L1 + L2** | **L1** | **L1 + L2** | **L1** | **L1 + L2** | **L1 + L2** | **L1 + L2** | **L1 + L2** | **L1 + L2** |
| Inglés | 0,085 | 0,222 | 0,283 | 0,568 | 0,192 | 0,437 | 0,499 | 0,309 | 0,305 | **0,390** |
| Chino | 0,218 | 0,205 | 0,081 | 0,062 | 0,125 | 0,108 | 0,030 | 0,248 | 0,194 | **0,141** |
| Español | 0,080 | 0,091 | 0,109 | 0,094 | 0,095 | 0,095 | 0,053 | 0,121 | 0,090 | **0,091** |
| Francés | 0,020 | 0,056 | 0,036 | 0,080 | 0,029 | 0,074 | 0,093 | 0,074 | 0,073 | **0,075** |
| ruso | 0,028 | 0,050 | 0,015 | 0,022 | 0,025 | 0,041 | 0,054 | 0,066 | 0,051 | **0,047** |
| Alemán | 0,019 | 0,031 | 0,028 | 0,037 | 0,026 | 0,039 | 0,084 | 0,042 | 0,039 | **0,046** |
| Portugués | 0,040 | 0,040 | 0,026 | 0,020 | 0,048 | 0,043 | 0,033 | 0,053 | 0,039 | **0,038** |
| Japonés | 0,036 | 0,034 | 0,031 | 0,024 | 0,038 | 0,032 | 0,030 | 0,043 | 0,046 | **0,035** |
| Árabe | 0,044 | 0,042 | 0,042 | 0,031 | 0,032 | 0,028 | 0,013 | 0,048 | 0,032 | **0,032** |
| Hindi | 0,031 | 0,039 | 0,020 | 0,020 | 0,024 | 0,027 | 0,003 | 0,035 | 0,033 | **0,026** |
| Malayo | 0,006 | 0,026 | 0,005 | 0,015 | 0,006 | 0,022 | 0,013 | 0,027 | 0,021 | **0,021** |
| Italiano | 0,009 | 0,009 | 0,014 | 0,011 | 0,011 | 0,010 | 0,041 | 0,012 | 0,010 | **0,016** |
| Coreano | 0,015 | 0,014 | 0,012 | 0,009 | 0,011 | 0,009 | 0,009 | 0,016 | 0,016 | **0,012** |
| Polaco | 0,009 | 0,008 | 0,010 | 0,008 | 0,009 | 0,008 | 0,030 | 0,009 | 0,008 | **0,012** |
| Urdu | 0,009 | 0,017 | 0,009 | 0,013 | 0,008 | 0,013 | 0,001 | 0,011 | 0,014 | **0,011** |
| **RESTO** | **0,351** | **0,368** | **0,279** | **0,236** | **0,321** | **0,265** | **0,264** | **0,135** | **0,279** | **0,258** |
| **TOTAL** | **1,000** | **1,250** | **1,000** | **1,250** | **1,000** | **1,250** | **1,250** | **1,250** | **1,250** | **1,250** |

La línea RESTO representa los resultados para el conjunto completo de todas las lenguas del mundo, excepto las 15 lenguas que aparecen en la tabla.

Debe quedar claro que la clasificación en términos de *poder* favorece las lenguas que tienen el mayor número de hablantes. Para clasificaciones que miden indicadores de lengua que son independientes del número de hablantes, las siguientes tablas ofrecen resultados interesantes.

TABLA 5 Lenguas ordenadas por capacidad

|  |  |  |
| --- | --- | --- |
|  | **Capacidad** | **Clasificación**  **Poder** |
| Hebreo | 5,40 | 35 |
| Finlandés | 5,40 | 38 |
| Holandés | 4,81 | 19 |
| Sueco | 4,46 | 28 |
| Inglés | 3,72 | 1 |
| Alemán | 3,40 | 6 |
| Danés | 3,30 | 49 |
| Italiano | 3,16 | 12 |
| Checo | 3,13 | 27 |
| Francés | 2,96 | 4 |
| Serbocroata | 2,83 | 22 |
| |  | | --- | | Húngaro | | 2,83 | 32 |
| Polaco | 2,64 | 14 |
| Japonés | 2,54 | 8 |
| Griego | 2,40 | 37 |

Es evidente que hay una correlación fuerte y consistente entre la capacidad y el porcentaje de personas conectadas a la Internet; Sin embargo, la correlación no es total como se muestra en la siguiente clasificación ordenada por el porcentaje de personas conectadas.

TABLA 6 Lenguas ordenadas por porcentaje de personas conectadas

|  |  |  |
| --- | --- | --- |
|  | **% Personas**  **conectadas** | **Clasificación**  **Poder** |
| Danés | 95.67% | 49 |
| Finlandés | 92.30% | 38 |
| Holandés | 92.27% | 19 |
| Sueco | 90.54% | 28 |
| Japonés | 90.43% | 8 |
| Alemán | 86.43% | 6 |
| Alemánico | 86,41% | 51 |
| Bávaro | 84.42% | 39 |
| Flamenco | 83.60% | 59 |
| Checo | 81,17% | 27 |
| Francés | 81,09% | 4 |
| Inglés | 78.05% | 1 |
| Eslovaco | 77.38% | 62 |
| Hebreo | 76.05% | 35 |
| Húngaro | 71.16% | 32 |

Y, finalmente, la última tabla, ordenada por *gradiente*, destaca el dinamismo de las personas conectadas.

TABLA 7 Lenguas ordenadas por gradiente

|  |  |  |
| --- | --- | --- |
|  | **Gradiente** | **Clasificación**  **Poder** |
| Hebreo | 2,62 | 35 |
| Finlandés | 2,16 | 38 |
| Holandés | 1,93 | 19 |
| Sueco | 1,81 | 28 |
| Inglés | 1,76 | 1 |
| Italiano | 1,73 | 12 |
| Serbocroata | 1,54 | 22 |
| Húngaro | 1,47 | 32 |
| Alemán | 1,45 | 6 |
| Checo | 1,42 | 27 |
| Polaco | 1,41 | 14 |
| Francés | 1,35 | 4 |
| Griego | 1,30 | 37 |
| Danés | 1,27 | 49 |
| Armenio | 1,20 | 60 |

Es de notar sin sorpresa que las lenguas nacionales de los países reconocidos por sus políticas proactivas para la sociedad de la información aparecen en las primeras posiciones de los últimos 3 macro-indicadores. Es notable que varias lenguas clasifican antes del inglés, a pesar de su ventaja estratégica en la Internet, por ser la lengua preferida para contenidos multilingües y la creencia de muchos que es la lengua franca de la Internet.

# Análisis de los resultados

Es importante saber que dos de los seis indicadores (el tráfico, medido por Alexa, y los contenidos que se basan principalmente en estadísticas de Wikimedia) pueden tener, a su manera, un sesgo muy significativa a favor del inglés, y en contra, primero, de las lenguas no occidentales (especialmente chino, hindi, urdu, bengalí, y en menor medida árabe), y segundo, aunque en menor medida, de algunas otras lenguas occidentales, en el caso de Alexa (portugués y, en menor medida española). Los diversos sesgos posibles deben ser analizados.

## 4.1 Acerca de Wikipedia

Las estadísticas de Wikimedia son impecables; sin embargo, se debe entender que, a pesar de ser una de las aplicaciones más global de la Internet, las cifras de algunas lenguas asiáticas están muy por debajo de sus presencias relativas en la Internet. La siguiente tabla compara las proporciones entre artículos de Wikipedia y usuarios de la Internet; aparecen enormes variaciones con valores anormalmente bajos para lenguas asiáticas.

TABLA 8 : Artículos Wikipedia vs. presencia en la Internet

|  |  |  |  |
| --- | --- | --- | --- |
| **LENGUA** | **% PERSONAS CONECTADAS** | **% ARTÍCULOS WIKIPEDIA** | **RELACIÓN** |
| Sueco | 0,2% | 8,8% | 44 |
| Holandés | 0,4% | 4,5% | 11 |
| Francés | 5,6% | 4,3% | 0,8 |
| Inglés | 22,6% | 12,6% | 0,6 |
| Chino | 20,9% | 2,8% | 0,13 |
| Urdu | 2,0% | 0,3% | 0,15 |
| Hindi | 3,9% | 0,3% | 0,08 |
| Bengalí | 1,3% | 0,1% | 0,08 |

## Comparación de los resultados con los de InternetWorldStats (IWS)

Es interesante comparar los resultados de este estudio con los datos ofrecidos por InternetWorldStats[[30]](#footnote-30), la fuente más citada desde varios años para datos de usuarios de la Internet por lengua (aunque limitada a sólo las 10 lenguas con el mayor número de usuarios). Utilizando en principio la misma fuente para las personas conectadas (UIT), los dos métodos deberían, a partir de los mismos datos demo-lingüísticos, dar resultados muy similares. Pero no es el caso. El análisis muestra tres causas diferentes para explicar diferencias significativas: 1) hipótesis muy diferentes para L2, 2) la gestión del multilingüismo, 3) un error que aparece en el porcentaje de francófonos conectados a la Internet (IWS presenta una cifra anormalmente baja de 26%, mientras que este estudio indica 81%).

Este estudio proporciona cifras de porcentaje de locutores conectados mucho más altas que los de IWS para el francés, el ruso y en menor grado para el español y el alemán. Las diferencias en los datos demo-lingüísticos utilizados explican en cierta medida estas diferencias. Sin embargo, una simulación del modelo utilizado en este estudio con las figuras demo-lingüísticas exactas utilizadas por IWS sigue mostrando diferencias significativas en lo que se refiere al porcentaje mundial de personas conectadas para el francés (75%) e el inglés (25%). El método de cálculo para L2 de este estudio podría explicar parte de estas diferencias (ya que puede sobrestimar los porcentajes mediante la determinación de las tasas de conexión para hablantes L2 idénticas a los hablantes L1) pero parece también que IWS ha cometido un error en alguna parte a lo largo de su cálculo con el francés[[31]](#footnote-31). Por otra parte, la gestión del multilingüismo con porcentajes de la población mundial conduciría rápidamente al absurdo sobre el resto de las lenguas de IWS si se extendiera su lista de lenguas.

TABLA 9 : Diferencias de hipótesis con IWS para L1 + L2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LENGUA** | **L1 + L2** | **L1 + L2** | **diferencias** | **Diferencias** |
|  | **IWS** | **Estudio** | **absolutas** | **%** |
| Inglés | 1404089957 | 955652967 | 448436990 | **47%** |
| [Chino](http://www.internetworldstats.com/stats17.htm) | 1417846280 | 1347846280 | 70000000 | **5%** |
| Español | 454877800 | 525 877 800 | 71000000 | **-14%** |
| Árabe | 388915929 | 333 915 929 | 55000000 | **16%** |
| [Portugués](http://www.internetworldstats.com/stats20.htm) | 267372370 | 228 372 370 | 39000000 | **17%** |
| Japonés | 125388300 | 125388300 | 0 | **0%** |
| [Malayo](https://en.wikipedia.org/wiki/Malay_language) | 293 998 640 | 228 998 640 | 65000000 | **28%** |
| [Ruso](https://en.wikipedia.org/wiki/Russian_language) | 146013300 | 243 287 120 | 97273820 | **-40%** |
| Francés | 394660700 | 230 146 470 | 164 514 230 | **71%** |
| Alemán | 95419000 | 122108000 | 26689000 | **-22%** |

TABLA 10 : Datos InternetWorldStats (junio de 2016)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LENGUA** | **USUARIOS**  **INTERNET** | **%**  **LOCUTORES** | **%** | **POBLACIÓN** |
|  |  | **CONECTADO** | **MUNDIAL** |
| Inglés | 948 608 782 | 67,76% | 26,27% | 1 400 052 373 |
| [Chino](http://www.internetworldstats.com/stats17.htm) | 751 985 224 | 53.12% | 20,82% | 1 415 572 934 |
| Español | 277 125 947 | 61.55% | 7,67% | 450 235 963 |
| Árabe | 168 426 690 | 43.37% | 4,66% | 388 332 877 |
| [Portugués](http://www.internetworldstats.com/stats20.htm) | 154 525 606 | 57.93% | 4,28% | 266 757 744 |
| Japonés | 115 111 595 | 91.02% | 3.19% | 126 464 583 |
| [Malayo](https://en.wikipedia.org/wiki/Malay_language) | 109 400 982 | 37.76% | 3,03% | 289 702 633 |
| [Ruso](https://en.wikipedia.org/wiki/Russian_language) | 103 147 691 | 70.48% | 2,86% | 146 358 055 |
| Francés | 102 171 481 | *25.94%* | 2,83% | 393 892 299 |
| Alemán | 83 825 134 | 88.26% | 2,32% | 94 973 855 |
| RESTO | 797 046 681 | *33.66%* | 22.07% | 2 367 750 664 |
| [TOTAL](http://www.internetworldstats.com/stats.htm) | 3 611 375 813 | 49.20% | 100.00% | 7 340 093 980 |

La siguiente tabla muestra los datos obtenidos en este estudio y la comparación con IWS.

TABLA 11: Comparación con IWS después de reducción a 100%

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **LENGUA** | **USUARIOS**  **INTERNET** | **% LOCUTORES** | **%**  **MUNDIAL** | **POBLACIÓN** | **DIFFE** | **RENCIAS** |
|  |  | **CONECTADO** |  |  | **en%** |
| Inglés | 746575851 | 78.18% | 23.94% | 954 971 600 | -2,32% | -9,71% |
| [Chino](http://www.internetworldstats.com/stats17.htm) | 692 230 352 | 51.36% | 22.20% | 1347846000 | 1,38% | 6.21% |
| Español | 306111437 | 58.21% | 9,82% | 525 877 800 | 2.14% | 21,83% |
| Árabe | 140 122 838 | 41.97% | 4,49% | 333 842 070 | -0,17% | -3,78% |
| [Portugués](http://www.internetworldstats.com/stats20.htm) | 134 142 584 | 58.74% | 4,30% | 228 372 370 | 0,02% | 0,54% |
| Japonés | 113372837 | 90.43% | 3,64% | 125376800 | 0,45% | 12,33% |
| [Malayo](https://en.wikipedia.org/wiki/Malay_language) | 87701961 | 38.30% | 2,81% | 228 998 640 | -0.22% | -7,70% |
| [Ruso](https://en.wikipedia.org/wiki/Russian_language) | 166904263 | 68.60% | 5,35% | 243 287 120 | 2,50% | 46.64% |
| Francés | 186641894 | 81.10% | 5,99% | 230 146 420 | 3.16% | 52.73% |
| Alemán | 105 539 688 | 86.43% | 3,38% | 122108000 | 1,06% | 31.42% |
| **Resto** | **438771072** | **16,05%** | **14,07%** | **2733112497** | **-8.00%** | -56.84% |
| [**TOTAL**](http://www.internetworldstats.com/stats.htm) | **3118114777** | **44.08%** | **100.00%** | **7073939317** | **0.00%** |  |

Algunas lenguas están experimentando una marcada diferencia, en orden descendente: el resto de los lenguas (- 56%), francés (+ 53%), ruso (- 47%), alemán (+ 31%), español (- 22%). La diferencia que aparece en el resto es una consecuencia clara de cómo IWS maneja el multilingüismo.

La siguiente tabla muestra los resultados cuando los mismos datos demo-lingüísticos de IWS se aplican al modelo de este estudio. Con estos datos el modelo establecido en el presente estudio encontró valores más altos del inglés (+ 25%) y del francés (+ 72%). También muestra un resto negativo, síntoma de un manejo inadecuado del multilingüismo. Las diferencias se explican por el tratamiento del multilingüismo (aplicar porcentaje de hablantes L1+L2 sobre la población mundial en lugar del total mundial de los locutores conduce a una contradicción en IWS la cual está oculta por el hecho de que se trata de un número limitado de lenguas). Si esto se corrigiera, seguiría quedando un error fuerte en el tratamiento del francés por IWS como se mencionó anteriormente.

TABLA 12 : Simulación con datos de IWS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LENGUA** | **USUARIOS  DE LA INTERNET** | **% LOCUTORES** | **%** | **POBLACIÓN** |  |
|  |  | **CONECTADO** | **GLOBAL** | **DIFERENCIAS** |
| Inglés | 1094530237 | 78.18% | 35.10% | 1400052373 | 25% |
| [Chino](http://www.internetworldstats.com/stats17.htm) | 727013732 | 51.36% | 23,32% | 1415572934 | 11% |
| Español | 262080616 | 58.21% | 8,41% | 450235963 | 9% |
| Árabe | 162994152 | 41.97% | 5,23% | 388332877 | 11% |
| [Portugués](http://www.internetworldstats.com/stats20.htm) | 156689591 | 58.74% | 5,03% | 266757744 | 15% |
| Japonés | 114 356 472 | 90.43% | 3,67% | 126 464 583 | 13% |
| [Malayo](https://en.wikipedia.org/wiki/Malay_language) | 110 950 392 | 38.30% | 3,56% | 289702633 | 15% |
| [Ruso](https://en.wikipedia.org/wiki/Russian_language) | 100 407 220 | 68.60% | 3,22% | 146 358 055 | 11% |
| Francés | 319 434 926 | 81.10% | 10,24% | 393892299 | 72% |
| Alemán | 82087259 | 86.43% | 2,63% | 94973855 | 12% |
| **Resto** | **12429819** | **-0.59%** | **-0,40%** | **2101596001** |  |
| [**TOTAL**](http://www.internetworldstats.com/stats.htm) | **3118114777** | **44.08%** | **100.00%** | **7073939317** |  |

## 4.3 Sensibilidad de los factores

Todo el modelo descrito se ha automatizado en una hoja de cálculo Excel y es posible obtener al instante los nuevos resultados de cualquier cambio de parámetro. Esta flexibilidad, que se usó para la comparación con IWS, también permitió probar la sensibilidad de cada factor sobre los resultados. Por lo tanto, apareció, en el caso del francés, que su cuarta posición en términos de *poder* se mantuviera si los datos L2 fuesen cambiados a un valor mucho más bajos. Al contrario, la sensibilidad del indicador de *tráfico* en relación con la elección de los sitios se encontró muy fuerte si se utiliza el promedio simple; eso condujo a la decisión de usar el promedio truncada al 20% para evitar el sesgo de selección.

Diferentes hipótesis acerca de las cifras de segundos lenguas, un método más preciso para segundas lenguas, la elección de una métrica diferente para determinar indicadores (promedio, promedio truncado, mediana) o diferentes combinaciones de micro-indicadores, obviamente, lleva a resultados de indicadores con algunas diferencias, relativamente menores[[32]](#footnote-32) pero que pueden cambiar el orden de las lenguas que son lo suficientemente cercanas en los resultados[[33]](#footnote-33).

Las diversas pruebas realizadas demostraron que el estudio proporciona una herramienta de trabajo y simulación capaz de ofrecer resultados inmediatos cuando se cambian los parámetros, lo que probablemente hace que sea un instrumento único y sin precedentes en el campo de la cibermetría de las lenguas. Es tentador ampliar el estudio para simulaciones basadas no solamente en cifras L1 + L2, pero añadiendo un concepto ad hoc de Li que se podría definir como *la población con suficiente dominio de un lenguaje para hacer un uso eficaz de las diversas características de la Internet,* aunque sigue siendo un desafío mayor encontrar cifras demo-lingüísticas fiables y consistentes para Li en cada lengua, inclusivo las más importantes...

# Limitaciones metodológicas, análisis de sesgo y controles realizados

## 5.1 ¡Ojo con los sesgos!

La demo-lingüística está lejos de ser una ciencia exacta y las divergencias entre los especialistas en cifras de L1, L2 y Li, son importantes y crecen de L1 a Li, cuando las definiciones se vuelven menos precisas. Por otro lado, la Internet es un universo en expansión que alcanzó cifras impresionantes para el número de personas conectadas, servidores y páginas web y para el cual no hay fuentes disponibles para el tamaño de la Web (y de los índices de los motores de búsqueda). Por otra parte, algunas fuentes toman mucha libertad con los datos para responder al marketing del clic y al dominio de la publicidad que ha destronado otras formas de modelo de negocio en la Internet más compatible con la teoría de la información.

La historia de la cibermetría lingüística estuvo marcada por el triunfo de los sesgos... En el año 2000, Inktomi, un motor de búsqueda ahora desaparecido, declaró, con un enorme poder de marketing, que más del 85% de la Web estaba en inglés, sobre la base de datos expuestos sin ningún tipo de información sobre el método de cálculo. Se mostró sólo 10 lenguas... y un total de 100%, como si las otras lenguas contarían para nada! Una estimación seria del resto de lenguas habría llevado la presencia del inglés a 70%.

TABLA 13 Distribución de lenguas en la Web según Inktomi (2000):

|  |  |
| --- | --- |
| LENGUA | PROPORCIÓN DE LA WEB (%) |
| Inglés | 86,54 |
| Alemán | 5,83 |
| Francés | 2,36 |
| Italiano | 1,55 |
| Español | 1,23 |
| Portugués | 0,75 |
| Holandés | 0,54 |
| Finlandés | 0,50 |
| Sueco | 0,36 |
| Japonés | 0.34 |

En 1997, una empresa canadiense, Alis Tecnología, publicó los resultados de un estudio llevado a cabo con el apoyo de la Internet Society y daba una cifra del 80% de inglés en la Web. Un grupo de investigadores de la OCLC[[34]](#footnote-34) reanudó el método en 1999 y en 2003 y logró el mismo resultado de 80% de inglés de una manera estable y eterna ... a pesar del crecimiento exponencial de la Web y la constante disminución en el porcentaje de angloparlantes en la Internet. Estos datos alimentaron la versión, retomada y amplificada por los medios de comunicación, de una Web estable con un 80% de páginas en inglés durante 10 años, entre 1987 y 2007. Sin embargo, en el mismo período, el observatorio de Funredes mostró una disminución del espacio del inglés en la Web del 80% en 1997 al 45% en 2007. Se necesitaron dos publicaciones de la UNESCO, en 2006 y 2009[[35]](#footnote-35), para poner fin a esta desinformación. ¿Cuál fue el sesgo mayúsculo del método Alis retomado por OCLC? El método consistió en la aplicación de un algoritmo de reconocimiento de lengua en **una muestra al azar de 3000 sitios**, seleccionados a partir de la generación aleatoria de números IP. Si el método había consistido en la repetición de la medición un centenar de veces (variando la muestra al azar) y había tratado a los resultados como la distribución de una variable aleatoria a la cual se aplicaba las leyes estadísticas (promedio, covarianza, ley de Fisher, intervalos de confianza...) habría sufrido sólo sesgos moderados (la elección de la página principal favorece el inglés, los algoritmos de reconocimiento de lengua favorecen el inglés y la ausencia de gestión del multilingüismo hace lo mismo), pero el problema fue que los resultados fueron el fruto de **uno y único sorteo al azar** y por lo tanto sólo sujetos a la única voluntad del azar!

Un estudio que va a combinar datos específicos de la Internet con datos demo-lingüísticos para la construcción de indicadores se encuentra en la intersección de estos dos mundos inciertos, lo cual, en términos estadísticos, implica la multiplicación de las incertidumbres y aumenta el tamaño de los intervalos de confianza. Por lo tanto, es necesario tener mucho cuidado e identificar los sesgos que pueden afectar a los datos o la metodología. Sólo así será posible validar la fiabilidad del método, inferir las posibles implicaciones de sesgos en los resultados y tratar de encontrar las correcciones apropiadas.

En este capítulo se intenta sistematizar el análisis de todos los sesgos o limitaciones que puedan ser causados ​​directa o indirectamente por las opciones metodológicas y/o las fuentes en ese estudio y muestra los controles que se han hecho en el tratamiento de los datos para asegurar resultados relativamente fiables, incluso si es imposible alcanzar la perfección en este campo.

## 5.1 Limitaciones y sesgos propios del método

El método se dedica a buscar, identificar, evaluar y utilizar micro-indicadores cuantitativos para construir indicadores que caracterizan adecuadamente la presencia de las lenguas en la Internet. Las fuentes por lengua siendo muy escasas, se tomó la opción de utilizar fuentes por país para superar esta limitación y gestionarla para obtener una visión estadística apoyada por suficientes micro-indicadores. Esta opción presenta un reto: el de la transformación de las cifras por país en cifras por lengua. Este desafío fue resuelto por el método de ponderación entre la distribución de hablantes de lenguas diferentes por país y los datos por país de algunas aplicaciones o espacios de la Internet. Es ahora necesario entender plenamente los criterios de validez y las limitaciones de este enfoque por ponderación.

Se creó una matriz que contiene los 7500 lenguas y 192 países. Por simple aritmética es posible repartir los datos por país de cada criterio medido entre cada lengua (en proporción de la presencia de las lenguas en cada país) y obtener de esta manera los datos por lengua de cada uno de los criterios medidos. ¿Cuál es la hipótesis que sustenta este método y ¿es correcto?

El supuesto implícito es que los datos acerca de la Internet en un país determinado son válidos de la misma manera para todas las lenguas que se hablan en este país. ¿Esto corresponde a la realidad o al menos se acerca?

Para tomar el ejemplo más obvio, el porcentaje de personas conectadas a la Internet en un país: ¿es este porcentaje igual para cada lengua hablada, y por lo tanto independiente de la lengua materna de los habitantes de ese país? Esto es por supuesto una hipótesis simplificadora que implica, en este ejemplo, que la brecha digital no se correlaciona con la lengua dentro del país. Esto es a menudo falso: es probable que los hablantes de ciertas lenguas tengan un porcentaje de conexión a la Internet significativamente más bajo o más alto que el promedio nacional. Este puede ser el caso, porque representan una población con niveles socioeconómico, cultural, o educativo muy diferentes de la media. Sin embargo, esta es una aproximación aceptable que no introduce distorsión violenta en los resultados bajo la condición de no aplicarla a espacios o aplicaciones que podrían agravar ese sesgo (por ejemplo, no se podría justificar esta simplificación para una aplicación que consiste a aprender una de las lenguas del estudio) y a condición también de **no tomarla como válida para diferenciar la situación respectiva de las lenguas dentro de un país[[36]](#footnote-36).**

Para cada lengua, además del país (o de los países) donde es lengua nacional se usará la distribución de locutores en diferentes países[[37]](#footnote-37) y la pregunta queda de saber si es correcto recíprocamente de considerar que los hablantes de una lengua que residen en un país donde no es lengua nacional tengan la misma tasa de conectividad a la Internet que los hablantes de la lengua nacional de ese país. Intuitivamente, uno puede entender que el riesgo de error (o el tamaño de la diferencia posible entre tasa nacional y tasa de una lengua no nacional) queda más grande cuando se trata de lenguas minoritarias con números bajos de hablantes. Por esta razón, la elección se hizo finalmente en lenguas que tienen más de 5 millones de hablantes (L1) como medida de prudencia.

Este problema sólo afecta a los cálculos para las lenguas nativas (L1); es probable que el tratamiento de L2 provoca otro tipo de riesgo de distorsión entre los cálculos y la realidad y merece un análisis más detallado que se hace en el siguiente capítulo.

La cuestión de las fuentes de datos incompletas se resolvió mediante la técnica de extrapolación, aplicada ya sea directamente en proporción directa a la tasa de conectividad de los respectivos países, ya sea con el método de cuartil. Este método resuelve la situación correctamente y se puede considerar que el sesgo introducido (en el caso en que los países tienen en realidad un valor para este micro-indicador mucho mayor o mucho menor que el calculado por la extrapolación, por razones particulares relativas a la naturaleza de este micro-indicador[[38]](#footnote-38)) serán compensados ​​en promedio y se convierten en marginales cuando el número de casos es lo suficientemente grande.

Ahora tenemos que cuestionar sobre limitaciones o sesgos para cada uno de los criterios relacionados con los micro-indicadores en cada elemento del complejo utilizados: lenguas, países y fuentes.

## 5.2 Lenguas

### 5.2.1 Selección de la fuente para el cálculo de L1

En esta edición, el Proyecto Joshua fue la principal fuente de los datos demo-lingüísticos en lengua materna (L1). En la próxima edición se contempla utilizar los datos de Ethnologue lo que sería un costo adicional para el estudio salvo si un acuerdo institucional puede ser logrado entre MAAYA y la SIL.

### 5.2.2 El caso de L2

El método elegido en este estudio para la L2 no llega a diferenciar por país y por lo tanto la corrección fina obtenida gracias a la ponderación no se aplica. Aplicar la misma tasa de crecimiento de L1 a L1+L2 a los resultados de los cálculos basados ​​en L1, implica, de hecho, la misma suposición simplificadora como la transformación de datos de país hacia lengua (los datos de conectividad a la Internet en cada país tienen el mismo comportamiento para todos los lenguas hablada en este país), pero esta vez, la implicación no está dentro de un país, pero entre todos los países del mundo. La hipótesis de simplificación en este caso es que la tasa de conexión a la Internet de personas en una lengua materna dada también es idéntica para los hablantes de esta lengua como L2, independientemente de sus países de residencia. Si para una determinada lengua las personas que la utilizan como L2 se encuentran principalmente en los países con bajos porcentajes de personas conectadas, mientras que los de las lenguas nativas (L1) se encuentran en países con altos porcentajes, entonces el método va a sobrestimar la población L1+L2 conectado a la Internet para esa lengua.

En el caso imaginario de un lenguaje L que tendría 100 millones de hablantes L1 en un país P1, con un porcentaje de conectividad a la Internet de 80%, y los 100 millones de hablantes de segunda lengua residen todos en un país P2 con una tasa de conectividad de 40%, entonces el error del método es de sobreestimar de un 50% el número de locutores L1 + L2 conectadas.

Nc = número calculado de locutores conectados

Nr = número real de locutores conectados

Nc (L1) = Nr (L1) = 100 millones x 80% = 80M

Nc (L1 + L2) = 80M x 2 = 160M

Nr (L1 + L2) = 100 M x 80% + 100M x 40% = 120M

Tamaño del error = 40M o 50% adicional

Esto es por supuesto una situación totalmente imaginaria y rebuscada y sólo se expone por necesidad pedagógica pues la realidad no tiene tales diferencias tan extremas.

Sería posible, sin embargo, para superar este potencial inconveniente, de utilizar el método de los cuartiles a condición de disponer de los datos correspondientes, por ejemplo:

* P0: la proporción de L2 que reside en países donde la tasa de conexión es menos de 15% de la población: primer cuartil (Q1 = nota más baja de personas conectadas: 0,08 con los valores actuales)
* P1: la proporción que reside en países donde la tasa es más de 15% pero menos del 35% segundo cuartil (Q2 = 0,34)
* P2: la proporción que reside en países donde la tasa es más de 35% pero menos del 65: tercer cuartil (Q3 = 0,56)
* P3: la proporción que reside en países donde la tasa es más de 65% pero menos del 85: cuarto cuartil (Q4 = 0,76)
* P4: la proporción residente en países donde la tasa es más de 85%: último cuartil (el más alto grado: Q5 = 1).

y la fórmula para volver a calcular la tasa de crecimiento a aplicar a los indicadores sería:

i = 4

Tc (j) = (T (j) x PL1 (j)) / Σ Pi x Qi

i = 0

donde

T (j) es la relación L1+L2/L1 para la lengua j

Tc (j) es la misma relación, pero corregida por el método de los cuartiles a aplicar a los resultados para L1 para pasar a resultados L1+L2

PL1 (j) es él % de personas conectadas para L1 (j).

En el ejemplo de francés, con L2 = 153 485 770, T = 3 y PL1 = 0,81, si la distribución fuese de 20% para cada cuartil, el valor corregido a aplicar seria de:

Tr(francés) = 3 x 0,81 / (20% x 0,08 + 20% x 0,34 + 20% x 0,56 + 20% x 0,76 + 20% x 1 = 0,548)

En ese caso en vez de 3 se usaría un factor de 2, lo que posiblemente es más cerca de la realidad.

Por supuesto, es también posible simplificar este método y sólo dividir en dos categorías, altos y bajos, todo depende en realidad de la claridad que podemos de la proporción de la población L2 entre esas dos categorías.

La calidad de los resultados, sin duda estaría mejor garantizada mediante la aplicación de este método simple. Mientras se implementa este método más fino (pero condicionado por la procuración de datos más detallados sobre poblaciones L2) es razonable considerar que el método actual podría causar una sobreestimación de los resultados del inglés y del francés que tienen grandes poblaciones L2 en países de conectividad Internet bastante menor que respectivamente los Estados Unidos (o el Reino Unido) y Francia (o Bélgica).

### 5.2.3 La reducción del número de lenguas

El principio fundamental de la independencia del proceso en comparación con el número de lenguas tratadas se verificó en pruebas lo que garantiza resultados idénticos cualquiera que sea el número de lenguas seleccionadas (por supuesto siempre cuidando el cálculo de la línea de "lenguas restantes"). Por tanto, es teóricamente posible tratar todas las lenguas con el método. En la práctica se trata de un impresionante volumen de cálculo (7500 x 200 = 1 millón y medio de células de Excel) y el tamaño del archivo es demasiado pesado para un manejo rápido (más de 100 MB) por lo que representa un enorme costo marginal para un beneficio marginal nulo porque sólo 500 lenguas entre 7500 tienen una existencia virtual. Por otra parte, las hipótesis simplificadoras invalidarían los resultados para las lenguas con cantidades de locutores que no se miden en millones. Después de varias pruebas con diferentes figuras, la selección se fijó finalmente con la lista de lenguas con más de 5 millones de locutores (140 lenguas) en esa primera edición, quedando una tarea futura tratar de bajar ese umbral en dirección de lenguas con más de 1 millón de locutores.

### 5.2.4 Comprobación de la invariancia de los resultados con el número de lenguas

Por supuesto, era necesario asegurarse de que, como la intuición dicta, operaciones aritméticas realizadas para transformar datos por país en datos por lengua obtuvieron los mismos resultados independientemente del número de lenguas procesadas. Esto ha sido comprobado varias veces a cambiando de 7500 a 390, 100, 28, 85 lenguas, la lista de lenguas de más de 10 millones de hablantes y, finalmente, a más de 5 millones de hablantes. Estas pruebas permitieron a veces descubrir errores en los datos y representaron una buena inversión para la fiabilidad de los resultados.

## 5.3 Países

Después de unas pocas iteraciones, se determinó la lista final de países a considerar de forma simplificada: la eliminación de algunos países con muy pocos habitantes y/o porque la UIT no proporciona datos y la agrupación en los principales países de estados satélites.

Los datos de la UIT se han modificado para reflejar los cambios en la lista de países retenidos, mediante la vinculación de los datos de población y la modificación de los datos del porcentaje de personas conectadas de forma proporcional a sus respectivas poblaciones.

Un buen ejemplo es China, que tiene un porcentaje de conexión a la Internet mucho más bajo que Hong Kong o Macao y, por tanto, la integración de estos dos estados en los datos de China se hará en proporción a sus respectivas tasas. Algunos países se han retirado de la lista (como Corea del Norte, San Marino o las Malvinas), otros se unieron a otro país (por ejemplo, las Islas Feroe y Groenlandia en Dinamarca). El Reino Unido es el país que fue objeto del mayor número de territorios adjuntos.

Para unos pocos países grandes la UIT no proporciona datos para los usuarios conectados a la Internet y los hablantes de estos países no se integraron en el modelo: este es el caso de Corea del Norte (que no está muy conectada a la Internet), Sahara Occidental, Kosovo y Sudán del Sur.

Por último, existe una lista de estados pequeños que no se han tenido en cuenta, ya sea porque la UIT no proporciona datos (en cursiva) o porque su población es muy limitado: *Bonaire, San Eustaquio y Saba, Curacao, Isla de Navidad*, Islas Cocos, Isla de Man, Isla Norfolk, Pitcairn, Malvinas, *Islas Marianas del Norte, Mayotte*, *Nauru, Palau*, *San Bartolomé*, San Marino, San Martín, *San Pedro y Miquelón*, la isla *Svalbard y Jan Mayen, Territorio Británico del Océano Índico, Ciudad del Vaticano.*

TABLA 14 Países: afiliación de territorios

|  |  |
| --- | --- |
| **País de afiliación** | **LISTA DE PAÍSES RELACIONADOS** |
| China | Hong Kong y Macao |
| Estados Unidos | Guam, las Islas Vírgenes de Estados Unidos, Puerto Rico y Samoa Americana |
| Dinamarca | Groenlandia y las Islas Feroe |
| Francia | Guadalupe, Guayana, Martinica, Nueva Caledonia, Polinesia Francesa, Reunión, y Wallis y Futuna |
| Nueva Zelanda | Islas Cook, Niue y Tokelau |
| Países Bajos | Aruba, Antillas holandesas |
| Reino Unido | Anguilla, Antigua y Barbuda, Ascensión, Bermudas, Gibraltar, Guernsey, Islas Caimán, Islas Vírgenes, Jersey, Montserrat, Santa Elena, San Cristóbal y Nieves, y Turcas y Caicos |

## 5.4 Fuentes

### 5.4.1 Conceptos básicos

Los tres principios siguientes se han respetado en general:

1. siempre ir a la primera fuente[[39]](#footnote-39);
2. siempre preferir las últimas fechas de las fuentes;
3. no asumir la responsabilidad de la corrección de las aproximaciones detectadas en las fuentes (excepto como se describe a continuación).

El primer principio se aplica obviamente a las fuentes que proporcionan porcentajes mundiales por país, no a las fuentes individuales para cada país (lo cual sigue siendo el trabajo de la primera fuente, la que compone las figuras mundiales). La celebración del primer principio implica a menudo el segundo principio, pero debe quedar claro que los cálculos pueden combinar diferentes fechas por país dentro de una cobertura mundial, sino también entre los elementos puestos en ecuación (véase *5.4.3 La cuestión de fechas*).

### 5.4.2 Excepciones a los principios básicos

Sin embargo, hay algunos casos en los que la mejora de las fuentes primarias se impone porque son datos críticos para todo el proceso: es el caso de los datos demográficos por países y personas conectadas por país.

Otra excepción se hizo en el caso de datos de suscripción ofrecidos por Statista para Netflix. Hay dos series: los datos actuales y una proyección a 2021. El descubrimiento de una reducción, en la proyección para el año 2021, del número de suscriptores en los EE.UU., país líder de lejos en esta aplicación, permitió la comprobación y el descubrimiento de que los datos para Estados Unidos no fueron documentados en la proyección para el año 2021. La técnica de extrapolación aplicada a esa categoría de fuente ha sido capaz de limitar el daño, pero el resultado seguía siendo insatisfactorio con una disminución de las suscripciones entre 2016 y 2021 para los Estados Unidos. Esto debía conducir a la eliminación de esta fuente, salvo al obtener datos fiables por otra fuente para el crecimiento de suscriptores en los EE.UU. en este mercado, lo que se obtuvo a través de la ayuda de Statista.

### 5.4.3 La cuestión de las fechas

La cuestión de las fechas de las fuentes es compleja; es común que algunas fuentes de datos por país utilizan diferentes fechas para diferentes países, una situación obligatoria debido a las diferencias en el nivel de la actualización de las fuentes de cada país. La presencia en la misma operación aritmética de las fuentes de datos con diferentes fechas no representa un obstáculo, aunque puede tener un impacto marginal en los resultados. Por lo tanto, la fuente de la UIT para la tasa de conectividad a la Internet por país es de 2015, pero los datos demográficos por país pueden ser de diferentes fechas; uno debe ser consciente de esta situación cuando procesa los datos, pero eso no les invalida.

En el caso de fuentes que combinan varias fuentes externas para construir indicadores compuestos, desafortunadamente es común que no utilizan los datos más actualizados; comprometerse a actualizar estas fuentes compuestas sale del alcance de este estudio. Este es el caso del indicador "índice de servicios de Gobierno en línea" que WebIndex utiliza para crear indicadores compuestos: da Francia con la primera calificación con una puntuación de 100. Se buscó la primera fuente después y la búsqueda reveló que WebIndex utilizaba una fuente de fecha de 2013 mientras que la primera fuente[[40]](#footnote-40) proporciona los datos de 2016, donde Francia se mueve al quinto lugar con un puntaje de 94.20. Los resultados no reflejan esta evolución y eso tiene efectos totalmente marginales en los resultados.

### 5.4.4 La cuestión del significado de la transformación país > lengua

Si el significado de esta transformación es intuitivamente obvio para los datos por país puede ser más difícil de entender, incluso incomprensible, acerca de porcentajes que no se expresan por país.

El ejemplo que tiene más sentido es el número de usuarios de la Internet por cada país. Si estos datos se expresan en términos de cantidad por país o por un porcentaje de personas conectadas por país (esta es la primera fuente ofrecido por la UIT) o como un porcentaje de usuarios de la Internet en un país en relación al total mundial de usuarios de la Internet, una aritmética simple permite pasar de un valor al otro. El producto de matriz (lengua x país) con las cantidades por país, por tanto, permite con una aritmética simple, de obtener el número de usuarios de la Internet por cada lengua (o el porcentaje de personas conectadas mundial para cada lengua, dependiendo del modo seleccionado de cálculo).

En el caso de los datos expresados ​​en términos de porcentajes nacionales, a veces se hace muy difícil hacer sentido de la transformación por lengua o incluso se hace imposible encontrar una técnica de extrapolación para este micro-indicador (los dos puntos son probablemente relacionados). En estos casos, la opción elegida fue de abandonar esta transformación y, por lo tanto, la integración de este micro-indicador en la producción de indicadores. Este es el caso por ejemplo del número de solicitudes por país de eliminación de datos a Google (el derecho a ser olvidado). La extrapolación no es posible porque la información está relacionada con la cultura jurídica del país (y si existe o no una ley) y no tiene relación directa con el porcentaje de personas en línea en el país. La transformación a lengua de este indicador, por otra parte, causa confusión, ya que podría deducirse que el borrado de datos se realiza en respeto a la lengua presentada por el cálculo cuando se trata de la lengua oficial del país. Si quisiéramos dar un sentido estricto y consistente para la transformación a lengua, en este caso específico, tendríamos que expresar: "el porcentaje de hablantes de una lengua en particular haciendo una petición de borraduras a Google en la lengua nacional de su país de residencia" .. . esto no es ni de interés obvio ni fácil de entender!

### 5.4.5 Las limitaciones debidas a las fuentes

Las limitaciones debidas a fuentes incluyen primero, por supuesto, las limitaciones y los posibles sesgos de los datos proporcionados por la fuente principal. En el caso del indicador de personas conectadas a la Internet por país, el análisis del método utilizado por la fuente[[41]](#footnote-41), muestra:

* Es el indicador HH7: proporción de individuos que utilizan la Internet con descripción precisa del método y de sus limitaciones[[42]](#footnote-42);
* los valores se establecieron a partir de encuestas de población;
* la definición establece que las personas que se conectan a la Internet son contabilizadas independientemente del dispositivo que utilicen (que no es necesariamente un ordenador y también puede ser un teléfono móvil, una tableta, un asistente digital personal, una consola de juegos, TV digital, etc.) y que el acceso se puede hacer por la red fija o la red móvil;
* la norma adoptada para contabilizar es la de personas de 16 a 75 años que se han conectado al menos una vez en los últimos 3 meses.

Un análisis más detallado muestra que:

* aproximadamente el 60% de las fuentes son estimaciones propias de la UIT y 15% se derivan de Eurostat utilizando exactamente las mismas definiciones y criterios[[43]](#footnote-43);
* para el resto de los países, los datos son proporcionados por diferentes autoridades, según el país. Dichas autoridades suelen utilizar diferentes criterios; por lo tanto, el rango de edad puede iniciar, a los 3, 5, 6, 7, 10, 12, 15 o 20 años y a veces no hay límite de edad;
* para algunos países los datos de los Estados asociados pueden ser contabilizados con los del país y en otros casos independientes[[44]](#footnote-44).

La UIT es la fuente más fiable, de lejos, para este tipo de datos y otras fuentes menudo lo utilizan sin citar y, a veces, mediante cambios indocumentados. Estos datos son, por otra parte, un elemento esencial en todo el proceso del modelo expuesto y el gran cuidado se ha tenido de considerar cuidadosamente todos los factores, en particular la distribución de cifras exactas en el caso de países con territorios asociados.

¿Cuáles son los posibles sesgos de los datos de la UIT? Por un lado, los países europeos, para los que la UIT uso los datos de Eurostat y los países en los que la UIT ha colocado sus propias estimaciones se pueden considerar datos particularmente fiables. Por el otro lado, una ligera sobreestimación de algunos de los países donde autoridades locales ofrecen los datos puede ser temida, ya sea porque los criterios de definición son más amplios, como ya se ha mencionado, o también debido a que algunos de estos países podrían haber hecho una estimación generosa para mostrar la buena validez de sus políticas contra la brecha digital. Es bueno saber este sesgo potencial, pero no es importante y puede ser considerado generalmente como bastante marginal.

### 5.4.6 Sesgo potencial de Alexa.com y W3Techs

La mayoría de los micro-indicadores son datos de tráfico a sitios web donde Alexa.com se ha utilizado como fuente. ¿Cuáles son los posibles sesgos de esta fuente?

Alexa ofrece estadísticas de tráfico de las aplicaciones de la Web a través de una barra de herramientas que una supuesta amplia muestra de usuarios de la Internet[[45]](#footnote-45) acordó instalar en su navegador. Alexa está informado de las visitas a los sitios a través del software que esta desencadenado a partir de la barra de herramientas y, con esos datos, realiza una clasificación de los 25 millones de los sitios web más concurridos. Alexa, como la mayoría de las compañías privadas de la Internet, no es transparente en cuanto a su metodología y los sesgos que pueden derivarse; en particular, no hay por desgracia estadísticas sobre las barras de herramientas instaladas que permitan conocer su total ni mucho menos su distribución por país o lengua.

Por supuesto, Alexa sólo ofrece sus estimaciones de datos de tráfico para los sitios que forman parte de la parte alta de la clasificación. La selección de sitios web para el estudio fue hecha con una buena posición en este ranking (por lo general con un ranking de Alexa de menos de 10 000, a excepción de algunos temas menos populares para los cuales clasificaciones hasta 50.000 han sido seleccionadas). Los sesgos que pueden surgir a partir del método de Alexa son:

* una concentración de la atención en los sitios más visitados (por definición);
* una subestimación en los países donde la barra de herramientas de Alexa no es muy descargada (especialmente una fuerte subestimación del tráfico interior chino ya que el 80% de los sitios no son accesibles por el sistema de nombres del ICANN[[46]](#footnote-46));
* una sobreestimación relativa de tráfico de los países que han instalado la barra de Alexa en mayor proporción (con toda probabilidad, en especial los países occidentales y más particularmente los Estados Unidos[[47]](#footnote-47)).

¿Cuáles son las implicaciones de estos sesgos en los resultados del estudio?

Se sospecha que existe una sobrerrepresentación de los países occidentales y entonces de las lenguas occidentales del resto de lenguas y una sobrerrepresentación del inglés en comparación con todas las lenguas, eso en un porcentaje muy alto en comparación con lenguas no occidentales.

Un estudio comparativo de los datos de suscripción de Facebook, Twitter y Linkedin ha confirmado esta sospecha.

El estudio comparó los datos en términos de suscripciones para estas aplicaciones con los datos en términos de tráfico proporcionados por Alexa. Las diferencias entre los dos resultados (después de tener cantidades cambiadas en porcentaje) son relativamente grandes y, como era previsible, en general, con datos de tráfico de Alexa sumamente altos en comparación con los números de suscriptores en el caso de las lenguas occidentales. Las diferencias entre el porcentaje de suscriptores y el tráfico están demostrando ser muy importantes en la mayoría de los países no occidentales.

Se nota que el portugués se comporta como las lenguas no occidentales en este tipo de comparaciones, y la razón no ha sido elucida aun (¿habría una razón específica en Brasil que haría que la descarga de la barra de herramientas de Alexa sea un problema o bien se trata de que el tiempo de conexión a los distintos los servicios son mucho más cortos debido a los precios, o tal vez será lo que los brasileños aman tanto las redes sociales que tienden a tener varias cuentas?). Este chequeo puso en relieve las favorables calificaciones del francés, no es el caso del español y mucho menos aún del finlandés (el último caso probablemente demuestra que la técnica de extrapolación no siempre es capaz de proporcionar correctamente los niveles reales).

La siguiente tabla muestra el resultado de dividir los valores propuestos por Alexa (en términos de tráfico) por los valores obtenidos a través de Statista (en términos de suscripciones), clasificadas en orden descendente, la dicotomía entre las lenguas occidentales (excepto portugués) y las lenguas asiáticas aparece con suma claridad. Un factor de 1 indicaría la equivalencia de los resultados, el resultado es superior a 1 cuando los valores propuestos por Alexa (tráfico) son superior a los de Statista (suscripciones), denotando un sesgo favorable de Alexa.

La explicación de las diferencias entre los datos de comparación se debe a la combinación de dos factores:

1) El ya mencionado: Alexa favorecerá significativamente los países occidentales como un reflejo natural de la distribución de la barra de herramientas.

2) Pero también un factor adicional que podría jugar un papel está relacionado con el tiempo promedio de los usuarios permanecen en sesión el cual, por razón de tarificación, debe ser generalmente mayor en los países industrializados.

Sin embargo, el segundo factor no puede explicar por sí solo el tamaño de las diferencias observadas y esto es sin duda el primer factor, el sesgo de Alexa[[48]](#footnote-48), que es el principal responsable.

TABLA 15 : Relaciones de tráfico / suscriptores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **FACEBOOK** | **TWITTER** | **LINKEDIN** | **PROMEDIO** |
| Francés | 2.81 | 4.40 | 2.49 | **3.23** |
| Inglés | 3.11 | 2.65 | 1.63 | **2.46** |
| Tailandés | 2.35 | 1.23 | 2.38 | **1.99** |
| Catalán | 1.69 | 1.82 | 2.25 | **1.92** |
| Japonés | 1.41 | 1.65 | 2.33 | **1.79** |
| Noruego | 3.57 | 1.46 | 0.22 | **1.75** |
| Sueco | 2.36 | 2.07 | 0.74 | **1.72** |
| Danés | 3.18 | 0.34 | 1.22 | **1.58** |
| Griego | 2.42 | 1.48 | 0.45 | **1.45** |
| Alemán | 1.19 | 1.57 | 1.33 | **1.36** |
| Polaco | 2.14 | 0.75 | 0.85 | **1.25** |
| Húngaro | 1.59 | 0.35 | 0.97 | **0.97** |
| Ruso | 0.68 | 0.48 | 1.33 | **0.83** |
| Checo | 1.76 | 0.32 | 0.39 | **0.82** |
| Holandés | 1.54 | 0.34 | 0.59 | **0.82** |
| Italiano | 0.85 | 1.17 | 0.41 | **0,81** |
| Coreano | 0.63 | 0,81 | 0.95 | **0.80** |
| Urdu | 0.41 | 0.44 | 1.46 | **0,77** |
| Rumano | 1.21 | 0.43 | 0.51 | **0,72** |
| Español | 0.80 | 0,72 | 0.62 | **0,71** |
| Turco | 0.28 | 0.88 | 0.52 | **0.56** |
| Árabe | 0.51 | 0.48 | 0.57 | **0.52** |
| Lahnda | 0.52 | 0.28 | 0.67 | **0.49** |
| Tamil | 0.20 | 0.18 | 0.99 | **0.46** |
| Marathi | 0.16 | 0.14 | 1.00 | **0.43** |
| Kannada | 0.15 | 0.13 | 0.96 | **0.41** |
| Chino | 0.55 | 0.23 | 0.44 | **0.41** |
| Bengalí | 0.27 | 0.25 | 0.68 | **0.40** |
| Hindi | 0.21 | 0,19 | 0.74 | **0.38** |
| Portugués | 0.33 | 0,19 | 0.27 | **0.26** |
| Cebuano | 0.10 | 0.12 | 0.57 | **0.26** |
| Javanés | 0.07 | 0.05 | 0.30 | **0.14** |
| Sondanés | 0.07 | 0.05 | 0.29 | **0.14** |

La conclusión general de estas comparaciones es que los resultados del macro-indicador *tráfico* exageran los valores para las lenguas occidentales y dentro de estas lenguas los del inglés[[49]](#footnote-49). Esa realidad se debe poner en perspectiva cuando se utilizan los resultados. ¿De qué orden de magnitud es el sesgo en contra de lenguas no occidentales?

Para tener una idea, tenemos que comprobar W3Techs que está utilizando un método de manera sistemática un algoritmo de identificación de lengua aplicado a los 10 millones de sitios mejor clasificados según Alexa y reflejará de forma natural estos sesgos en su resultado[[50]](#footnote-50). En realidad, W3Techs no sólo reflejará estos sesgos, pero también a veces los amplificara o a veces los deformará ya que otros factores se suman:

* los algoritmos de reconocimiento de lengua también tienen un sesgo a favor del inglés, aunque esto podría ser un factor marginal de alrededor del 10%[[51]](#footnote-51);
* la elección de la detección de lengua en la página principal de los sitios web, la cual a menudo es multilingüe, con el inglés como lengua adicional, incrementará sustancialmente la puntuación del inglés (podría llegar a una sobreestimación el orden de 1/3);
* la consideración no efectiva del multilingüismo.

La determinación para saber cuál de los factores es más influente es un ejercicio muy difícil. Sin embargo, hay un "truco" que permitirá entender la situación en cierta medida. En estudios realizados por FUNREDES y la Unión Latina, en el período 1998-2007, el macro-indicador *productividad* (relación entre el porcentaje de contenido en línea para una lengua y el porcentaje de personas conectadas por esta lengua) había sido estudiado y supervisado durante un largo periodo. Este indicador de la propensión a la creación de contenidos por lengua había mostrado cierta estabilidad y consistencia en el tiempo con valores muy claramente centrados alrededor de 1. Las diferencias estaban en el orden de más o menos 50% dependiendo de si la lengua fue objeto de un gran esfuerzo de producción de contenidos (1.5) o por el contrario si había debilidad en esa dirección (0.5).

Estos son los valores de este indicador en 2005 y 2008 (fuente FUNREDES [[52]](#footnote-52)):

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Francés** | **Inglés** | **Español** | **Portugués** | **Alemán** | **Italiano** | **Rumano** | **Catalán** |
| **2005** | 1.14 | 1.57 | 0.66 | 0.55 | 1.06 | 0,81 | 0.70 |  |
| **2008** | 0.87 | 1.42 | 0.43 | 0.34 | 1.16 | 0.98 | 0.66 | 0.74 |

Se ve que el indicador permanece en la ventana de 0,3 a 1,5. Aquí hacemos la misma comparación, usando los datos de la UIT (convertidos en términos de lengua) y los de W3Techs para contenidos.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Francés** | **Inglés** | **Español** | **Portugués** | **Alemán** | **Italiano** | **Rumano** | **Catalán** |
| **2017** | 0,71 | 2.34 | 0.53 | 0.62 | 1.72 | 2.41 | 0.92 | 0.44 |

Los resultados para inglés, alemán e italiano no son creíbles. Muestran una gran sobreestimación de los valores asignados por W3Techs a los contenidos en esas lenguas. Las siguientes dos tablas muestran los mismos indicadores calculados a partir de W3Techs para una muestra de otras lenguas.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Chino** | **Hindi** | **Árabe** | **Malayo** | **Urdu** | **Bengalí** | **Coreano** | **Vietnamita** | **Tailandés** |
| **2017** | 0.09 | 0.02 | 0,19 | 0.23 | 0.05 | 0.07 | 0.57 | 0.45 | 0.36 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Ruso** | **Japonés** | **Polaco** | **Hebreo** | **Húngaro** | **Sueco** | **Turco** | **Ucranio** | **Checo** |
| **2017** | 1.27 | 1.60 | 2.00 | 1.65 | 1.59 | 2.2 | 1.58 | 0.17 | 3.22 |

Es de notar que la subestimación de ciertas lenguas alcanza proporciones gigantescas (chino, hindi, urdu, bengalí), lo que invalida por completo las cifras presentadas por W3Techs para estas lenguas, o demasiado grandes para ser creíbles (árabe, malayo, ucraniano, tailandés), mientras que otras están claramente sobreestimadas (italiano, checo, sueco, polaco). Cifras inferiores a 0,1 muestran que el déficit es gigantesco y absurdo; cifras entre 0,1 y 0,3 muestran que es enorme y no es creíble, finalmente, los valores superiores a 2 muestran un superávit difícil de creer.

Estas comparaciones muestran que W3Techs no es creíble como fuente de medición de contenidos. Esto no es para dudar de la capacidad de la empresa para llevar a cabo estas acciones, pero entender que intervienen dos sesgos cumulados en los resultados:

* El sesgo específico a la muestra de Alexa que es utilizado por W3Techs para apoyar las medidas.
* El sesgo del método de W3Techs que favorece claramente el inglés mediante la medición de las páginas de inicio y no teniendo en cuenta el multilingüismo.

El análisis de los datos muestra que si un sesgo notable propio a W3Techs existe y debe inflar los resultados del inglés en una cierta proporción (no tomar en cuenta los sitios multilingües y trabajar a partir de las páginas de entrada, puede provocar una sobreestimación del inglés del orden de 30%), el sesgo principal, de lejos, es inducido por Alexa. ¡Hasta cierto punto, W3Techs es un revelador del sesgo de Alexa! De hecho, parece que hay países cuya presencia en la base de datos de Alexa se subestima en un factor de aproximadamente el 500%: China e India juntas representan más de mil millones de personas conectadas a la Internet (es decir, casi un tercio de los usuarios de la Internet), cómo podrían tener juntos sólo un 6,6% de los contenidos de la Internet, según lo indicado por W3Techs?

### 5.4.7 La corrección de los sesgos de W3Techs

Hay una técnica que se utilizó con frecuencia en los primeros estudios de FUNREDES / Unión Latina, en el período 1998-2007, para determinar el peso del inglés[[53]](#footnote-53). Fue de tratar de obtener resultados coherentes de los indicadores para el resto de las lenguas y a partir de ahí se fijaba todos los resultados. Si tenemos en cuenta las primeros 15 lenguas clasificadas en este estudio, el porcentaje de personas conectadas del resto de las lenguas es de aproximadamente de 35%. Para este conjunto de lenguas W3Techs atribuye un resto a la altura de un poco más del 10% de los contenidos mundial, lo que no es consistente. Partiendo de la suposición lógica de una fuerte sobreestimación del contenido en inglés por W3Techs (medido en 53%), la fuerte subestimación de algunas lenguas asiáticas y del indicador para el resto de las lenguas, se puede tratar de hacer una aproximación razonable... aunque puramente especulativa, de lo que es la cifra más probable para la distribución de contenidos por lengua en la Internet; sería un intento manual de corrección de los sesgos basado en el conjunto de observaciones y en el método de cuadrar un resto coherente.

La siguiente tabla proporciona los siguientes datos por columna:

1: Indicador de contenidos

2: Porcentaje de personas conectadas

3: Cifra propuesta para los contenidos por W3Techs

4: Relación entre datos de W3Techs y personas conectadas (calcula como 3 sobre 2)

5: Capacidad medida (1 sobre 7)

6: Capacidad especulativa (8 sobre 7)

7: Porcentaje de la población mundial

8: Especulaciones sobre el porcentaje de contenido

9: Productividad de contenidos resultante de la especulación (8 sobre 2)

10: Especulación sobre el porcentaje de contenidos presentado sobre la base de 100%

TABLA 16 : Clasificación especulativa

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|  |  |  | % |  | **W3TECHS/** |  |  |  | **CONTENUS** | **PRODUCT.** | **CONTENUS** |
|  |  | **PODER** | **CONECT..** |  | **CO\NECT.** | **CAP.** | **CAP.** | **POP.** | **ESPEC.** | **CONT** | **ESPEC.** |
|  |  | **L1+L2** | **L1+L2** | **W3T** | **L1+L2** | **L1+L2** | **Espec.** | **MUNDIAL** | **L1+L2** | **L1+L2** | **/100%** |
| 1 | inglés | 0,404 | 0,226 | 0,519 | 2,299 | 3.74 | 3.33 | 0,108 | **0,360** | 1,59 | 29% |
| 2 | chino | 0,140 | 0,209 | 0,020 | 0,096 | 0.92 | 1.18 | 0152 | **0,180** | 0,86 | 14% |
| 3 | español | 0,093 | 0,093 | 0,051 | 0,551 | 1.56 | 1.35 | 0,059 | **0,080** | 0,86 | 6% |
| 4 | francés | 0076 | 0,056 | 0,041 | 0,726 | 2.93 | 2.31 | 0,026 | 0,060 | 1,06 | 5% |
| 5 | ruso | 0,047 | 0,050 | 0,065 | 1,288 | 1.71 | 1.45 | 0028 | **0,040** | 0,79 | 3,2% |
| 6 | portugués | 0,038 | 0,041 | 0,026 | 0,641 | 1.49 | 1.36 | 0,026 | **0,035** | 0,86 | 2,8% |
| 6 | alemán | 0,046 | 0,032 | 0,055 | 1,723 | 3.31 | 2.53 | 0,014 | **0,035** | 1,10 | 2,8% |
| 6 | japonés | 0,033 | 0,034 | 0,056 | 1,633 | 2.34 | 2.47 | 0,014 | **0,035** | 1,02 | 2,8% |
| 9 | hindi | 0,027 | 0,039 | 0,001 | 0,023 | 0.48 | 0.54 | 0,056 | **0,030** | 0,76 | 2,4% |
| 9 | árabe | 0,032 | 0,042 | 0,007 | 0,165 | 0.86 | 0.79 | 0,038 | **0,030** | 0,71 | 2,4% |
| 11 | malayo | 0,021 | 0,027 | 0,060 | 2,262 | 0.80 | 0,77 | 0,026 | **0,020** | 0,75 | 1,6% |
| 12 | coreano | 0,013 | 0,014 | 0,009 | 0,637 | 1.41 | 1.69 | 0,009 | **0,015** | 1,06 | 1,2% |
| 12 | urdu | 0,014 | 0,020 | 0,001 | 0,044 | 0.51 | 0.55 | 0,027 | **0,015** | 0,74 | 1,2% |
| 14 | italiano | 0,016 | 0,009 | 0,023 | 2,515 | 3.22 | 1.97 | 0,005 | **0,010** | 1,09 | 0,8% |
| 15 | bengalí | 0,008 | 0,013 | 0,001 | 0,069 | 0.29 | 0.37 | 0,027 | **0,010** | 0,76 | 0,8% |
| 16 | polaco | 0,012 | 0,009 | 0,017 | 1,996 | 2.55 | 1.94 | 0,005 | **0,009** | 1,06 | 0,7% |
|  | RESTO | 0,230 | 0,335 | 0,048 | 0,144 | 0.48 | 0.45 | 0,630 | **0,286** | **0,85** | 23% |
|  | TOTAL | 1,250 | 1,250 | 1,000 |  |  |  | 1,2500 | 1.2500 |  | 100% |

Vemos en esta tabla que en las cifras especulativas la productividad del indicador de contenidos se ha reducido a valores coherentes para el resto de lenguas. La especulación se basa precisamente en un equilibrio razonable entre la productividad de contenido para el resto de las lenguas, y el mantener las lenguas tratadas dentro de una ventana razonable entre 0,7 y 1,3 (con excepción del inglés, que puede alcanzar un valor más alto). Hay una bastante fuerte sensibilidad de estos dos factores; por lo que si pasamos el inglés en un 30% (sin cambiar el resto) la productividad del resto de lenguas pasa a ser superior a 1 lo que no se corresponde con lo que se puede esperar; si empujamos el inglés hasta el 40%, entonces la productividad del inglés llegaría a 1,8. Un valor especulativo de un tercio de la proporción de contenidos en inglés daría una productividad para el resto por encima de 0.9 lo que es demasiado importante. Por lo tanto, la especulación se ajustó a este valor para el inglés: 36%.

La escasez de fuentes de información sobre las lenguas en la Internet ha creado una situación alarmante en el que las únicas fuentes existentes no son fiables, ¡pero son referenciadas por una gran cantidad de interlocutores serios que construyen inferencias distorsionadas por datos sesgados con una sobreestimación del papel del inglés en la Internet**[[54]](#footnote-54)**!

### 5.4.8 Limitaciones/sesgos relacionados con el grado de localidad de las fuentes

Muchos usuarios de la Internet en los países occidentales podrían pensar que las aplicaciones famosas como Wikipedia, Facebook, Google, Twitter o YouTube son estándares internacionales cuya fama y utilización es igual en todos los países del mundo. La realidad es muy diferente: algunos países (como China o Rusia) han desarrollado aplicaciones alternativas que drenan una gran parte del tráfico potencial nacional para este conjunto de aplicaciones grandes de la Internet. Este es el caso por ejemplo de Vkontakte que sustituye a Facebook en Rusia o Badiu que sustituye a Google en China. Si pudiéramos asignar una calificación cualitativa de "globalidad" (o por el contrario grado de "localidad") para aplicaciones en la Internet vamos a ser conscientes de que este criterio puede variar significativamente y casi nunca alcanzará su horizonte de 100%. Entre las aplicaciones más "global"[[55]](#footnote-55) se encuentra probablemente Twitter y sin duda Wikipedia (y no es casualidad que esta última es la de mayor diversidad lingüística de lejos) y tal vez Google, con algunas reservas, pero esto es menos el caso de Facebook, YouTube, Linkedin, incluso si estas aplicaciones son todavía mucho más "global" que sus competidores asiáticos Vkontakte o Badiu.

En este contexto, ¿cuál sería la validez del análisis utilizando diferentes sitios de Alexa? Si bien es claro que cada indicador de país para una aplicación o un área determinada debe ser tomado con cuidado para evitar que la calificación de una lengua sea determinada por un alto grado de localidad de los sitios seleccionados, no es menos cierto que la multiplicación de sitios (más de 300) proporciona, gracias a la media truncada, un medio de neutralización de tales efectos. El suavizado de la curva también se logra por la técnica de extrapolación que proporcionará valores para los países que faltan en la fuente y que funciona con respecto a criterios globales.

En este sentido, es interesante identificar los sitios que son fuertemente local para una lengua en la selección y comprobar 1) si la distribución de estas lenguas no es demasiado sesgada y 2) si la media truncada al 20% corrige el sesgo de la selección.

La siguiente tabla cuantifica, por cada lengua, el sesgo de localidad en la muestra de sitios utilizados para la medición del tráfico.

TABLA 16 : Distribución de sitios web con fuerte localidad

|  |  |
| --- | --- |
| LENGUA | Número de sitios |
| Alemán | 2 |
| Árabe | 2 |
| Chino | 10 |
| Coreano | 3 |
| Español | 9 |
| Francés | 12 |
| Hindi | 6 |
| Inglés | 25 |
| Portugués | 3 |
| Ruso | 4 |

El *Anexo IV* enumera los sitios fuertemente locales de las principales lenguas. Es de notar que la media truncada al 20% eliminará, a partir de la muestra de 316 sitios web, cerca de 30 de las puntuaciones más bajas y igual número de las más altas para cada lengua, lo que cubre ampliamente los valores expuestos en la tabla precedente.

Queda por comprobar la sensibilidad de este factor en los resultados. Parece ser que la sensibilidad es muy alta si se utiliza la media simple o la mediana. Por ejemplo, el valor medio del indicador de tráfico para el francés es 10,31 y con ese valor el macro-indicador *poder* del francés sería 8,07. Si eliminamos el sitio Fnac.com de la lista a continuación la media (de tráfico) se reduce a 10,00; si también eliminamos Gameblog.com, la media baja a 9,69; si también eliminamos Fun-mooc.fr, Openclassrooms.com, Qwant.com, Skyrock.com, Theses.fr y Viadeo.com la media del indicador *tráfico* se reduce a 8,09 y el macro-indicador *poder* disminuye a 7,70. Este análisis demuestra que ni la media ni la mediana son eficaces y hay una necesidad de descartar los valores extremos. Este análisis condujo a la solución finalmente adoptada para el indicador de tráfico: la media truncada en 20%.

Esta situación afecta aún más en el esfuerzo de diferenciación temática; por esta razón este enfoque sólo se aplicará a los datos de tráfico que son los más numerosos, y los resultados no se considerarán en el caso de temática con muy pocos micro-indicadores, ya que sólo reflejaría el sesgo de selección de los sitios.

### 5.4.9 Sobre el principio de ponderación

Si bien el principio de ponderación es fácil de entender para los conceptos tales como contenidos, usos, tráfico y especialmente las personas conectadas puede llegar a ser bastante oscuro o misterioso, cuando se aplica a las calificaciones de 0 a 100, como en el caso de los índices o a porcentajes para un total de aplicaciones, como en el caso de interfaces. Esto merece una explicación.

¿Por qué mecanismo una puntuación de 0 a 100 atribuido a una lengua como el resultado de una evaluación de un país en un criterio determinado se pueden convertir en porcentaje mundial y cuál es el sentido de este nuevo valor? ¿Por qué mecanismo un porcentaje que representa el número de veces que una lengua está presente en 24 interfaces para aplicaciones se puede convertir en un porcentaje mundial y cuál es el significado de este porcentaje?

El mecanismo y el significado son los mismos en ambos casos. El mecanismo es el de la ponderación de los valores obtenidos para los países por el porcentaje de personas conectadas por país. Esto se calcula de esta manera:

Pinter(j) = Tg x INTER(j) x UIT(j) / Po

i=L

con Po = ∑ UIT (i) x INTER(i) (factor de ponderación)

i=1

donde:

Pinter(j) es el peso del indicador del interfaz calculado para la lengua j

Tg es la tasa global del multilingüismo (para normalizar el total de L1 + L2)

INTER (j) es el porcentaje encontrado de lenguaje j en relación con interfaces

UIT (i) es el porcentaje de personas con acceso a la Internet para la lengua i.

L es el número total de lenguas (incluyendo el resto).

El sentido de este porcentaje mundial es de traer el valor del porcentaje de las interfaces en una lengua dada hacia el porcentaje de personas conectadas a la Internet a esta lengua, con variaciones en función de la distribución de los porcentajes de interfaz para todas las lenguas. Si ese valor es exactamente el promedio ponderado de las notas de índices (o porcentajes para las interfaces), el resultado sería idéntico al porcentaje de personas conectadas. Si este valor es mucho mayor que este promedio, entonces el resultado aumentará; de lo contrario, se reducirá. Se debe entender que el efecto no sólo depende de la nota (o porcentaje) sino de la distribución entera de los valores del porcentaje de personas conectadas de todas las lenguas.

Para desarrollar una aproximación más intuitiva de este mecanismo, lo mejor es simular la situación en el caso más simple de las interfaces. El francés está presente en 23 de los 24 elementos medidos dándole un porcentaje de 23/24 = 95,83%, lo que va empujar por arriba su porcentaje mundial ponderado. De hecho, este micro-indicador lleva el valor de las personas conectadas de lengua francesa (5,64%) a un valor de 7,37%. La simulación jugando con diferentes valores del porcentaje de presencia en las interfaces permite sentir intuitivamente lo que está pasando con el indicador con este método y como impacta los macro-indicadores.

TABLA 17 : Simulación para interfaces

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Número de interfaces** | 0 | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 23 | 24 |
| **Porcentaje de**  **las interfaces** | 0% | 12,5% | 25% | 37,5% | 50% | 62,5% | 75% | 87,5% | 95,8% | 100% |
| **Valor**  **calculado por ponderación** | 0% | 1.01% | 2.01% | 2.99% | 3,96% | 4,9% | 5,8% | 6,8% | 7,4% | 7,7% |
| **Cuota de poder** | 6,4% | 6,6% | 6,7% | 6,9% | 7,1% | 7,2% | 7,4% | 7,5% | 7,6% | 7,7% |

Por lo tanto, si no hay ninguna interfaz en francés entre los 24 de la muestra, el indicador de "interfaz" tomaría el valor de 0% y afectaría el valor de poder que se reduciría de un 1,2%. Por el contrario, si las 24 interfaces existieran en francés (100%), el indicador interfaz gana 0,3% y el poder gana 0,1%. Se ve que el valor de 5,6% de las personas francoparlantes conectadas es muy cercano al valor de 17 para el número de interfaz en francés y por lo tanto ahí está el umbral de equilibrio (que depende de la distribución de estos valores) que va mover el indicador por encima o por debajo del valor de referencia del porcentaje de hablantes de francés conectados.

El mismo ejercicio con el indicador de e.gobierno da los siguientes resultados.

TABLA 18 : Simulación para índice

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Clasificación de Francia en el índice** | 0 | 25 | 50 | 75 | 96.1 | 100 |
| **Clasificación del francés** | 19 | 38 | 57 | 76 | 91 | 94 |
| **Porcentaje computado para el francés en el índice e.gov** | 0,18% | 3,5% | 5,1% | 6,7% | 8,0% | 8,3% |

El efecto sobre el valor del poder es bastante marginal, ya que es uno de los 5 micro-indicadores cuya media da la nota para el indicador de índice; con una puntuación de 50, la potencia se reduce a 7,5% y con una puntuación de 0 a 7,4%. En este caso, el umbral de equilibrio está cerca de 57. Estos dos ejemplos muestran de manera intuitiva cómo funciona el mecanismo de ponderación y cuál es su sensibilidad a las condiciones iniciales.

# Conclusiones y perspectivas

El seguimiento y la mejora de este método pragmático de medición de las lenguas en la Internet requiere, en el mejor caso de:

* una alianza entre MAAYA, Ethnologue y la UIT para sincronizar una observación anual después de recibir actualizaciones de la UIT y Ethnologue;
* la ampliación de la lista de micro-indicadores y el seguimiento continuo de la disponibilidad de nuevas opciones;
* la tarea de separar los datos L2 por país, o al menos por grupo de países, a fin de mejorar el paso a los indicadores L1+L2;
* la investigación de micro-indicadores adicionales que podrían servir a la delicada cuestión de los contenidos;
* la profundización de los controles y las comparaciones en el lado del modelo;
* la sistematización de la captura de datos de fuentes por país y lengua el modelo de manera a reducir los riesgos de errores;
* la migración de la implementación del modelo de Excel a APL que sería el lenguaje de programación ideal para esas operaciones matriciales.

Este método "artesanal" no anula en modo alguno las ambiciones mantenidas por el proyecto DILINET para dar respuestas a una amplia gama de situaciones y áreas relacionadas con las lenguas en el ciberespacio. Sin embargo, ese modelo proporciona una aproximación razonable a una importante cuestión de la Internet que ha sido totalmente abandonado en manos del sector privado, con la enorme desventaja de la pérdida de dominio de un problema societal importante. En los últimos años la obtención de datos sobre las lenguas en la Internet fue el resultado de la convergencia de intereses privados, pocos transparentes sobre los métodos utilizados, obligando a las personas interesadas ​​en estos datos a aceptarlos los ojos cerrados. Es importante y urgente que el mundo de la investigación a recuperar el control de un sujeto cuyas implicaciones no sólo son comerciales y cuyas apuestas para la diversidad son esenciales.

# Bibliografía

[1] - D. Pimienta, D. Prado, "Étude sur la place des langues de France dans l'Internet", Ministère de la culture 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>

[2] - "Le français dans l'Internet", Rapport 2014 "La langue française dans le monde", pp. 501-541, OIF, Nathan, 2014 - <http://www.francophonie.org/Rapports-Publications.html>

[3] - D. Pimienta, D. Prado, "Exploring the status of languages ​​of France in the Internet: methods and reflection as possible of approaches for --other groups of languages", Proc. of third international conference on linguistic and cultural diversity in cyberspace, IFAP/UNESCO,Yakutsk, 7/2014

[http://mcbs.ru/files/2015/hmao\_2015/yak\_mling\_2015.pdf#page=140](http://mcbs.ru/files/2015/hmao_2015/yak_mling_2015.pdf%23page=140)

[4] - D. Pimienta, Chapter “Language and content” in “Accelerating Development Using the Web: Empowering Poor and Marginalized Populations”, W3C Foundation, 2012

http://g3ict.org/download/p/fileId\_928/productId\_241

[5] - MAAYA, “NET.LANG: vers un cyberespace multilingue”, C&F Editions, 2012 - <http://net-lang.net>

[6] - D. Pimienta, D. Prado, Á. Blanco, Douze années de mesure de la diversité linguistique sur l’Internet : bilan et perspectives, UNESCO, 2009 -

<http://unesdoc.unesco.org/images/0018/001870/187016f.pdf>.

[7] - D. Pimienta, Chapter “Target 9 contents and languages" in Measuring the Information Society: the ICT Development Index, ITU, ISBN 92-61-12831-9, 2009.

<http://www.itu.int/ITU-D/ict/publications/idi/2009/index.html>

[8] - J. Paolillo, D. Pimienta, D. Prado, et al. (2005), Mesurer la diversité linguistique sur Internet, UNESCO, 12/2005- <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>.

[9] - I. Suzuki, Y. Mikami, et al. (2002), “A Language and Character Set Determination Method Based on N-gram Statistics”, in ACM Transactions on Asian Language Information Processing, Vol.1, No.3, 9/2002, pp.270-279.

[10] - D. Pimienta, D. Prado, "Medición de la presencia de la lengua española en la Internet: métodos y resultados", en Revista Española de Documentación Científica 39 (3), julio-septiembre 2016, e141

ISSN-L:0210-0614. doi: <http://dx.doi.org/10.3989/redc.2016.3.1328>

# Anexo I. Lista de micro-indicadores

|  |
| --- |
| **Micro-indicadores USOS** |
| Penetración 3G |
| Suscripciones móviles |
| Video-llamadas mediante dispositivo móvil |
| Video-llamadas mediante dispositivo móvil - otra fuente |
| Video-llamadas mediante dispositivo móvil - Según Statista |
| Cuentas de banda ancha fija |
| Cuentas de banda ancha móvil |
| Coste de banda ancha |
| Hogares conectados a la Internet |
| Cajeros Bitcoin |
| Infecciones de PC |
| Teléfonos fijos |
| Mercado e.Commerce |
| Mensajería instantánea a través de móvil |
| Banca electrónica |
| E-Wallets |
| Peticiones formales a Google (datos) |
| Peticiones formales a Google (Borrado) |
| Servidores Internet seguros |
| Teléfonos inteligentes |
| Descargas OpenOffice |
| Telefonía IP mediante dispositivo móvil |
| Tráfico móvil |
| Twitter a través de móvil |
| Uso servidores por sitios Web |
| Usuarios de Facebook |
| Usuarios de Facebook según Owloo |
| Usuarios de Facebook según Statista |
| Usuarios de Linkedin |
| Usuarios Netflix- según Netflix |
| Usuarios Netflix- según Statista |
| Usuarios de redes sociales |
| Usuarios de redes sociales (proyección 2021) |
| Usuarios de Twitter |
| Usuarios de YouTube |
| Usuarios de YouTube según Alexa |
| Usuarios de YouTube según Statista |
| Usuarios de Google |

|  |
| --- |
| **Micro-indicadores ÍNDICES** |
| Indicador de Acceso Universal |
| e.gov indicador |
| e.participation indicador |
| WebIndex-macro-indicador |
| Infraestructura macro-indicador |

|  |
| --- |
| **Micro-indicadores INTERFAZ** |
| Android |
| Apple |
| Lenguas de traducción - Bing |
| Interfaz Cortana- |
| Dictionnary.com - lenguas de traducción |
| DMOZ - Lenguas contenido |
| DuckDuckGo - Interfaz |
| Duolingo - lenguas de traducción |
| Facebook - Interfaz |
| FreeTraductor - lenguas de traducción |
| Google - Interfaz |
| Google Académico - Interfaz |
| GoogleTranslate - lenguas de traducción |
| IM Translator - lenguas de traducción |
| Ios |
| Linux |
| On-Line - Lengua de traducción |
| Preferencias Bing |
| Preferencias Google |
| Preferencias Yahoo |
| Reverso -lenguas de traducción |
| SDL - lenguas de traducción |
| Skype - Interfaz |
| Systran - Lenguas de traducción |
| Telegram- interfaz |
| Windows |

|  |
| --- |
| **Micro-indicador CONTENIDO** |
| Amazon EEUU - número de libros |
| W3Techs - contenido |
| Wikilibros - Interfaz |
| Wikilibros - Número de usuarios |
| Wikilibros - número de elementos |
| Wikilibros - número de editores |
| Wikipedia - Editores casuales |
| Wikipedia - Artículos lenguas |
| Wikipedia - Editores confirmados |
| Wikipedia - Interfaz |
| Wikipedia - Profundidad |
| Wikinoticias - Artículos |
| Wikinoticias - Interfaz |
| WikiQuote - Interfaz |
| WikiQuote - Artículos |
| Wikiversidad - Interfaz |
| Wikiversidad - Artículos |
| Wikcionario - Interfaz |
| Wikcionario - Artículos |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Micro-indicador TRÁFICO** |  |  |  |  |
| 1and1.com | Dmoz.org | Jurn.org | Reverbnation.com | Yelp.com |
| 4shared.com | Doaj.org | Justanswer.com | Rumble.com | Youku.com |
| 500px.com | Douban.com | Kaixin001.com | Rutube.ru | Youtube.com |
| A2hosting.com | Draugiem.lv | Kakao.com | Sapo.pt | Zhihu.com |
| Abilogic.com | Dreamhost.com | Kongregate.com | Scielo.org | Zoosk.com |
| Sobre mi | Dreamwidth.org | Last FM | Scienceopen.com |  |
| Academia.edu | Dropbox.com | Librarything.com | Search.com |  |
| Acfun.tv | Drupal.org | Linkedin.com | Secondlife.com |  |
| Adam4Adam.com | Duckduckgo.com | Liquidweb.com | Semanticscholar.org |  |
| Addictinggames.com | DXY.cn | Liveleak.com | Sharecare.com |  |
| Adobe.com | Eclipse.org | Logoslibrary.eu | Similarweb.com |  |
| Adultfriendfinder.com | Edx.org | Mamba.ru | Sitebuilder.com |  |
| Alexa.com | Egnyte.com | Match.com | Skyrock.com |  |
| Alivedirectory.com | Eharmony.com | Mediafire.com | Slideshare.net |  |
| Anastasiadate.com | Etoro.com | Meetic.fr | Smugmug.com |  |
| Angel.co | Europa.eu/ | Meetup.com | Snapchat.com |  |
| Anobii.com | Exalead.com | Mega.nz | Socolar.com |  |
| Answers.com | Excite.com | Mendeley.com | Sogou.com |  |
| Apple.com | Experienceproject.com | Metacafe.com | Somuch.com |  |
| Archive.org | Fetlife.com | Metafilter.com | Sony.com |  |
| Archives-ouvertes.fr | Filefactory.com | Microsoft.com | Soso.com |  |
| Armorgames.com | Fileserve.com | Miniclip.com | Soundcloud.com |  |
| Arvixe.com | Filmaffinity.com | Mixi.jp | Spaces.ru |  |
| Arxiv.org | Filmow.com | Mocospace.com | Spip.net |  |
| Ashleymadison.com | Flickr.com | Moodle.org | Spotify.com |  |
| Ask.com | Flipboard.com | Mouthshut.com | Squarespace.com |  |
| Ask.fm | Flixster.com | Mozilla.org | Stackexchange.com |  |
| Atom.io | Fotki.com | Mubi.com | Startpage.com |  |
| Avvo.com | Fotolog.com | Myheritage.com | Steampowered.com |  |
| Badoo.com | Foursquare.com | Mylife.com | Straightdope.com |  |
| Baidu.com | Fun-mooc.fr | Myspace.com | Stumbleupon.com |  |
| Bandcamp.com | Funnyordie.com | Napster.com | Sublimetext.com |  |
| Base-search.net | Futurelearn.com | Naver.com | Tagged.com |  |
| Beyond.com | G2a.com | Netbeans.org | Taringa.net |  |
| Bilibili.com | Gaiaonline.com | Netcraft.com | Theses.fr |  |
| Bing.com | Gameblog.com | Netflix.com | Tinyurl.com |  |
| Bit.ly | Gamefaqs.com | Netlog.com | Trombi.com |  |
| Bitbucket.org | Geni.com | Newgrounds.com | Tudou.com |  |
| Bitshare.com | Gfycat.com | Nicovideo.jp | Twitch.tv |  |
| Blackle.com | Gigablast.com | Ning.com | Twoo.com |  |
| Bluehost.com | Gigasize.com | Notepad-plus-plus.org | Udacity.com |  |
| Blurtit.com | Girlsaskguys.com | Novoed.com | Udemy.com |  |
| Box.com | Github.com | Oatd.org | Uploaded.net |  |
| Brackets.io | Godaddy.com | Odnoklassniki.ru | Uploading.com |  |
| Business.com | GOG.com | Office.com | Veoh.com |  |
| Busuu.com | Goodreads.com | Okcupid.com | Viadeo.com |  |
| C9.io | Google.com | Openclassrooms.com | Vimeo.com |  |
| Cafemom.com | Gotinder.com | Opengrey.eu | Vine.co |  |
| Cairn.info | Grindr.com | Openlibrary.org | Visualstudio.com |  |
| Care2.com | Hi5.com | Openoffice.org | Vk.com |  |
| Caringbridge.org | Hightail.com | Openthesis.org | Wattpad.com |  |
| Chacha.com | Hostgator.com | Opera.com | Wayn.com |  |
| Chrome.com | Hulu.com | Origin.com | Wdl.org |  |
| Classmates.com | Icloud.com | Paypal.com | Webcrawler.com |  |
| Codeanywhere.com | Infinit.io | Periscope.com | Webometrics.info |  |
| Codepen.io | Influenster.com | Periscope.tv | Weebly.com |  |
| Commonsensemedia.org | Inmotionhosting.com | Photobucket.com | Weheartit.com |  |
| Contentful.com | Instagram.com | Pinterest.com | Wetransfer.com |  |
| Couchsurfing.com | Iqiyi.com | Playstation.com | Wikimedia.org |  |
| Coursera.org | Italki.com | Plurk.com | Wistia.com |  |
| Crunchyroll.com | Itch.io | Qq.com | Wix.com |  |
| Cyworld.com | Ixquick.com | Quora.com | Wolframalpha.com |  |
| Dailymotion.com | Ixquick.eu | Qwant.com | Wordpress.com |  |
| Dart-europe.eu | Jasminedirectory.com | Raptr.com | Worldcat.org |  |
| Daum.net | Jetbrains.com | Ravelry.com | Worldwidescience.org |  |
| Deezer.com | Joinhouse.party | Reddit.com | Xbox.com |  |
| Delicioso | Joomla.com | Rediff.com | Xing.com |  |
| Depositfiles.com | Journalseek.net | Renren.com | Yacy.net |  |
| Deviantart.com | Jstor.org | Researchgate.net | Yahoo.com |  |

# Anexo II. Fuentes seleccionadas

|  |  |  |
| --- | --- | --- |
| **NOMBRE DE LA FUENTE** | **DIRECCIÓN (HOME)** | **NÚMERO DE INDIC. PREVISTO** |
| Alexa. datos sobre tráfico hacia sitios web[[56]](#footnote-56) | http://www.alexa.com/ | 310 |
| Informe de la Comisión de banda ancha 2016. Datos sobre conectividad y acceso a la información | [http://broadbandcommission.org/](http://www.broadbandcommission.org/Documents/reports/bb-annualreport2016.pdf) | 5 |
| CIA. Datos demográficos y acceso a la información | [https://www.cia.gov/](https://www.cia.gov/library/publications/the-world-factbook/fields/2150.html#xx) | 1 |
| Owloo. Porcentaje de uso de Facebook | https://www.Owloo.com/ | 1 |
| Social Progress Imperative. | http://socialprogressimperative.org/ | 1 |
| Statista. Datos sobre consulta de sitios[[57]](#footnote-57) | https://www.statista.com/statistics/ | 31 |
| Apache OpenOffice. Número de descargas de OpenOffice | http://openoffice.org/stats/countries | 1 |
| Internet World Stats. Datos sobre el número de usuarios por país e lengua | [http://www.Internetworldstats.com/](http://www.internetworldstats.com/) | 2 |
| Translated. Índice sobre interés económico por lengua | <http://www.translated.net/> | 1 |
| W3Techs. contenidos por lengua | https://W3Techs.com/ | 1 |
| WebIndexData 2014. Índices sobre Sociedad de la Información | <http://thewebindex.org/> | 9 |
| Wikipedia. Demografía y demo-lingüística | <https://wikipedia.org/> | 2 |
| Amazon. Datos sobre Amazon | https://www.amazon.com/ | 2 |
| Wikimedia.Datos sobre los diversos wikis del universo Wikimedia | <https://stats.wikimedia.org/> | 11 |
| Proyecto Joshua. datos demo-lingüísticos y demográficos | http://legacy.joshuaproject.net/ | 1 |
| Etnólogo. datos demo-lingüísticos. | https://www.ethnologue.com/ | 1 |

# Anexo III: Valores elegidos para L2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **código ISO** | **Lengua** | **L2** | **L1** | **L1 + L2 / L1** |
| awa | Awadhi | **45 400** | 11 821 647 | 1,0038 |
| ben | Bengalí | **19 202 880** | 248 094 781 | 1,0774 |
| bho | Bhojpuri | **160 000** | 48 656 324 | 1,0033 |
| ces | Checo | **2540000** | 10 121 000 | 1,251 |
| deu | Alemán | **52 689 000** | 69 419 000 | 1,759 |
| ell | Griego | **57 000** | 11 455 800 | 1,005 |
| eng | Inglés | **611 563 010** | 344 179 157 | 2,7769 |
| fas | Persa | **18 434 114** | 43 495 306 | 1,4238 |
| fra | Francés | **153 485 770** | 76 660 700 | 3,0021 |
| hau | Hausa | **19 500 000** | 45 494 090 | 1,4286 |
| hin | Hindi | **121 230 000** | 374 692 040 | 1,3235 |
| ita | Italiano | **3 085 000** | 41769700 | 1,0739 |
| jpn | Japonés | **11 500** | 125 376 800 | 1,0001 |
| kan | Kannada | **9 000 000** | 43 729 877 | 1,2058 |
| khm | Khmer | **1 000 000** | 16 030 000 | 1,0624 |
| mar | Marathi | **3 000 000** | 82 230 980 | 1,0365 |
| msa | Malayo | **175 200 000** | 53 798 640 | 4,2566 |
| mya | Birmano | **10 000 000** | 31 704 180 | 1,3154 |
| orm | Oromo | **170 000** | 33 129 700 | 1,0051 |
| pol | Polaco | **454 000** | 40 479 300 | 1,0112 |
| por | Portugués | **11 180 000** | 217 192 370 | 1,0515 |
| rus | Ruso | **113 273 820** | 130 013 300 | 1,8712 |
| sin | Cingalés | **2 000 000** | 11 838 680 | 1,1689 |
| som | Somalí | **95 600** | 19 801 000 | 1,0048 |
| spa | Español | **91 308 400** | 434 569 400 | 1,2101 |
| tam | Tamil | **8 000 000** | 74 551 992 | 1,1073 |
| tel | Telugu | **5 000 000** | 87 584 764 | 1,0571 |
| tha | Tailandés | **40 000 000** | 28 549 500 | 2,4011 |
| tur | Turco | **380 300** | 64 175 300 | 1,0059 |
| urd | Urdu | **94 045 800** | 99 137 906 | 1,9486 |
| yor | Yoruba | **2 000 000** | 38 104 300 | 1,0525 |
| zha | Zhuang | **20 000** | 16 305 000 | 1,0012 |
| zul | Zulú | **15 700 000** | 12 159 200 | 2,2912 |
| **zzz** | **RESTO** | **236 200 000** | **891 034 646** | **1,2651** |
|  | **TOTAL** | **1824031594** | **7296012569** | **1,2500** |

notas:

- La cifra de malayo incluye indonesio

- Valores de L1 para el chino siendo mayores que los valores L1+L2 indicados por Ethnologue se dejó a 0 L2 para los chinos.

- Por precaución se registró un valor de 1.265 para el resto de lenguas: el efecto marginal sobre los cálculos es de evitar empujar demasiado hacia arriba los resultados de las 140 lenguas seleccionadas. El valor fue finalmente elegido para permitir un redondeado a aproximativamente 25% del efecto del multilingüismo.

# Anexo IV: Sitios totalmente locales

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SITIO** | **lengua dominante** | **% lengua**  **dominante** |  | **SITIO** | **lengua dominante** | **% lengua**  **dominante** |
| Gmx.net | Alemán | 57 |  | Filmaffinity.com | Español | 72 |
| Xing.com | Alemán | 53 |  | Fotolog.com | Español | 52 |
| Adam4adam.com | | Inglés | 50 |  | Gigasize.com | Español  Hindi | 21  17 |
| Adictinggames.com | | Inglés | 53 |  | Megaupload.com | Español | 21 |
| Aim.com | | Inglés | 62 |  | Rapidshare.com | Español | 17 |
| Amazon.com | | Inglés | 50 |  | Taringa.net | Español | 78 |
| Avvo.com | | Inglés | 67 |  | Tuenti.com | Español | 65 |
| Beyond.com | | Inglés | 68 |  | Archives-ouvertes.fr | Francés | 21 |
| Cafemom.com | | Inglés | 51 |  | Bnf.fr | Francés | 45 |
| Caringbridge.org | | Inglés | 68 |  | Cairn.info | Francés | 41 |
| Chacha.com | | Inglés | 48 |  | Deezer.com | Francés | 36 |
| Classmates.com | | Inglés | 67 |  | Exalead.com | Francés | 36 |
| Commonsensemedia.org | | Inglés | 59 |  | Fnac.com | Francés | 70 |
| Crunchyroll.com | | Inglés | 50 |  | Fun-MOOC .com | Francés | 48 |
| Eharmony.com | | Inglés | 62 |  | Gameblog.com | Francés | 69 |
| Excite.com | | Inglés | 53 |  | Openclassrooms.com | Francés | 52 |
| Fetlife.com | | Inglés | 52 |  | Qwant.com | Francés | 68 |
| Flixter.com | | Inglés | 50 |  | Skyrock.com | Francés | 47 |
| Gaiaonline.com | | Inglés | 56 |  | Theses.fr | Francés | 51 |
| Gfycat.com | | Inglés | 55 |  | Viadeo.com | Francés | 51 |
| Influenster.com | | Inglés | 64 |  | Yacy.com | Francés | 34 |
| Joinhouse.party | | Inglés | 57 |  | Mouthshut.com | Hindi | 43 |
| Justanswers.com | | Inglés | 53 |  | Novoed.com | Hindi | 14 |
| Match.com | | Inglés | 53 |  | Rediff.com | Hindi | 40 |
| Metafilter.com | | Inglés | 54 |  | Rediffmail.com | Hindi | 44 |
| Mocospace.com | | Inglés | 55 |  | Somuch.com | Hindi | 36 |
| Mylife.com | | Inglés | 66 |  | Anobii.com | Italiano | 38 |
| Oovoo.com | | Inglés | 51 |  | Nicovideo.jp | Japonés | 90 |
| Ravelry.com | | Inglés | 57 |  | 4shared.com | Portugués  Árabe  Español | 22  16  11 |
| Rumble.com | | Inglés | 64 |  | Filmow.com | Portugués | 83 |
| Sharecare.com | | Inglés | 51 |  | Sapo.pt | Portugués | 71 |
| Smugmug.com | | Inglés | 54 |  | Scielo.org | Portugués  Español | 52  21 |
| Straightdope.com | | Inglés | 55 |  | Icq.com | Ruso | 31 |
| Telegram.com | | Inglés | 54 |  | Mail.ru | Ruso | 41 |
| Yelp.com | | Inglés | 65 |  | Mamba.ru | Ruso | 43 |
| Jurn.org | | Árabe | 13 |  | Spaces.ru | Ruso | 47 |
| Baidu.com | | Chino | 88 |  | Playstore.com | Turco | 41 |
| Douban.com | | Chino | 85 |  |  |  |  |
| Ikiyi.com | | Chino | 88 |  |  |  |  |
| Kaixin001.com | | Chino | 85 |  |  |  |  |
| Odnosklassniki.ru | | Chino | 75 |  |  |  |  |
| Plurk.com | | Chino | 64 |  |  |  |  |
| Qq.com | | Chino | 79 |  |  |  |  |
| Renren.com | | Chino | 84 |  |  |  |  |
| Sogou.com | | Chino | 92 |  |  |  |  |
| Weibo.com | | Chino | 89 |  |  |  |  |
| Youku.com | | Chino | 89 |  |  |  |  |
| Zihu.com | | Chino | 86 |  |  |  |  |
| Cyword.com | | Coreano | 78 |  |  |  |  |
| Kakao.com | | Coreano | 84 |  |  |  |  |
| Naver.com | | Coreano | 86 |  |  |  |  |

1. En particular, el ambicioso Language Observatory Project (LOP) dirigido por Yoshiki Mikami de la Universidad de Nagaoka. Ver [9]. [↑](#footnote-ref-1)
2. Parece que el tsunami sufrido por Japón en marzo de 2011 desencadeno la parálisis del proyecto LOP. [↑](#footnote-ref-2)
3. Ver [https://web.archive.org/web/20180809042117/http://dilinet.org/](https://web.archive.org/web/20180809042117/http:/dilinet.org/) y <http://funredes.org/lc/dilinet/es/> [↑](#footnote-ref-3)
4. En enero de 2012, convocatoria de propuestas ICT-2011.4.4 Gestión inteligente de la información; en enero de 2013, llamado ICT-2013.4.1 Análisis de contenido y tecnologías del lenguaje - Análisis de contenido cruzado. [↑](#footnote-ref-4)
5. Este análisis se ha realizado para el francés, sin embargo, este proceso debe tratarse con precaución debido a la reducción de datos estadísticos que implica segmentar la muestra de micro-indicadores. [↑](#footnote-ref-5)
6. Por ejemplo, la distribución porcentual mundial de personas conectadas por país se transforma en distribución porcentual mundial por lengua. [↑](#footnote-ref-6)
7. Se prefiere el término "cuota" a "porcentaje" porque la comprensión consistente del multilingüismo implicará que los cálculos ya no se relacionan con la población mundial sino con el total de hablantes por lengua, valor que es igual a la población mundial aumentada por el número de personas bilingües. [↑](#footnote-ref-7)
8. Amazon (<http://www.amazon.com>) es una compañía, líder mundial en ventas de libros en línea y transformada en una plataforma para el comercio electrónico de muchos tipos de productos. [↑](#footnote-ref-8)
9. La **media truncada** es el promedio de los valores del vector después de eliminar los valores extremos más altos y más bajos. Se utiliza como un recurso para reducir considerablemente los sesgos resultando de la selección de fuentes. Por ejemplo, ese método permite reducir, para el indicador *tráfico*, el sesgo debido a la elección de aplicaciones que favorecen ciertos idiomas (ver *7.4.7 Limitaciones / sesgos relacionados con el grado de localidad de las fuentes*). [↑](#footnote-ref-9)
10. Un portal tecnológico para encuestas y estadísticas. <https://w3techs.com/> [↑](#footnote-ref-10)
11. http://thewebindex.org/ [↑](#footnote-ref-11)
12. Las únicas fuentes disponibles con datos para todos los países del mundo son las personas conectadas a la Internet, las cuentas de banda ancha fijas o móviles y el número de descargas de Open Office. [↑](#footnote-ref-12)
13. En inglés : open source, open data, open knowledge. [↑](#footnote-ref-13)
14. Alexa (<http://alexa.com>) es una empresa que proporciona estadísticas y clasificaciones en sitios web basadas en la estimación del tráfico. [↑](#footnote-ref-14)
15. Statista (<http://statista.com>) es un portal generalista de estadísticas que ha sido organizado por ellos. [↑](#footnote-ref-15)
16. Más del 80% de los micro-indicadores se miden con la ayuda de este servicio. [↑](#footnote-ref-16)
17. Un portal para datos demo-lingüísticos. <https://joshuaproject.net/> [↑](#footnote-ref-17)
18. Cuando una cantidad (o un porcentaje mundial) mide los locutores de una lengua (L1 o L2) que tienen acceso a la Internet, se debe entender que si se hace el total para todas las lenguas del mundo será lógicamente superior a la población total de usuarios de la Internet (o si es expresado en porcentaje mundial va a exceder los 100%). Desde luego, la misma persona, si es bilingüe va a aparecer dos veces en esta contabilidad (3 veces si habla 3 idiomas, 4 veces si habla 4 idiomas, etc.). [↑](#footnote-ref-18)
19. También hay que hacer la simplificación de contar los bebes como hablantes de su lengua materna desde su nacimiento, aunque será realmente el caso después de unos años de aprendizaje. [↑](#footnote-ref-19)
20. Un portal para datos demo-lingüísticos acerca de todos los idiomas del mundo (<http://ethnologue.com>), un proyecto de SIL International (<http://sil.org>). [↑](#footnote-ref-20)
21. Esto significa que el 25% de la población mundial tiene un segundo idioma. [↑](#footnote-ref-21)
22. APL, "**A** **P**rogramming **L**anguage" en inglés, es tanto un formalismo matemático como su implementación en forma de lenguaje de programación, diseñado por Kenneth Iverson. Para obtener más detalles, consulte https://es.wikipedia.org/wiki/APL. [↑](#footnote-ref-22)
23. Pero no olvidemos que, si bien Wikipedia es una de las aplicaciones más globalizadas de Internet, no está muy presente en algunos países importantes de la Internet, como China o India, y que ofrece información en un máximo de 300 idiomas. [↑](#footnote-ref-23)
24. El análisis realizado demuestra que W3Techs no puede considerarse una medida confiable, ni mucho menos. [↑](#footnote-ref-24)
25. Obviamente, también podría medir la propensión de los hablantes de otros idiomas a producir contenido en inglés. [↑](#footnote-ref-25)
26. http://thewebindex.org/ [↑](#footnote-ref-26)
27. The Economist acaba de publicar un informe con macro-indicadores para medir el "grado de inclusión" de la Internet (<https://theinclusiveInternet.eiu.com>). La ONG Open Knowledge publica un índice para medir los datos abiertos (<http://index.okfn.org/>). La UNESCO tiene la intención de desarrollar el grado de universalidad de Internet en el país ([http://unesco.org/new/en/media-services/single-view/news/unesco\_call\_for\_proposals\_defining\_the\_Internet\_universality\_in/](http://www.unesco.org/new/en/media-services/single-view/news/unesco_call_for_proposals_defining_the_Internet_universality_in/)). [↑](#footnote-ref-27)
28. Por ejemplo, Wikpedia en chino tendría unos 700,000 artículos, en contraste con las dos opciones chinas de enciclopedia (Baidu, Baike y Hudong) que juntas representan casi 12 millones de artículos, es decir más de 17 veces más (fuente: <https://fr.wikipedia.org/wiki/Wikipédia_en_chinois>). [↑](#footnote-ref-28)
29. Cursos masivos y abiertos en línea. [↑](#footnote-ref-29)
30. Este es un servicio gratuito en línea de Miniwatts Marketing Group, una compañía que ofrece en su sitio web (http://InternetWorldStats.com) estadísticas de población y uso de la Internet interesantes y bien organizadas y algunas cifras para lenguas (http://Internetworldstats.com/stats7.htm) y ha ganado una merecida notoriedad. [↑](#footnote-ref-30)
31. De hecho, dado que los porcentajes de personas conectadas en los principales países francófonos (Francia, Bélgica, Suiza, Canadá) son respectivamente 83%, 85%, 87% y 88% y que estos países representan 115 millones de hablantes del francés, para obtener un promedio ponderado de 26% tal cómo IWS indica, se requiere que los otros 279 millones de hablantes de francés (IWS indica 394 millones como población total de hablantes de francés) tengan menos de 2% de personas conectadas. Eso lo indica la ecuación: 26% = (115M \* 85% + 279M \* x%) / 394M ==> x = 1,7%. Aunque la África francófona sufre de la brecha digital, sus tasas de conectividad por país varían entre una ventana 4-7% (Burundi, Centroáfrica, Chad), pasando por una ventana 10-14% (Benín, Mali, Madagascar, Níger) y muchos con más de 45% (Costa Marfil, Gabón, Marruecos, Túnez) por lo cual esta cifra de 1,7% no tiene sentido. [↑](#footnote-ref-31)
32. Excepto por la decisión de usar la media truncada, otros cambios probablemente afectarían el decimal en una amplitud máxima de +- 0.3% si consideramos solo los métodos de micro-indicadores o agrupación de cálculos y más si las desviaciones en L2 son importantes. [↑](#footnote-ref-32)
33. Esto es especialmente cierto para el orden del alemán y del ruso en términos de *poder* y *capacidad*. [↑](#footnote-ref-33)
34. Online Computer Library Center, una red académica líder que presta servicios a bibliotecarios. [↑](#footnote-ref-34)
35. Ver [6] y [8]. [↑](#footnote-ref-35)
36. Por lo tanto, esta limitación impide sacar conclusiones para las lenguas no nacionales que se hablan dentro de un país. En el caso de Francia, por ejemplo, donde el Ministerio de Cultura está muy interesado en esta cuestión, este estudio no aporta ninguna luz particular y es preferible referirse al estudio apoyado por la Dirección de lenguas de Francia de este Ministerio (ver [1]). [↑](#footnote-ref-36)
37. Los idiomas que tienen el mayor número de países en los que se distribuyen, para aquellos con más de 50 países, en orden: inglés (151 países), árabe (117), francés (105), chino (94) alemán (81) Griego (74), Rusia (70), español (62) italiano (58), hindi (56), armenio (55) y portugués (51). [↑](#footnote-ref-37)
38. Debe quedar claro que a veces la extrapolación se va a aplicar en unos países donde la aplicación no se usa y, en otros casos, va a subestimar los valores reales en algunos países donde la aplicación está muy utilizada. [↑](#footnote-ref-38)
39. En otras palabras, siempre preferir las fuentes primarias. [↑](#footnote-ref-39)
40. <https://publicadministration.un.org/egovkb/en-us/reports/un-e-government-survey-2016> [↑](#footnote-ref-40)
41. [http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals\_la Internet\_2000-2015.xls](http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals_Internet_2000-2015.xls) [↑](#footnote-ref-41)
42. Manual on the Measurement of access of households and individuals to information and communications technology (ICT) and use of these technologies, 2014 (see page 74) - <https://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ITCMEAS-2014-PDF-F.pdf> [↑](#footnote-ref-42)
43. <http://ec.europa.eu/eurostat/statistics-explained/index.php/Information_society_statistics_-_households_and_individuals/fr> [↑](#footnote-ref-43)
44. Por ejemplo, Martinica, Guadalupe, Guyana y Reunión se incluyen en Francia, pero no es el caso de Polinesia Francesa, Nueva Caledonia, Wallis y Futuna y Saint Pierre y Miquelón. [↑](#footnote-ref-44)
45. Alexa no informa con claridad acerca de los "millones de barras de herramientas instaladas" (http://www.alexa.com/about) y un artículo de Wikipedia menciona la cifra de 10 millones en 2005. [↑](#footnote-ref-45)
46. Según los expertos de OpenRoot, ¡esto concierne al 80% de los sitios chinos lo que lleva la subestimación del contenido (o tráfico) chino en un factor de 5! [↑](#footnote-ref-46)
47. Esta declaración proviene de un razonamiento simple y no puede ser respaldada por los datos en ausencia de datos estadísticos en la barra de herramientas; sin embargo, el análisis de los resultados de W3Techs lo confirmará más adelante. [↑](#footnote-ref-47)
48. Con altas puntuaciones de Inglés en el ranking de Alexa, un mero factor de sobreestimación de 1,7 como en el caso de LinkedIn, representan más del 20% extendida sobreestimación sobre otros lenguas, lo que es enorme. [↑](#footnote-ref-48)
49. Y también puede ser para el francés, a juzgar por las pruebas de comparación realizadas. [↑](#footnote-ref-49)
50. <https://W3Techs.com/technologies/overview/content_language/all> [↑](#footnote-ref-50)
51. Marginal ... pero casi siempre a favor del inglés. [↑](#footnote-ref-51)
52. <http://funredes.org/lc> [↑](#footnote-ref-52)
53. Desde luego, el método utilizado en ese tiempo producía resultados en porcentajes relativos al inglés, y la determinación del porcentaje absoluto de los contenidos en inglés (que permitía, en secuencia, deducir los otros porcentajes absolutos) era un paso delicado que podría solamente ser resuelto en aproximación mediante el cruce de diversos datos. [↑](#footnote-ref-53)
54. Véase, como ejemplo típico, el reciente informe de Mozilla sobre la salud de la Internet que presta especial atención a los aspectos lingüísticos, por desgracia, a partir de supuestos sesgados ([\https://Internethealthreport.org/v01/](/https:/Internethealthreport.org/v01/)). [↑](#footnote-ref-54)
55. Aunque un análisis más profundo mostró que hay que matizar esta evaluación debido a que la representación de las principales lenguas asiáticas en Wikipedia es muy por debajo de su presencia demográfica. [↑](#footnote-ref-55)
56. Los datos se pagan. [↑](#footnote-ref-56)
57. Statista fue elegido como la fuente principal pagada después de estudiar las posibles fuentes. [↑](#footnote-ref-57)