C@lea
Cadernos de aulas do LEA
Número Especial - IV SIMC, Março de 2022, p. 4-13

INDICADORES DE LA PRESENCIA DE LAS LENGUAS EN LA INTERNET

Daniel Pimienta*

Resumen: Se requiere la disponibilidad de indicadores del espacio de las lenguas en la Internet para apoyar políticas públicas adecuadas. Las fuentes actuales son escasas y fuertemente sesgadas. Este método permite calcular indicadores para las 140 lenguas con más de 5 millones de hablantes L1. Se basa en la recopilación de un gran conjunto de microindicadores que miden lenguas o países en varios espacios o aplicaciones de la Internet. Se aplican métodos estadísticos para producir 6 indicadores: *usuarios de la Internet, tráfico, usos, contenidos, índices societales e interfaces*, de los cuales se deducen 4 macroindicadores: *potencia, capacidad, gradiente y productividad de contenidos*. Se presentan algunos resultados y se analizan los sesgos de los métodos existentes.

Palabras clave: Lenguas. Internet. Indicadores. Sesgos.

Resumo: São necessários indicadores de espaço linguístico na Internet para apoiar políticas públicas apropriadas. As fontes actuais são escassas e muito tendenciosas. Este método permite calcular indicadores para as 140 línguas com mais de 5 milhões de falantes de L1. Baseia-se na compilação de um grande conjunto de microindicadores que medem línguas ou países em vários espaços ou aplicações da Internet. São aplicados métodos estatísticos para produzir 6 indicadores: utilizadores da Internet, tráfego, utilização, conteúdo, índices sociais e interfaces, dos quais derivam 4 macro-indicadores: potência, capacidade, gradiente e produtividade do conteúdo. Apresentam-se alguns resultados e analisam-se os preconceitos dos métodos existentes.

Palavras-chave: Línguas. Internet. Indicadores. Enviesamentos.

* Diretor de FUNREDES, Associação de Redes e Desenvolvimento para a promoção da Internet no Caribe. Observatorio de lenguas y culturas en la Internet (http://funredes.org/lc); Red Mundial para la Diversidad Lingüística (http://maaya.org). E-mail: [1]pimienta@funredes.org

Introducción

Durante el período 1998-2007, el Observatorio de la Diversidad Lingüística y Cultural en la Internet[1] ha sido un proyecto de la Fundación Redes y Desarrollo (FUNREDES[2]) y ha colaborado con la Unión Latina[3] para el diseño de métodos de medición de lenguas en la Internet que puedan proporcionar indicadores reproducibles y fiables; al mismo tiempo otras iniciativas[4] existía con los mismos objetivos (PIMIENTA, 2009). A partir de 2007, los cambios en el tamaño de la Web y el comportamiento de los motores de búsqueda han dejado obsoletos los métodos y ha creado un vacío en la producción de indicadores de lenguas en la Internet. Un nuevo método artesanal, basado en la observación del comportamiento del lenguaje en una amplia variedad de espacios y aplicaciones de la Internet, fue propuesto en 2012 y abrió nuevos estudios del Observatorio, con el sombrero institucional de la Red Mundial para la Diversidad Lingüística[5] y con el apoyo de la OIF[6]. Dos estudios preliminares brindan resultados en términos de clasificación del francés en la Internet. El segundo, realizado en 2013, alimentó el capítulo sobre la Internet del informe de 2014 "Le français dans le monde" (OIF, 2014) y fue seguido por un estudio similar del español en la Internet (PIMIENTA D., Prado D., 2016). El último estudio, financiado por la OIF, más ambicioso, que inspira este artículo, logró, mediante la aplicación de un enfoque estadístico, autorizado por el mayor número de fuentes, lograr resultados en términos de indicadores lingüísticos en la Internet para una amplia gama de lenguas.

El método se basa en recopilar información cuantitativa sobre el uso de la lengua en el mayor número posible de aplicaciones y espacios de la Internet. El proceso estadístico de fuentes permite medir la presencia de lenguas en la Internet y poner en perspectiva los resultados construyendo una serie de indicadores de la participación de lenguas en la Internet. Se extrae una síntesis en forma de una serie de macroindicadores que combinan todos los indicadores. El marco metodológico es utilizar fuentes bien directamente cuando se dispone de cifras relativas a lenguas, lo que lamentablemente es poco frecuente, o bien indirectamente, utilizando cifras por país y transformándolas en cifras por lengua. Esta transformación hace de

[1] http://funredes.org/lc
[2] http://funredes.org
[3] http://unilat.org
[4] En particular, el ambicioso Proyecto LOP (Mikami et al. 2006)
[5] http://maaya.org
[6] http://francofonía.org

este método un enfoque sin precedentes con la capacidad de producir datos lingüísticos sobre la Internet, en un contexto donde los indicadores linguísticos son pocos confiables y muy escasos.

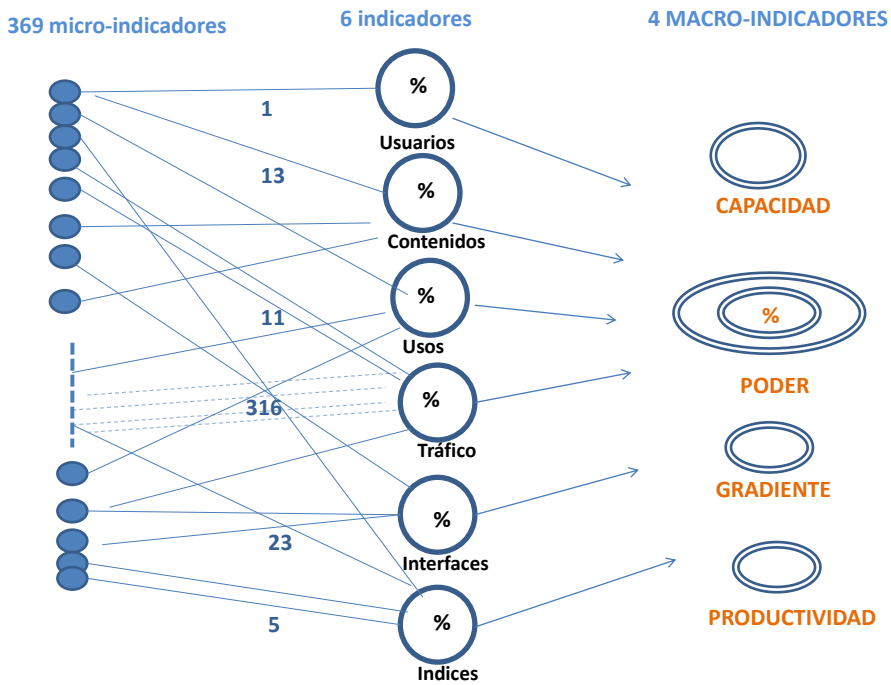
Este enfoque está respaldado por supuestos implícitos que deben hacerse explícitos y evaluarse para garantizar la coherencia, la confiabilidad y exponer los sesgos correspondientes. Los resultados se comparan con las 2 fuentes existentes: (W3TECHS, 2019) e (InternetWorldStats, 2019) y se analizan las diferencias notables bajo el foco de los respectivos sesgos.

Los detalles de la metodología, una recopilación de los resultados y el análisis completo de sesgos de los 3 métodos existentes se pueden consultar en (PIMIENTA, 2017) y en http://funredes.org/lc2017. En particular, una versión extendida de este documento está disponible en http://funredes.org/lc2017/Alternativa%20Lengua%20Internet.docx.

Indicadores

El siguiente diagrama muestra todos los indicadores que se procesan para cada lengua y la cantidad de fuentes correspondientes.

Figura 1: Diagrama de indicadores



La siguiente tabla muestra para cada indicador sus fuentes y cómo se calcula. Todos los indicadores se expresan en términos de cuota mundial, sobre la base de la población total de hablantes de L1+L2.

Tabla 1: Descripción de los indicadores

INDICADOR	DEFINICIÓN	PROCESO	FIABILIDAD
A: INTERNAUTS	Indicador único de la UIT: % mundial de personas conectadas por país.	Ponderación país -> lengua	Muy fuerte Solo sesgo marginal
B: USOS	Incluye 11 micro indicadores: Líneas telefónicas; mercado de comercio electrónico; Descarga de Open Office; Usuarios de redes sociales + Proyección 2021; Varias redes sociales suscriptores y proyecciones.	Ponderación C-> L Extrapolación por proporción Media truncada al 20%	Fuerte fiabilidad. Bajo sesgo. Pero sería necesario ampliar el número de microindicadores para dar más sentido a la media.
C: TRÁFICO	Alexa.com midió el tráfico de una selección de 316 sitios web.	Ponderación C-> L Proporción de extrapolación. Media truncada al 20%	relativamente bueno Pero enorme sesgo occidental de Alexa
D: SOCIEDAD DE LA INFORMACIÓN	Incluye 5 índices de WebIndex para los siguientes criterios: E.gov, Universal Access, E.participation...	Ponderación C -> L, media Transformar a % mundial ponderando con datos de la UIT	Bueno (datos subjetivos por organismo competente). Debería extenderse.
E: CONTENIDO (Wikipedia y libros)	Incluye 13 microindicadores Número de libros en Amazon; W3Techs; 11 indicadores de lengua de Wikimedia	Uso directo de cifras por lengua. Media truncada al 20%	Muy fuerte. Pero fuerte sesgo negativo para las lenguas asiáticas. Necesita ser extendido
F: INTERFAZ (e lenguas de traducción)	23 micro indicadores binarios 12 interfaces, 1 lengua de contenido, 10 aplicaciones de traducción	% de presencia transformado en % de palabra por ponderación con cifras de la UIT.	Perfecto.

La siguiente tabla muestra los macro-indicadores definidos.

Tabla 2: Descripción de macro-indicadores

POTENCIA	Mide la participación global de la lengua en la Internet	Media de los 6 indicadores (mundo L1+L2 %).
CAPACIDAD	Mide la fuerza de la lengua en la Internet independientemente de su número de hablantes.	Ratio de potencia vs. % mundial de hablantes. Sin dimensión, normalizado a 1.
GRADIENTE	Mide la potencia de los altavoces conectados independientemente de su número.	Ratio de potencia frente al % mundial de altavoces conectados. Sin dimensión, normalizado a 1.
PRODUCTIVIDAD	Mide la propensión de los hablantes conectados a producir contenido en su lengua.	Ratio de % de contenidos vs % de altavoces conectados. Sin dimensión, normalizado a 1.

Cálculos

El modelo se basa en 3 categorías de datos: 1) la gran lista de fuentes relacionadas con la Internet por lengua o país 2) datos demo-lingüísticos 3) datos de la UIT sobre el porcentaje de personas conectadas a la Internet por país (UIT, 2009).

1) Todos los microindicadores se expresan en porcentaje mundial. La transformación de país a lengua se realiza ponderando con el número de hablantes de cada lengua en cada país. Las fuentes rara vez cubren todos los países del mundo entonces se utilizan algunas técnicas de extrapolación, ya sea en proporción al porcentaje de personas conectadas por país o utilizando el método de cuartiles. Cuando la extrapolación carece de sentido se rechaza el microindicador.
2) Hoy en día, existen dos fuentes que proporcionan la matriz de cantidad de hablantes de L1 de cada lengua para cada país: el proyecto Joshua (sin cargo) y Ethnologue (de pago). La primera edición del modelo utilizó Joshua. En cuanto a los hablantes de L2, se usan los datos de Ethnologue y las mediciones futuras intentarán usar Ethnologue también para L1.

3) datos de la UIT, considerados como confiables y esenciales para el método, se actualiza de forma gratuita cada año.

La matriz LOC1 cumple con la siguiente definición, para todos los lenguas seleccionados y todos los países seleccionados: LOC1 (i,j) = Número de hablantes de L1 para la lengua i en el país j.

La fuente proporciona cifras para 7500 lenguas, pero solo se procesará un subconjunto. El número estimado de lenguas presentes en la Internet es de alrededor de 500. Una posibilidad es apuntar a ellos. Otra posibilidad es seleccionar las lenguas para los que Wikipedia ofrece estadísticas (cerca de 300). Después de varias pruebas, la elección finalmente se asentó en la lista de las 140 lenguas con más de 5 millones de hablantes. La decisión se tomó con el fin de reducir los sesgos resultantes de los supuestos implícitos.

En cuanto a L2, la primera prioridad es la de tener en cuenta coherentemente el multilingüismo. Las personas computadas en L2 obviamente también tienen una primera lengua y por lo tanto el conjunto de hablantes de L1+L2 incluye a las mismas personas más de una vez. La evidencia dice que las cifras necesariamente deben basarse en el total de hablantes de lenguas en el mundo y no en relación con la población mundial. Lamentablemente, esta evidencia es ignorada por muchas fuentes y provoca errores. En el escenario que se adoptó, la participación mundial se calculará sobre la base del 125% de la población mundial (cifra calculada a partir de las entradas demo-lingüísticas). Esta noción es igualmente aplicable a todos los conceptos: usuarios, tráfico, uso, contenido, interfaces e índices (por ejemplo, los sitios web se pueden hacer en varias lenguas, lo mismo para el flujo de correos electrónicos). El método ideal para tratar el caso de L2 sería obviamente producir, como para L1, una matriz LOC2 (i, j) = número de hablantes L1+L2 de la lengua i en el país j. Desafortunadamente, estos datos no están disponibles. Luego se propone otro enfoque cuyo principio simple consiste, para cada lengua, en obtener un número que represente el aumento a aplicar a las cantidades L1 para obtener el valor L1+L2 y utilizar un enfoque lineal para los resultados. La tasa de incremento global (1,25) es el resultado de la siguiente operación de ponderación: para obtener un número que represente el aumento que se aplicará a las cantidades L1 para obtener el valor L1+L2 y utilizar un enfoque lineal para los resultados. La tasa de incremento global (1,25) es el resultado de la siguiente operación de ponderación: para obtener un número que represente el aumento que se aplicará a las cantidades L1 para obtener el valor L1+L2 y utilizar un enfoque lineal para los resultados.

$$Rg = \sum_{j=1}^{j=L} L1(j) \times R12(j)$$

donde L es el número total de lenguas, L1(j) el número de hablantes de L1 para la lengua j y R12(j) la tasa de aumento de L1 a L1+L2 para el lengua j. El valor de los micro indicadores para L1+ L2 se calcula de esta forma a partir del valor para L1:

$$ML1+L2(i) = Rg \times ML1(i) \ / \ \sum_{j=1}^{j=L} ML1(j) \times R12(j)$$

El método L1+L2 se aplica a todos los indicadores excepto *Índice, Contenido e Interfaz*, que por naturaleza están destinados a aplicarse directamente a L1+L2. Este método es menos preciso que una solución que podría funcionar a nivel de país y genera entonces algunos sesgos.

En cuanto al cómputo de los indicadores, sólo requieren cómputo los expresados por países. El principio para convertir las cifras expresadas en porcentajes por país en porcentajes por lengua es el producto matricial entre la matriz LOC y el vector MCn que contiene las cifras fuentes por país para el microindicador n. El microindicador expresado en porcentaje por lengua (MLn) es entonces:

$$MLn(i) = \sum_{j=1}^{j=p} LOC(i, j) \times MCn(j)$$

donde P es el número total de países, LOC (i, j) es el número de hablantes de la lengua i en el país j y MCn(j) es el valor medido para el microindicador n en el país j.

El producto de matriz ML = LOC +. x MC en APL[7]o ML=SumProduct (LOC; MC) en notación Excel, es una operación de ponderación de los valores del microindicador en cada país con la presencia de cada lengua en cada país. Los totales del MLn son los mismos que los del MPn pero esta vez la distribución se hace por lengua en lugar de por país. Como la mayoría de los cálculos se basan en ponderaciones, es útil identificar los diferentes tipos utilizados en el proceso y hacer explícitos los supuestos simplificadores que subyacen a la validez de los resultados obtenidos por estas ponderaciones, supuestos que guiarán la comprensión de los sesgos.

[7]APL, "Un lenguaje de programación", un formalismo matemático y un lenguaje de programación.

Tabla 3: Diferentes ponderaciones aplicadas

	Demo-lingüística	**L2**	**Usuarios**
TIPO	P → L	L1 → L1+L2	Criterio % → mundo %
APLICACION	Datos por País	Resultados L1	% por criterio
RESULTADO	Datos por Lengua	Resultados L1+L2	% Mundial
PONDERACIÓN	matriz LOC	L1+L2/L1 por L	datos UIT
ALCANCE	Todas las fuentes por país	Usuarios, tráfico y uso	Índice e interfaces
SUPOSICIÓN IMPLÍCITA	Tasa de conexión idéntica para todos los L1 en el mismo País	Tasa de conexión idéntica para L2 que para L1	Modulación según tasa de conexión a la Internet

Resultados

La siguiente tabla muestra los resultados corregidos de sesgo para los primeros 10 lenguas de contenido y permite la comparación con las otras dos fuentes existentes, mostrando fuertes discrepancias que se entienden bien cuando se analizan cuidadosamente los sesgos (PIMIENTA, 2017).

Tabla 4: Comparaciones para las diez principales lenguas

	CONTENIDO	**W3TECH**	**INTERNAUTAS**	**IWS**
inglés	**32,0%**	51,9%	20,4%	26,3 %
chino	**18,0%**	2,0%	20,0%	20,8 %
español	**8,0%**	5,1%	9,1%	7,7%
francés	**6,5%**	4,1%	4,9%	2,8%
alemán	**3,8%**	5,5%	2,7%	2,3%
portugués	**3,5%**	2,6%	4,1%	4,3%
japonés	**3,5%**	5,6%	4,5%	3,2%
ruso	**3,5%**	6,5%	4,9%	2,9%

hindi	**3,0%**	< 0,1%	4,6%	n / A
árabe	**3,0%**	0,7%	3,0%	4,7%
Restante	***40,2%***	*15,9%*	*46,6%*	*25,0 %*
TOTAL	125,0%	100,0 %	125,0%	100,0 %

La siguiente tabla muestra el ranking por *capacidad* y no es de extrañar ver arriba las lenguas de países con fuertes políticas para la Sociedad de la Información.

Tabla 5: Diez lenguas principales para *capacidad*

	Capacidad	**Clasificación potencia**	**% conectados**
hebreo	5.40	35	76.05
finlandés	5.40	38	92.30
holandés	4.81	19	92.27
sueco	4.46	28	90.54
inglés	3.72	1	78.05
alemán	3.40	6	86.43
danés	3.30	49	95.67
italiano	3.16	12	64.20
checo	3.13	27	81.17
francés	2.96	4	81.09

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

Tabla 6: Nueve lenguas principales para *gradiente*

	Gradiente	**Clasificación potencia**
hebreo	2.62	35
finlandés	2.16	38
holandés	1.93	19
sueco	1.81	28
inglés	1.76	1
checo	1.42	27
inglés	1.76	1
italiano	1.73	12
serbocroata	1.54	22

Agradecimientos

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

El estudio fue financiado por la OIF (http://francophonie.org).

REFERENCIAS

Ethnologue. **Languages of the World**. Disponible en https://www.ethnologue.com, 2019.

Internet World Stats**Internet world users per language, top 10 languages**. Disponible en https://www.internetworldstats.com/stats7.htm, 2019.

ITU. **Percentage of individuals using the Internet per country**. Disponible en https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2019/Individuals_Internet_2000-2018_Jun2019.xls, 2019

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

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

Pimienta D., An alternative approach to produce indicators of languages in the Internet in **Proc. of Global Expert Meeting Multilingualism in Cyberspace for Inclusive Sustainable Development**, Khanty-Mansiysk, Russian Federation. Disponible en http://funredes.org/lc2017/Alernative%20Languages%20Internet.docx http://funredes.org/lc2017/Alternativa%20Lengua%20Internet.docx (en español), June, 2017.

Pimienta D., Prado D. 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. Disponible en http://dx.doi.org/10.3989/redc.2016.3.1328, 2016.

Pimienta, D., Prado D. et al, (2009), Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, in **UNESCO Publications for the World Summit on the Information Society**, CI.2009/WS/1. Disponible en http://unesdoc.unesco.org/images/0018/001870/187016e.pdf, 2012.

W3Techs. **Usage of content languages for websites**. Disponible en https://w3techs.com/technologies/overview/content_language/all, 2019.