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

ESTUDIOS / *RESEARCH STUDIES*

Medición de la presencia de la lengua española en la Internet: métodos y resultados

Daniel Pimienta*, Daniel Prado**

*Presidente de la Fundación Redes y Desarrollo (Funredes) y Secretario Ejecutivo de la Red mundial por la diversidad lingüística (Maaya)
**Consultor y miembro del Comité Ejecutivo de la Red mundial por la diversidad lingüística (Maaya)
Correos-e: pimienta@funredes.org, dhprado@gmail.com

Recibido: 13-08-2015; 2ª versión: 21-12-2015; Aceptado: 22-01-2016.

Cómo citar este artículo/Citation: Pimienta, D.; Prado, D. (2016). Medición de la presencia de la lengua española en la Internet: métodos y resultados. *Revista Española de Documentación Científica*, 39(3): e141. doi: http://dx.doi.org/10.3989/redc.2016.3.1328

Resumen: En esta comunicación se resumen estudios recientes que exploran y analizan la situación de la lengua española en la Internet e infiere lecciones metodológicas que podrían ser útiles para estudios similares para otros idiomas. El método ha sido utilizado por primera vez para obtener resultados acerca de la situación de la lengua francesa en la Internet, en un estudio financiado por la *Organisation internationale de la Francophonie*. Al tomarse en cuenta una selección de espacios y aplicaciones de mucho uso, la lengua española podría ser considerada entre la tercera y la cuarta lengua dentro de los usos de la Internet (cuando es la tercera en número de internautas) ocupando los primeros lugares en materia de P2P, videos y blogs, apreciándose un rol menor en materia de libros, páginas web y aplicaciones. Es así como el español en la Internet parece actuar más dinámicamente como vector de comunicación que como vector de información.

Palabras clave: Cibermetría de lenguas; diversidad lingüística; medición de lenguas; multilingüismo en el ciberespacio; español en la Internet; presencia de lenguas en la Internet.

Measuring the presence of the Spanish language on the Internet: methods and results

Abstract: This paper summarizes recent studies that explore and analyze the situation of the Spanish language on the Internet and infers methodological lessons that could be useful for similar studies for other languages. The method was first used to obtain results for the state of French on the Internet in a study funded by the *Organisation internationale de la Francophonie*. By taking into account a selection of web spaces and commonly used applications, Spanish could be ranked between third and fourth place in terms of Internet usage (whereas it is third in terms of Internet users), occupying the top positions for P2P, videos and blogs, but having a minor role in terms of books, webpages and applications. Thus Spanish on the Internet appears to act more dynamically as a vector for communication than for information.

Keywords: Cybermetric language; Linguistic diversity; measuring languages; multilingualism in cyberspace; Spanish on the Internet; presence of languages on the Internet.

1. INTRODUCCIÓN

La Red Mundial por la diversidad lingüística, Maaya (Maaya, 2015), ya sea directamente o a través de algunos de sus miembros, como Funredes (Funredes, 2014), el *Language Observatory Project* (LOP, 2012) o Unión Latina (Unión Latina, 2015), ha llevado a cabo una serie de estudios para analizar la situación de las lenguas en la Internet desde 1988 hasta el presente.

Por un lado, Funredes y la Unión Latina desarrollaron inicialmente una metodología de medición específica para un grupo de lenguas[1], permitiendo realizar una serie de campañas de medición en diferentes espacios de la Internet, entre 1988 y 2008 (Estudio lenguas y ciberespacio, 2012 y Observatorio de lenguas y culturas en la Internet, 2012).

Por otro lado, el *Language Observatory Project* (LOP, 2012) estudió, en la segunda parte del mismo periodo, el comportamiento de las lenguas en la mayoría de los dominios de Internet de nivel superior (conocidos como ccTLD) de países de Asia y África, con la intención de medir la situación particular de las lenguas minoritarias.

Mientras el *Language Observatory Project* basaba sus estudios en el rastreo sistemático de las páginas Web del dominio elegido mediante un algoritmo de reconocimiento de idiomas, Funredes y la Unión Latina utilizaban los dispositivos de conteo de número de ocurrencias de los motores de búsqueda, con un muestreo de palabras suficientemente específicas que pudiesen permitir comparación entre ciertos idiomas.

Pero en 2008 Funredes / Unión Latina dejaron de utilizar dicha metodología como consecuencia del cambio de comportamiento de los motores de búsqueda (porcentajes bajos de indexación, cifras diferentes en función del usuario, preferencia por los sitios que más beneficios económicos aportan al motor, ninguna confiabilidad en los conteos propuestos, etc.) y el *Language Observatory Project* dejó de realizar su medición sistemática, impidiendo así contar con medios serios de supervisión de la evolución de las lenguas en la Internet, dejando lugar a la circulación de cifras provenientes de fuentes poco fiables, de carácter promocional o comercial (Pimienta y otros, 2009).

Para superar esta situación, Maaya diseñó un proyecto de investigación muy ambicioso, Dilinet (Dilinet, 2014), aun en espera de financiación, sin la cual no se puede pretender una investigación profunda del cada vez más *insondable* ciberespacio. En paralelo, editó un compendio de artículos con grandes nombres mundiales especializados en la diversidad lingüística en Internet, algunos de cuyos artículos han facilitado y orientado la investigación aquí presentada (Net.Lang, 2012).

En espera del éxito del proyecto Dilinet, se exponen en este trabajo estudios cuyos métodos representan una alternativa, quizás menos ambiciosa, pero válida para alcanzar un nivel aceptable de estimación sobre la situación de las lenguas estudiadas en las zonas más visibles de la Internet, y emprender así una sistematización en la observación de la situación de las lenguas.

El método descrito aquí se ha utilizado por primera vez para la lengua francesa (OIF, 2014), solicitado por la Organización Internacional de la Francofonía (*Organisation internationale de la Francophonie* – www.francophonie.org) y podría servir de inspiración para estudios sobre otros idiomas con un gran número de hablantes, como el portugués, el ruso, el árabe o el alemán. Los resultados mostrados aquí han sido los obtenidos para la lengua española.

2. ANTECEDENTES Y ENFOQUE

Aunque diversos indicadores podrían evaluar de forma precisa la presencia de la lengua española (en educación, en ciencia, en organismos internacionales, en materia de traducción de idiomas, en materia de edición, etc. [Unión Latina, 2010]), quedan dudas al hablar de la presencia española en el ciberespacio, a pesar de que los medios de comunicación le atribuyen, sin mayor reflexión, un hipotético tercer lugar que ocuparía la lengua española en la Internet.

La voluntad, aparentemente simple, de medición de la "presencia" de una lengua en la Internet choca con un malentendido permanente, que es la consecuencia de la escasez de datos sobre el tema, causa de discrepancias en las cifras facilitadas por diversas fuentes.

Dos indicadores diferentes suelen confundirse en los resultados publicados:

- El porcentaje estimado *de usuarios de* la *Internet* en una lengua determinada;
- El porcentaje estimado *de contenidos en la Internet* para un determinado idioma.

La medición del número de usuarios de la Internet hispanohablantes o la del número de páginas web en español son fundamentalmente diferentes, y reflejan diferentes realidades que merecen una mirada también diferente: la primera medida está relacionada con lo que llamamos *brecha digital* (es decir, el acceso físico a la Internet) y el segundo, con *la brecha de contenidos* (es decir la presencia de información en el idioma estudiado), una brecha mucho menos conocida pero más decisiva.

Es así, cuando se encuentran en los periódicos o en algunos informes las cifras de la "presencia del español", se debe entender y diferenciar si se refiere a la lengua del internauta o a la de los contenidos. Por lo tanto, la afirmación de que el español es la tercera lengua de la Internet, información ampliamente difundida por los medios de comunicación, sólo tiene sentido si se especifica que la tercera población de usuarios de Internet es la hispanohablante. De ninguna manera eso implica que el español es el tercero en términos de contenidos como veremos adelante: nuestro estudio establece que, en términos de páginas web, el español sería la quinta lengua y si se ponderan debidamente todos los factores de presencia relacionados con la Internet (el acceso, los contenidos y los múltiples usos posibles) llegaríamos a una posición mediana entre 3 y 4.

Desde luego, este nuevo estudio introduce una nueva variable que no había sido estudiada en los trabajos anteriores (más centrados en contar páginas web), la de medir los usos en función de aplicaciones.

Medición de usuarios de Internet por idioma

Los datos sobre el número de internautas por lengua provienen de la fuente más consultada en este ámbito: InternetWorldStats (InternetWorldStats, 2013), fuente que, si bien está lejos de satisfacer las expectativas de estadísticas rigurosas, al menos tiene el mérito de ser la única que existe (para medir los usuarios por lenguas) y de actualizarse periódicamente. Su metodología consiste en determinar los principales idiomas que se utilizan en cada país y cruzar esta información con los datos de la Unión Internacional de Telecomunicaciones sobre el número total de usuarios de Internet en cada país. Es de notar que las cifras dejaron de ser actualizadas entre mayo del 2011 y finales del año 2013, tomadas como referencia para este estudio.

Sin embargo, los datos de la Unión Internacional de Telecomunicaciones son producidos por los gobiernos, lo que no es necesariamente un criterio de fiabilidad y de homogeneidad, puesto que algunos países tienden a inflar las cifras dadas a la Unión Internacional de Telecomunicaciones para demostrar el éxito de sus esfuerzos para luchar contra la brecha digital (UIT, 2010). Por otra parte, no hay ninguna mención de la metodología utilizada por InternetWorldStats para contar el número de usuarios de Internet para las lenguas estudiadas. Al parecer, el único criterio es el del (o los) idioma(s) oficial(es) de cada país. InternetWorldStats también utiliza, para cada país, diversas fuentes de mercadeo, sin metodología común entre los países. Por último, este estudio se limita a los diez principales idiomas de los usuarios de Internet, en contraste con la compañía que proporcionaba cifras comparables antes de 2007 (GlobalStats, desaparecida hoy día), consultable solo mediante el *Wayback Machine* del sitio https://archive.org.

Medición de contenidos por idioma

En materia de medición de la lengua de los contenidos a nivel mundial, hay, y ha habido, profusión de publicaciones (por lo general, originadas por empresas de mercadotecnia) con cifras diversas cuyo método ha sido raramente revelado, lo que imposibilita validar los resultados. Sólo un estudio puntual realizado en los albores del siglo mantenía un rigor cierto y debidamente documentado, de la sociedad Xerox (Grefenstette y Nioche, 2001).

Más recientemente, las "Encuestas de Tecnología W3Techs" (W3Techs, 2014), probablemente la fuente de información más fiable y completa sobre los usos de Internet, tiene la ventaja de ofrecer datos sobre la lengua de los contenidos[2]. W3Techs calcula sus estadísticas a partir de los elementos suministrados por Alexa (Alexa, 2014), una empresa capaz de proporcionar estadísticas sobre la frecuentación de sitios o páginas web a través de una barra de herramientas que los internautas aceptan instalar en su navegador y que reporta a Alexa las etapas de su navegación. Se trata de un muestreo compuesto por un número de usuarios selectos. Alexa presenta así los 25 millones de sitios más visitados sorteados por el número de visitas. A sabiendas de que la web cuenta (en marzo 2015) con casi 877 millones de sitios Internet (Netcraft, 2015), de los cuales se considera que casi 177 millones son sitios realmente en funcionamiento, esa muestra solo representa la punta de una pirámide. De esos 25 millones clasificados por Alexa, W3Techs toma los primeros 10 millones y determina los idiomas identificados en los metadatos o por algoritmo de reconocimiento de la lengua. Lo interesante es notar que W3Techs actualiza diariamente sus datos, lo que permite apreciar la evolución desde la fecha de inicio del servicio, junio de 2013, hasta hoy día[3].

Otro elemento de interés en el trabajo de W3Techs es su capacidad de cruzar diferentes datos:

- http://w3techs.com/technologies/cross/top_level_domain/content_language permite cruzar los nombres de dominio y las lenguas de la página (por ejemplo, se sabe que el 30% de los sitios en francés pertenecerían al dominio de nivel superior .fr[4], datos que no existen desafortunadamente para ningún país hispanohablante);

- http://w3techs.com/technologies/cross/content_language/top_level_domain realiza el cruce recíproco (así el 93% de los sitios de dominio .fr estarían en francés);
- http://w3techs.com/technologies/cross/content_language/ranking cruza la clasificación jerárquica con el parámetro de idioma (así el 60% de los sitios clasificados en los 1.000 primeros rangos estarían redactados en inglés).

Hoy día, la estimación de W3Techs sobre el lugar del inglés (55,5% de todas las páginas Web) es mucho más alta que la indicada por el estudio de Funredes / Unión Latina en 2007 (44%) y mucho más alta que lo que sería nuestra proyección actual (en torno al 34%). Y es que hay varias razones para obtener, con ese método, cifras de la lengua inglesa muy por encima de la realidad:

1. Funredes / Unión Latina (Paolillo, 2015) o el *Language Observatory Project* (Suzuki y otros, 2002) centraban su investigación en el contenido de *las páginas web,* lo que permitía medir, dentro de cada sitio web, el idioma de cada página y no solo de la principal (*home page*); en cambio W3Techs centra su investigación *en los sitios web*, lo que provoca la toma en cuenta para el sitio entero de la página principal, muchas veces conteniendo inglés (lengua de entrada principal de muchos sitios multilingües), incluso si el resto está escrito en otros idiomas.
2. Dado que Alexa se instala voluntariamente por el usuario, es un buen instrumento para medir los sitios que visitan los usuarios que lo han instalado, pero es un instrumento muy utilizado en ciertas regiones del planeta, y mucho menos en otras, lo que introduce muchos sesgos en materia de poblaciones estudiadas.
3. Alexa, por su naturaleza, mide el uso y no la existencia; de modo que las páginas no visitadas o poco visitadas por sus usuarios no llegan a ser identificadas y, por lo tanto, contabilizadas por W3Techs.
4. Pero la principal razón es que W3Techs sólo considera los 10 millones de sitios web más visitados según Alexa, es decir, un poco más del 1 % de los sitios existentes o, para ser más justos, un poco menos del 6% de los sitios considerados válidos. Por lo tanto, los sitios visitados incluirán necesariamente los medios de comunicación y los sitios comerciales más reputados de los diferentes países, mayoritariamente los países occidentales, con los Estados Unidos en primer lugar. Es así como aquellos sitios que presentan un interés mundial menor (sitios científicos, culturales, administrativos nacionales, sitios localizados en lenguas menos utilizadas, etc.), no figurarán en dicha clasificación. Entre los posibles sesgos lingüísticos de este estudio, cabe señalar que la República Checa tendría más páginas que Corea, o bien que habría menos páginas en chino, en alemán o en ruso, con una población en línea casi 10 veces más grande.

A pesar de sus limitaciones, W3Techs representa la fuente más atractiva entre los indicadores disponibles en la actualidad y se deben aceptar con satisfacción los progresos que representa para el campo de estudio.

Evolución de los motores de búsqueda desde 2008

A partir de 2008 se ha reducido el número de motores de búsqueda y aquellos que han quedado (Google, Yahoo, Bing / Live Search, Ask, AOL, Lycos, Excite, Exalead, Teoma[5]) han evolucionado de la misma manera:

- reducción significativa del porcentaje de la parte indexada de la Web (más del 80% a menos del 5% del espacio total);
- pérdida total de credibilidad de las cifras publicadas por el número de apariciones de una palabra clave determinada;
- implementación de la *"búsqueda inteligente"* por palabra-clave dando lugar a la pérdida de asociación entre palabras-clave y resultados (ya sea causada por la traducción automática, o por algoritmos de sinonimia o de corrección ortográfica).

A partir de 2008, y de manera gradualmente amplificada, el tamaño de la Web se ha convertido en incontrolable y puede considerarse en términos prácticos, como tendente hacia el infinito. Es decir que los motores de búsqueda no pueden, por razones de costo, llevar a cabo un rastreo sistemático integral de toda la web, contentándose con un aproximado 5% de páginas indexadas en total[6]. Si hay un área en donde la falta de transparencia es la regla, es en el tamaño de los índices. En función del motor de búsqueda, se utilizan varios trucos para ocultar esta limitación de exploración de todas las páginas de un sitio web, la cual no se aplica de la misma manera a todos los idiomas en clara desventaja de las lenguas minoritarias, menos susceptibles de atraer publicidad o audiencia.

Junto con el auge de la Web 2.0 la naturaleza de la Internet ha cambiado y las páginas estáticas (HTML de base) han dejado un mayor espacio para las páginas dinámicas. En el mismo período, la topología de la Internet en materia de lenguas ha cambiado radicalmente (Prado y otros, 2010), con una reducción en términos de crecimiento de las lenguas inicialmente bien representadas (lenguas occidentales y japonés, en particular) y crecimientos exponenciales para los idiomas de Asia y, más recientemente, de la lengua árabe. Al mismo tiempo, la naturaleza del contenido ha evolucionado mediante la reducción de la proporción de información textual y un aumento masivo de documentos multimedia, significando, para el video en particular, un 79% por ciento de todo el tráfico estimado en 2018, cuando representaba un 66% en 2013. Ver *Cisco Visual Networking Index* para las predicciones 2014-2019 (CISCO, 2014).

En tales condiciones, el porcentaje de páginas textuales sigue siendo un indicador de cierta importancia pero, frente a una realidad más compleja, se deben crear otros indicadores que reflejen mejor dicha complejidad y aceptar que se deben tratar separadamente elementos parciales de un mosaico, en lugar de tratar con un número limitado de indicadores integrales.

En 2010, la Unión Latina, en colaboración con Funredes, llevó a cabo un primer intento de captar la percepción de la realidad compleja de las lenguas en diferentes espacios de la Web, y aunque dicha experiencia no dio pie a una publicación específica, permitió contribuir indirectamente a algunas publicaciones (Pimienta, 2011, UIT, 2010).

El trabajo del año 2010 exploró la presencia de un gran número de lenguas en las aplicaciones más conocidas (redes sociales, blogosfera, peer-to-peer, motores de búsqueda, VOIP[7], Wikipedia, Youtube, etc.). A menudo, en ausencia de otras alternativas, los datos se construyeron a partir del origen geográfico del tráfico de dichas aplicaciones, cruzando los datos por país con fuentes demolingüísticas. Estos principios que han contribuido a perfilar un nuevo enfoque metodológico se ampliaron y extendieron en este nuevo estudio.

3. METODOLOGÍA

La metodología propuesta se basa en dicho enfoque del 2010 y se aplica a un determinado idioma, el español en este caso, comparándolo con los principales idiomas utilizados en la Internet, extendiéndola al mayor número posible de espacios y de aplicaciones. El enfoque inicial que estaba basado a menudo en el tráfico por país se complementa con la investigación intensa y sistemática sobre el contenido lingüístico de esas aplicaciones.

La metodología se basa entonces en un trabajo amplio y abierto de colecta de datos sobre el uso de las lenguas en diferentes aplicaciones y espacios de Internet, a partir de fuentes diversas y múltiples, seguido por una recopilación y organización de esos datos, luego, su evaluación y validación, finalmente el cruce de los datos organizados con diferentes estudios de referencia considerados como fiables, y en última instancia, la puesta en perspectiva de los resultados obtenidos con el fin de comprobar las tendencias y los indicadores compuestos que dan cuenta sobre los nuevos acontecimientos.

Los elementos metodológicos clave de este enfoque son:

- espacios y aplicaciones seleccionados;
- búsqueda, selección y análisis de fuentes de información ofreciendo datos sobre el lugar del español en relación con esos espacios y aplicaciones;
- cruce con datos demo-lingüísticos que permiten poner en perspectiva los datos recogidos;
- síntesis a partir del conjunto de datos dispersos obtenidos permitiendo informar, de manera significativa, sobre el lugar del español en la Internet en comparación con otras lenguas.

Se debería advertir que:

1. Los datos expuestos son extraídos de fuentes analizadas entre finales del 2013 e inicios del 2014 y no reflejan la situación de hoy ni el espacio abrumador que ha tomado la Internet móvil y sus aplicaciones propias en el 2015.
2. Como sea, la confianza que se le puede atribuir a las numerosas fuentes procesadas no es homogénea según las aplicaciones o espacios, y en muchos casos no es muy alta. Se deben entonces tomar los resultados específicos con cierta precaución, aún más cuando una de las categorías comporta muy pocas fuentes.

Lo que sí consideramos es que el conjunto de todas esas fuentes ofrece globalmente una estimación relativamente confiable sobre lo que es la situación del español en la Internet.

Se han identificado un centenar de aplicaciones y espacios de Internet como los más apropiados para informar sobre el espacio de las lenguas en la Internet. Treinta de ellos no ofrecían datos fiables o utilizables de manera uniforme, por lo que han sido excluidos provisoriamente y se integrarán tan pronto como lo permitan. Igualmente, ciertos espacios utilizados estudiados no presentan datos utilizables para el objetivo de este artículo por lo que se han descartado, como ha sucedido con los navegadores y los motores de búsqueda.

La tabla I muestra la lista inicial de los espacios y aplicaciones elegidos (incluyendo los que se desecharon), organizada según el tipo de aplicación o el espacio estudiados.

4. LAS FUENTES ANALIZADAS

Se necesitó un importante esfuerzo para recopilar fuentes sobre la presencia de las lenguas en la Internet en general, y en español, en particular. Este capítulo está dedicado a las fuentes que alimentaron los resultados presentados.

La falta de producción de datos en materia de cibermetría de lenguas ha creado un vacío en donde la gran mayoría de las fuentes provienen de empresas de negocios o de mercadeo en línea, que filtran información parcial de forma gratuita (y por lo general sin revelar la metodología) como medio para promover sus servicios de pago. Fuera de las estadísticas tradicionales y confiables que proporcionan cierta información útil sobre las lenguas (Organización de las Naciones Unidas, UNESCO, Unión Internacional de Telecomunicaciones, Organización para la Cooperación y Desarrollo Económicos, Unión Europea, etc.), no sin

Tabla I. Espacios y aplicaciones de Internet seleccionados para analizar la presencia de las lenguas en Internet
En cursiva, las fuentes que han sido descartadas en las ponderaciones finales para la lengua española o que no han sido utilizadas en el presente artículo

Infraestructura	**Libros y artículos en línea**	**Teléfonos y tabletas**	**Mensajería y telefonía IP**
Usuarios de Internet por lengua	*Biblioteca virtual*	Smartphones	Skype
Computadoras por país	*GoogleBooks*	*Tabletas G3*	*QQ*
Servidores por habitantes	Amazon	*Data Sims*	AIM
Sitios web por internautas	*GoogleScholar*	3G	ICQ
Penetración de Internet			Yahoo!
Penetración de banda ancha			
Trafico de traducción automática			
Herramientas lingüísticas			
Telefonía móvil por habitante			
Descarga de archivos y P2P	***Redes sociales y videos***	***Blogs***	***Conteo de páginas Web***
Megaupload	Wikipedia	*Technorati*	W3 Techs
Rapidshare	Facebook	Blogs	*Webboar*
Filefactory	Twitter	Wordpress	Internetarchive
Depositfiles	Linkedin	*Google Blogs*	*Google*
Hotfile	Viadeo	*Blogger*	*Baidu*
Uploading	*Xing*	*Blogspot*	*Wolfram Alpha*
Uploaded	*Yahoo*	*Sina Weibo*	MSN
Fileserve	Google+	*Technorati*	*Bing*
Mediafire	Windows Live Profile	Tumblr	Foofind
Gigasize	*Myspace*		*Rtbot*
Bitshare	Livejournal		Ccsearch
4shared	*Secondlife*		*Altavista*
	Ning		*Yandex*
	Tuenti		*Wikia Search*
	Hi5		Open Directory Project
	Orkut		
	Badoo		
	Instagram		
	Sonico		
	Qzone		
	Youtube		
	Googleplay		

Correo electrónico	***Motores de búsqueda***	***Navegadores***	***Sistema operativo y aplicación ofimática***
Gmail	*Google*	*Chrome*	*Windows*
Hotmail	*Bing*	*Firefox*	*Linux*
Yahoo	*Yahoo!*	*IE*	*Mac*
Yandex Mail	*Yandex*	*Safari*	*Ios*
Icloud	*Baidu*	*Opera*	*Android*
Outlook	*Otros*	*Otros*	OpenOffice
			Microsoft Office
			Otros

En cursiva, las fuentes que han sido descartadas en las ponderaciones finales para la lengua española o que no han sido utilizadas en el presente artículo

dificultades en el momento de discriminar el aspecto lingüístico en la Internet, algunos consultores o expertos difunden la recopilación de información sobre un espacio o aplicación determinados, con el fin de promover sus conocimientos. Es el análisis del conjunto de esos sitios profesionales, aunque a menudo imperfectos, lo que permite, no sin cierta dificultad, recopilar un número interesante de datos relativamente confiables segmentados por aplicación (Wikipedia, Twitter, Youtube, Facebook, etc.) o por espacio (motores de búsqueda, correo electrónico, VoIP, etc.).

Otra limitación a tener en cuenta en este tipo de análisis es que el grado de globalización de los espacios y aplicaciones es cada día más variable, puesto que hay países o regiones que adoptan aplicaciones locales específicas en detrimento de las principales aplicaciones con vocación mundial (Baidu en lugar de Google en China y Vkontakte y Odnoklassniki en lugar de Facebook en Rusia, por ejemplo). Es importante tener en cuenta este factor cuando se establecen los datos cuantitativos sobre el uso de las lenguas (o sobre el uso por países) para cada aplicación o espacio determinado.

Por ejemplo, concluir que en las redes sociales no profesionales, el francés sería la cuarta o la primera lengua a partir de sus resultados respectivos en Facebook o Viadeo, debe interpretarse de manera relativa puesto que ninguna de estas aplicaciones tiene una penetración geo-equilibrada o *linguo-equilibrada*. Para obtener un indicador creíble sobre el lugar que cada lengua ocupa en una jerarquía de uso, cada aplicación debe ser ponderada en función de su distribución en el mundo (peso total y presencia relativa de un país).

Una visión rápida permite apreciar que las aplicaciones más globalizadas son Wikipedia, Twitter y YouTube[8]; otras son relativamente globalizadas, como Facebook y Google (ésta, con competencias de motores locales en China, Rusia, Kazajstán, Corea, etc.), o que muestran diferentes hábitos (en Japón, Yahoo! es mucho más utilizado que Google, por ejemplo).

Para la mayoría de las aplicaciones y espacios estudiados se debe tener cuidado y matizar las conclusiones lingüísticas de los resultados en función de posibles sesgos en los usos.

Desde 2010, se han analizado y almacenado varios cientos de fuentes potenciales, y se ha mantenido una observación permanente mediante búsqueda de palabras clave o consultando los enlaces externos de las fuentes ya seleccionadas. Muchas fuentes han sido excluidas por alguna o algunas de las causas siguientes:

- el campo de estudio es demasiado pequeño o parcial
- los datos parecen ser demasiado parciales
- las estadísticas no han sido actualizadas, o bien no había diferencias significativas entre dos fechas de consulta
- la metodología utilizada no permite comparar los datos
- la metodología utilizada es poco pertinente, no adecuada o poco creíble
- quedan dudas sobre la fiabilidad de la fuente.

Finalmente se han seleccionado, clasificado, evaluado y valorado cerca de 200 fuentes diferentes (URL, artículos, libros u otros) que permitirían identificar distintos indicadores sobre la presencia de las lenguas en diferentes áreas.

Clasificación de las fuentes

Cada elemento de este muestreo de fuentes fue notado y clasificado de 5 (promedio) a 10 (excelente) y aquéllos que obtenían menos puntaje fueron rechazados automáticamente. Las notas han sido asignadas en base a varios criterios: relevancia, confianza, alcance, transparencia de método, etc.

Para cada fuente se registraron los siguientes parámetros:

- El último año de publicación
- El ámbito (mundial, América Latina, España ...)
- Si la fuente se actualiza con frecuencia o no
- El tipo de fuente (por ejemplo, meta-información)
- El área de aplicación de la fuente (por ejemplo, Facebook)
- Si es específica para el idioma estudiado o no.

Las diferentes fuentes consultadas serán divulgadas próximamente en un metasitio con los parámetros correspondientes a fin de mantener un observatorio permanente. Mientras tanto, el grado de rápida obsolescencia de las fuentes y la dinámica de creación y eliminación de páginas de Internet es tal que no es apropiado citarlas, puesto que esta encuesta se terminó de realizar a finales del año 2013. Es el caso, por ejemplo, del sitio Socialbakers (Socialbakers, 2015), utilizado para medir la utilización de algunas redes sociales (LinkedIn, Facebook, Twitter, Google+), que ha cambiado la disposición de sus páginas desde entonces.

Datos demográficos y demo-lingüísticos

Como bien se sabe, los países monolingües son excepción, siendo el multilingüismo la regla, incluso si la adopción de una lengua oficial en la mayoría de los países pueda crear confusión. Ya sea en España, México, Estados Unidos, China o Camerún, o incluso en micro-estados como Mónaco, Malta o Singapur, el multilingüismo está siempre presente.

La mayoría de los estudios que han sido consultados proporcionan datos por país o región y tienden a extrapolar los resultados a la lengua oficial de cada país, sin tomar en cuenta las otras lenguas habladas[9], lo que es una simplificación que conlleva a errores importantes. Repartir los datos de número de hablantes proporcionalmente a su presencia en el país sería otra simplificación, por supuesto, porque la brecha digital no se distribuye uniformemente en la población y los inmigrantes a menudo tienen menos acceso a la Internet. Sin embargo es un error más aceptable, pues así se asegura una mejor toma en cuenta de los idiomas de un país, en vez de ocultar totalmente esta diversidad.

En la mayoría de los países del mundo se necesita contar los hablantes de diferentes lenguas si se quiere transformar los datos por país (los cuales son las fuentes primarias en la mayoría de los casos) en datos por idioma (necesarios para este estudio).

Para poder comparar una lengua (en este caso el español) con las otras, es necesario establecer estadísticas fiables, no necesariamente para todos los idiomas del mundo, pero por lo menos aquellos que son hablados por una amplia mayoría. Existen estadísticas confiables para algunas lenguas habladas en territorios bien definidos y que conocen un importante desarrollo (por ejemplo, en Europa), siempre que tengan un estatus oficial o instituciones de tutela[10] y que la diáspora que aún lo habla esté bien estudiada (es el caso del español en los Estados Unidos, en Francia, en Brasil, por ejemplo). Pero esta tarea se complica para los idiomas que no cumplen ninguna de estas condiciones.

Dificultades con lenguas sin organismo de tutela

Podemos, por ejemplo, señalar el caso de lenguas como el occitano (hablado en España, Francia e Italia), el piemontés (hablado en Italia) o el franco-provenzal (hablado en Francia, Italia y Suiza), para los cuales, a pesar de que se hablan en zonas desarrolladas y en territorios bastante bien definidos, la falta de institución de tutela para el conjunto de la lengua[11] afecta a la calidad de las cifras sobre el número de hablantes.

Dificultades con lenguas habladas en territorios extensos

Para algunos idiomas hablados en extensos territorios y diversas condiciones socioeconómicas, hay estadísticas relativamente fiables; este es el caso del francés o español, por ejemplo, no así para otros idiomas con características similares, tales como el inglés, el chino, el portugués[12] o el árabe para los cuales hay diferencias muy importantes entre las fuentes, en particular en el conteo de los que lo hablan como segundo idioma.

Fuentes demo-lingüísticas en conflicto

A las complicaciones ya mencionadas se suman las divergencias en las metodologías de conteo de los hablantes en los numerosos estudios demo-lingüísticos. Poder comparar todos los estudios disponibles sobre el número de hablantes de todas las lenguas en el mundo para crear una tabla comparativa completa es una tarea enorme, por encima de los recursos de este estudio. Por lo tanto, es impensable poner en paralelo los resultados de diferentes estudios sin un análisis detallado, porque pondría al mismo nivel cifras obtenidas por definiciones no comparables o métodos no homogéneos.

Tipología de lenguas

Un problema adicional que hay que resolver es la tipología de las lenguas. ¿Debe ser tomado el idioma alemán como un todo incluyendo los distintos

dialectos, a veces muy distantes unos de otros? ¿Debemos considerar el árabe literario -o sólo el clásico- o todos los árabes dialectales (árabe argelino, árabe levantino, árabe egipcio, etc.)?

Algunos especialistas aprobarán unas, otros aprobarán otras, pero no están calificados los autores de este estudio para tomar decisiones al respecto.

Segunda lengua (L2)

Pero sin duda, el dilema más agudo es la forma de tener en cuenta los hablantes "L2" (quienes, sin ser su lengua materna, la utilizan seguido o la dominan en un grado satisfactorio) para las lenguas vehiculares. En este estudio, anotamos L1 para los hablantes de lengua materna y L2 para los hablantes que utilizan con fluidez ese idioma, no materno para ellos.

Este aspecto es particularmente importante para una serie de lenguas que permiten un mayor acceso a la información o su utilización posibilita una mayor difusión de ideas, puesto que aquellos que la dominan suficientemente, sin que sea su propia lengua, la utilizarán en prioridad en ciertos contextos específicos, e Internet no escapa a dicha situación.

Hay muchas lenguas vehiculares en el mundo, pero son aquéllas que han tenido un pasado colonial o imperial importante las que han adquirido un lugar relativo de *prestigio* (entendido como la influencia o autoridad que le confiere la comunidad a dicha lengua) y, por lo tanto, interesa medir la población que pueda utilizarla en la Internet como lengua de comunicación. Es por dicha razón que el inglés se posiciona en primer lugar entre las lenguas mundiales (a pesar de tener muchos menos hablantes vernáculos que el mandarín o el español) y que el francés, el portugués, el ruso, el alemán, el mandarín o el malayo indonesio son porcentualmente mucho más utilizados de lo que el número de hablantes vernáculos dejaría suponer.

Ahora bien, el número de hablantes vernáculos es consecuencia del grado de penetración de una lengua en el territorio que fuera dominado. Es sabido que el español tuvo un grado de penetración mucho mayor en los territorios colonizados que el inglés, el francés o el portugués, por ejemplo. Es así como habría unos 400 millones de hablantes de español lengua materna en los países en que la misma es oficial (en los que viven unos 440 millones de habitantes), mientras que el francés, por ejemplo, es lengua materna de unos 70 millones de individuos, pero es oficial en países en los cuales viven unos 800 millones de habitantes, lo que le permite ser lengua primera y segunda (entendida esta última como lengua de uso sostenido) de unos 230 millones de individuos.

Es así que para lenguas como el inglés, el portugués, el ruso o el alemán (aun muy utilizado como lengua segunda en el Este de Europa), es importante el número de hablantes vehiculares a la hora de medir su uso en la Internet. Es mucho menos pertinente su cálculo para la lengua española, puesto que el número de hablantes de lengua segunda (L2) es poco significante (aunque numeroso en cifras) con respecto al número de hablantes de lengua materna (segunda lengua materna en el mundo luego del mandarín).

Por esta razón, al realizar mediciones que comparen los hablantes de lengua materna exclusivamente, los resultados muestran una neta predominancia del español con respecto a otras lenguas (como el francés, el portugués, el alemán, el ruso, etc.), pero no así en ciertos segmentos especializados en los cuales otras lenguas pueden ser más utilizadas por su carácter vehicular en grandes regiones (y el reservorio de hablantes de lengua segunda que eso representa) y la lengua española puede retroceder en uso global. Es importante notar que este aspecto es válido para segmentos especializados pero mucho menos para segmentos en los cuales el discurso es informal y familiar, en los cuales el uso de la lengua materna se afianzará.

Opciones demo-lingüísticas

Para tomar mejor cuenta de todos estos elementos se ha seleccionado una serie de opciones para el mejor equilibrio entre la homogeneidad y fiabilidad de los datos demo-lingüísticos.

1) Ethnologue para contabilizar los hablantes de lengua materna (L1)

Al contabilizar los hablantes de la lengua materna, la elección se hizo de forma natural en Ethnologue (Ethnologue, 2013), la única fuente que proporciona cifras actualizadas continuamente sobre todas las lenguas del mundo. Esta fuente es a menudo inexacta en sus cifras, frecuentemente incompleta y sus actualizaciones no son homogéneas, sin embargo es la única que puede proporcionar datos dinámicos de todos los idiomas del mundo mediante la aplicación de una metodología relativamente consistente[13]. La resolución expuesta y explicada[14] por los autores del "Barómetro Calvet" (Calvet, 2012) inspiró esa decisión.

Wikipedia (Wikipedia, lista de países por población, 2014) era una opción, pero la famosa enciclopedia no ofrece una metodología estable

para el recuento de los hablantes (y por una buena razón, ya que la naturaleza de la Wikipedia no es proporcionar información centralizada, cada artículo es independiente de los otros). Mientras que Wikipedia está tomando Ethnologue como su fuente principal para la mayoría de los idiomas inventariados, para otros idiomas toma como fuentes diversos estudios, ofreciendo diferentes metodologías y, por tanto, no comparables entre sí. Por otra parte, para algunos idiomas, Wikipedia es cauteloso y da rangos y figuras no precisas[15]. El caso de Albania, por citar una lengua que podría tener datos más precisos debido a la antigüedad de los estudios que se han dedicado, es esclarecedor. Ethnologue menciona 15.000 hablantes en Turquía, mientras que el artículo de Wikipedia, en diciembre de 2013, mencionaba cerca de tres millones de hablantes en este país, al mismo tiempo que incluía Ethnologue entre las fuentes.

Los profesionales de la lengua son, a menudo, muy críticos con los datos de Ethnologue[16] y haberlo elegido para este estudio no acarrea, obviamente, sólo ventajas; sin embargo, los inconvenientes afectan solo marginalmente este estudio porque se aplican sobre todo en idiomas identificados en estadísticas para las lenguas que no son objeto de este estudio.

2) Wikipedia para los datos demográficos

Después de estudiar varias fuentes para obtener buenas cifras de datos demográficos de todos los países del mundo, se utilizó Wikipedia en francés, que parecía federar (a fines de 2013) las mejores y más actualizadas fuentes[17].

3) Situaciones específicas

Se tomaron ciertas decisiones arbitrarias en determinados casos específicos, por falta de precisiones por parte de las fuentes consultadas, puesto que algunas de ellas tuvieron en cuenta la macrolengua, mientras que otras sólo una de las lenguas de la familia. A modo de ejemplo, el alemán, el árabe, el chino, etc. En este caso la macrolengua se ha tomado como única referencia con el fin de permitir la comparación. Por lo tanto, cuando hablamos de chino, será todos los idiomas hablados en China, así como para el árabe, alemán, malayo, etc.

Aunque estas consideraciones tienen un impacto pequeño en los resultados de este estudio, centrado en el español, se reportan desde un punto de vista crítico porque estas decisiones podrían parecer contradictorias con un proceso normal de tratamiento de variantes lingüísticas.

5. MÉTODO DE EVALUACIÓN GLOBAL

En las páginas precedentes se han presentado muchos registros diferentes y, en general, independientes, de la presencia del español en la Internet en relación con una selección de espacios o aplicaciones. Lo que sigue es un intento de dar un resultado significativo y completo sobre el lugar que ocuparía el español en la Internet, a partir de esta amplia colección de datos. En otras palabras, la intención de ese capítulo es la de ofrecer una evaluación global de la presencia de la lengua española en la Internet, tratando de resumir el conjunto de datos en un resultado único, a sabiendas de que cada dato, tomado individualmente, sólo expresa el rango que la lengua española tiene en el espacio o aplicación estudiado.

Una aproximación simplista de determinar un coeficiente común consistiría en calcular el **promedio del rango** de la presencia del español a partir de la suma de datos y tomarlo como la expresión resumida y global del rango del español en la Internet.

Una aproximación más satisfactoria consistiría en atribuir a cada espacio o aplicación un peso relativo, de acuerdo con su importancia como indicador de la presencia de la lengua en la Internet, y calcular el **promedio ponderado de los rangos**, lo que daría más sentido a una estimación global.

Una última posibilidad, más satisfactoria aún, sería la de establecer una serie de parámetros de calificación para cada resultado de la serie de datos, en base a elementos de credibilidad, y calcular así el **promedio ponderado multi-criterio** a partir de una ecuación que tome debidamente en cuenta este conjunto de parámetros.

Se decidió entonces mostrar las tres aproximaciones, para poder compararlas entre sí, dando paso a una clasificación global que pudiese integrar todos los resultados y reflejar con mayor precisión el lugar estimado del español en la Internet.

Las clasificaciones obtenidas se presentan en la tabla de presentación de los resultados (tabla II), ordenadas por aplicación o espacio. La clasificación del español en la Internet varía, según la aplicación o espacio estudiado, **entre la posición 1 y la 7** con un **promedio simple de 3,12**, lo que, en definitiva, no la aleja mucho de su posición de tercera lengua en términos de número de internautas. Esto está graficado en la columna L1 de la tabla II, la cual expresa el rango medido del español en relación con cada espacio o aplicación. Así, por ejemplo, 1 indica que el español es la primera lengua para la aplicación Gigasize o 7 indica que el español es la séptima lengua, en el caso de Amazon.

Una ponderación simple permitiría asignar un peso para cada elemento (indicada "**P"** en la tabla de presentación de los resultados), para reflejar la importancia relativa de tal espacio o aplicación como indicador de la presencia de la lengua en la Internet. Así, por ejemplo, se ha asignado una nota de 10 a los estudios específicos que analicen directamente la lengua de los usuarios de Internet o a aquéllos que mencionen un "porcentaje de páginas en español" y un peso de 3, a aplicaciones como Hi5 o ccSearch, que no cuentan con especificación lingüística específica.

Con los valores propuestos, se puede establecer un **promedio ponderado**, suma de los productos (peso x rango) dividido por la suma de los pesos. En este caso, el **promedio ponderado sería de 3,27.**

Una ponderación relativamente más complicada (que llamamos **multi-criterio**) y que aporta una mayor precisión sobre la importancia de los parámetros, sería la de calcular "**I**" (indicador del lugar de la lengua española en la Internet) a partir de la ecuación: **I** = A x B x C x D / 100, graficada en la Tabla II, en donde:

A = Grado de globalización del espacio o la aplicación considerado (0 a 10). Parámetro que refleja la relevancia del espacio o aplicación en cualquier lugar del globo. Así, Viadeo, red social especialmente usada en Francia tiene una nota de 2 en este estudio sobre la lengua española mientras que Facebook que es utilizado a nivel casi mundial (con las excepciones notables de China y Rusia, entre otros), llega a una nota de 8.

B = Grado de fiabilidad de los valores de este parámetro (0 a 10).

C = Nivel de confianza de los datos obtenidos para el español (0 a 10).

D = Relevancia del parámetro para la lengua española (de 0 a 10).

L1 = Clasificación del español para cada espacio o aplicación estudiada.

P = Ponderación simple del espacio o aplicación.

En este caso, el promedio resultando de la ecuación sería de 3,52. Por tanto, es razonable concluir que, en base a los parámetros establecidos y los resultados reunidos, el lugar relativo del español respecto a las otras lenguas en la Internet, según todos los criterios incluidos, sería **entre la tercera y la cuarta lengua**.

Si se comparan estos resultados con el lugar que ocupa la lengua española respecto a otras lenguas en el concierto mundial, se aprecia una situación aparentemente holgada gracias tanto a la importancia demográfica como a la relativa buena penetración de Internet en los países en los cuales se habla.

Sin embargo, incluso si la lengua española es la segunda en términos demográficos, y tercera tanto en número de países en los cuales es oficial como en cuanto lengua de enseñanza, se encuentra aún rezagada en términos relativos en el mundo virtual y queda trabajo si se desea que haya una coherencia entre uso y contenido.

Incluso si las instituciones encargadas de su tutela afirman sostenidamente un presunto segundo rol en el concierto de lenguas internacionales, es difícil poder afirmarlo en materia de relaciones internacionales o económicas, de comunicación científica, de normativa internacional o de patentes industriales, así como en materia de número de traducciones tanto literarias como especializadas, de industrias culturales o de turismo. Las lenguas francesa, alemana, china, rusa o japonesa ocupan muchas veces una situación superior a la castellana en esos sectores. Lo mismo pasa con la posición del español en la Internet, potencialmente importante como vector de comunicación pero que requiere mas esfuerzos en términos de producción de contenidos especialmente en materia de presencia de libros y de páginas web como lo muestra la tabla III.

El control regular de todos estos resultados permitiría una observación del lugar del español por Internet, muy útil para determinar las políticas a seguir por el apoyo y la mejora de su presencia.

6. CONCLUSIÓN

Los enfoques metodológicos, propuestos en un campo caracterizado por una crisis prolongada, permiten superar parcialmente las lagunas existentes en la información sobre la presencia de las lenguas en la Internet y dan una contribución original al tema. Hay una probabilidad razonable de que puedan adaptarse, sin muchos cambios, para otros idiomas fuera del francés y el español.

Los resultados del español en la Internet son prometedores dada la gran proporción de internautas hispanohablantes,[18] pero la presencia de contenidos en lengua española parecería no estar acorde con el número de utilizadores. Es probable que sean necesarias ciertas acciones voluntaristas (digitalización de libros, incentivos a la creación de artículos científicos en español, incentivos a crear artículos Wikipedia, alfabetización digital, e-gobierno o e-administración, etc.) tal como lo han promovido ciertos grupos lingüísticos (francés, lenguas nórdicas, lenguas minoritarias, etc.) permitiendo una presencia lingüística proporcional o superior a la de usuarios.

Tabla II. Posición de la lengua española en Internet. Evaluación general por espacios y aplicaciones

Espacio o aplicación	A	B	C	D	I	L1	P	L1x I	L1 x P	Tipo
3G	10	6	9	6	32	4	3	130	12	Infraestructura
Amazon	7	9	9	8	45	7	6	318	42	Libros
AOL/AIM	5	7	7	6	15	3	3	44	9	Aplicaciones
Badoo	6	5	7	5	11	2	3	21	6	Redes sociales
Banda ancha	10	9	9	7	57	4	4	227	16	Infraestructura
Bitshare ++	7	7	7	6	21	3	4	62	12	P2P
Blogs.com	6	7	7	5	15	3	5	44	15	Blogs
Ccsearch	6	7	7	6	18	3	3	53	9	Aplicaciones
Facebook	8	7	7	6	24	2	7	47	14	Redes sociales
Foofind	7	7	7	6	21	3	3	62	9	Aplicaciones
Gigasize	7	7	7	6	21	1	4	21	4	P2P
Gmail	7	5	8	6	17	3	6	50	18	Aplicaciones
Google +	8	7	7	6	24	3	7	71	21	Redes sociales
Hi5	7	6	7	4	12	1	3	12	3	Redes sociales
Hotmail	5	5	6	6	9	2	4	18	8	Aplicaciones
Icq	5	7	7	5	12	5	3	61	15	Aplicaciones
Instagram	7	5	7	6	15	2	5	29	10	Redes sociales
Internet archive	9	7	7	9	40	3	7	119	21	Contenidos
Internetworldstats	10	6	10	10	60	3	10	180	30	Usuarios
Ixquick	6	7	7	6	18	2	3	35	6	Aplicaciones
Linkedin	7	7	7	7	24	3	6	72	18	Redes sociales
Live journal	8	7	7	7	27	3	5	82	15	Blogs
MSN	7	7	7	6	21	3	5	62	15	Aplicaciones
Ning	7	7	7	8	27	3	5	82	15	Redes sociales
Open directory	9	10	7	9	57	5	7	284	35	Contenidos
Open office	9	9	9	8	58	5	5	292	25	Aplicaciones
Orkut	2	5	7	3	2	3	6	6	18	Redes sociales
Rapidshare	7	7	7	6	21	2	4	41	8	P2P
Servidores / hab.	10	9	9	7	57	2	5	113	10	Infraestructura
Skype	8	7	7	8	31	4	7	125	28	Aplicaciones
Smartphones	10	6	8	9	43	4	4	173	16	Infraestructura
Telefonia móvil	10	9	9	7	57	2	3	113	6	Infraestructura
Tumblr	6	6	7	6	15	2	4	30	8	Redes sociales
Twitter	9	8	7	9	45	2	8	91	16	Redes sociales
Viadeo	2	5	7	10	7	5	6	35	30	Redes sociales
W3techs	10	7	10	10	70	5	10	350	50	Páginas
Wikipedia	9	10	10	10	90	6	8	540	48	Contenidos
Windows live	7	7	7	6	21	3	4	62	12	Redes sociales
Wordpress	8	7	7	7	27	2	5	55	10	Blogs
Yahoo!	5	5	6	6	9	3	4	36	16	Aplicaciones
Youtube	8	7	7	8	31	2	7	63	14	Video
Promedios						**3,12**		**3,27**	**3,52**	

Ver definición de las columnas en el texto.

 Rev. Esp. Doc. Cient., 39(3), julio-septiembre 2016, e141. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.3.1328

Tabla III. Posición de la lengua española en Internet. Valores por tipo de segmento

TIPO DE SEGMENTO	Posición	Número de aplicaciones o espacios estudiados
Aplicaciones	3,7	11
Blog	2,6	3
Contenidos y páginas	5	4
Infraestructura	3,1	5
Libros	7	1
P2P	2	3
Redes sociales	2,8	12
Usuarios (Número de)	3	1
Video	2	1

7. AGRADECIMIENTOS / ACKNOWLEDGEMENTS

Esta investigación se realizó en base a un trabajo específico para la lengua francesa, financiado originalmente por la *Organisation internationale de la Francophonie* (OIF) y consultable en línea. Álvaro Blanco colaboró en la etapa inicial de recopilación de fuentes.

This research is a result of a specific work for the French language, originally funded by the *Organisation internationale de la Francophonie* (OIF) and avialable online. Álvaro Blanco contributed in the initial stage of gathering sources.

8. NOTAS

1. Lenguas romances (catalán, español, francés, italiano, portugués, rumano) así como el inglés y el alemán.
2. Ver la rúbrica «Usage of content languages for websites».
3. Ver la secuencia del español en http://w3techs.com/technologies/details/cl-es-/all/all o también la de las lenguas con más de 0,1% de presencia en: http://w3techs.com/technologies/history_overview/content_language
4. En 2007, el trabajo de Funredes / Unión Latina calcula un valor de poco más de 26% para este indicador, y un valor de poco más de 57% para el porcentaje de francés en las páginas situadas en Francia (incluyendo las del .fr tanto como las de los dominios genéricos como .com o .org).
5. Según Netmarketshare Google sería, en abril 2015, el motor de búsquedas más utilizado con 62% pero con tendencia a la baja desde 2010: http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0&qpcustom=
6. En 2008, varias fuentes mencionaban unas 127 mil millones de páginas indexadas (en particular el motor de búsqueda CUIL, ahora desaparecido, que afirmaba explorar la Web entera: http://web.archive.org/web/20100916001435/http://www.cuil.com/) y se puede encontrar fácilmente varias fuentes que mencionan la cifra de 40 mil millones para las páginas indexadas por Google (http://www.openinnova.es/como-hace-google-para-indexar-las-paginas-web/), cuando esta cifra es de 2008.
7. Voz por Internet
8. Aunque Youtube tiene un competidor en segunda posición, Dailymotion, con fuerte presencia francesa, puesto que pertenece a capitales franceses.
9. Es el caso de InternetWorldStats, por ejemplo, y uno de los escasos contraejemplos es una vez más Wikipedia que provee datos lingüísticos de calidad en http://meta.wikimedia.org/wiki/List_of_Wikipedias/sortable
10. Entendemos por organismos de tutela, toda aquella entidad pública, parapública o asociativa que tenga por misión la de promover la difusión, promoción y/o desarrollo de la lengua en conjunto, como lo pueden ser diferentes direcciones de políticas lingüísticas, academias de la lengua, asociaciones locales o regionales de promoción cultural, etc.
11. Mencionamos la falta de organismo de tutela para el conjunto de la lengua, dado que muchas veces hay entidades que cumplen con dicha función, pero a nivel local y no global. Es el caso del aranés en España, variante del occitano, que cuenta con una protección de parte de las autoridades de Cataluña, pero no existen entidades de tutela para el conjunto de la lengua occitana.
12. El instituto Camões ofrece desde hace poco cifras con un margen de error aceptable.
13. Ethnologue publicó una nueva versión en línea con muchos cambios en los datos estadísticos después que terminamos ese estudio. Debe quedar claro que los datos demolingüísticos de Ethnologue tomadas en este estudio son aquellos publicados en mayo 2013.
14. http://wikilf.culture.fr/barometre2012/tmpl.php?data=doc/methodologie/index
15. Es el caso el inglés por el cual indica 309 < L1 < 380 millones.
16. Varios problemas también han sido identificados y registrados durante el curso de este estudio.
17. http://fr.wikipedia.org/wiki/Liste_des_pays_par_population
18. Según InternetWorldStats, el porcentaje de hispanohablantes conectados sería de 50,6% en 2013 a comparar con las cifras de 39% por la población mundial, de 58,4% para el inglés, 46,7% para el portugués y de 20,9% para el francés.

9. REFERENCIAS

Nota: Salvo indicación específica, las fuentes web citadas en este artículo se entienden consultadas entre finales del año 2013 e inicios del año 2015.

Alexa - http://alexa.com [diciembre 2014].

Calvet L.-J. (2012). Poids des langues (Baromètre Calvet) - http://wikilf.culture.fr/barometre2012/ [marzo 2014].

CISCO, *Cisco Visual Networking Index* - http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/ [diciembre 2014].

Dilinet - http://dilinet.org [diciembre 2014].

Estudio lenguas y ciberespacio en *Unión Latina* http://dtil.unilat.org/LI/2007/index_es.htm [julio 2012].

Ethnologue: Languages of the World - http://www.sil.org/ethnologue [mayo 2013].

Funredes, Fundación Redes y Desarrollo - http://funredes.org [diciembre 2014].

Grefenstette y Nioche J. (2001). Estimation of English and non English Language use on the WWW. Technical report from Xerox Corporation Center Europe. - http://arxiv.org/ftp/cs/papers/0006/0006032.pdf [marzo 2014].

InternetWorldStats - http://www.internetworldstats.com/stats7.htm [diciembre 2013].

LOP, Language Observatory Project - http://gii2.nagaokaut.ac.jp/gii/blog/lopdiary.php [julio 2012].

Maaya, Red Mundial por la diversidad lingüística - http://maaya.org [mayo 2015].

Netcraft - http://news.netcraft.com/archives/category/web-server-survey/ [marzo 2015].

Net.Lang : Réussir le cyberespace multilingue / Towards a multilingual cyberspace (2012), Paris, C&F Éditions. Consultable igualmente en http://net-lang.net/ [diciembre 2014].

Observatorio de lenguas y culturas en la Internet en Funredes - http://funredes.org/lc [julio 2012].

OIF (Organisation internationale de la francophonie). La langue française dans le monde (2014) - http://www.francophonie.org/Langue-Francaise-2014/ [junio 2016].

Paolillo J.; Pimienta D.; Prado D.; Mikami Y.; Zaki abu Bakar A.; Sonlertlamvanich V.; Vikas O.; Pavol Z.; Zaidi abdul Rozan M.; Nagy János G.; Takahashi T.; Fantognan X. (2005). *Mesurer la diversité linguistique dans l'Internet*, Paris, UNESCO – Consultable igualmente en http://unesdoc.unesco.org/images/0014/001421/142186f.pdf [diciembre 2014].

Pimienta D. (2011). Chapter 12, Language and Content, pp183-197, in *Accelerating Development Using the Web*. Gorge Sadowsky (ed.), Word Wide Web Foundation. http://g3ict.org/download/p/fileId_928/productId_241.

Pimienta D.; Prado D.; Blanco A. (2009). *Douze ans de mesure de la diversité linguistique dans l'Internet: bilan et perspectives*, Paris, UNESCO – Consultable igualmente en http://unesdoc.unesco.org/images/0018/001870/187016f.pdf [diciembre 2014].

Prado D.; Pimienta D.; Lemoulinier A. (2010). Diversité linguistique et cyberespace : état de l'art, enjeux et opportunités en *Cosmopolis,* - http://agora.qc.ca/cosmopolis.nsf/Articles/no2010_1_Diversite_linguistique_et_cyberespace___etat_de_l?OpenDocument [diciembre 2014].

Socialbakers - http://www.socialbakers.com/statistics/ [diciembre 2015].

Suzuki I.; Mikami Y.; Ohsato A.; Chubachi Y. (2002). A Language and Character Set Determination Method based on N-gram Statistics, ACM Trans. on Asian Language Information Processing, Vol 1 N3, pp. 270-279. http://dx.doi.org/10.1145/772755.772759

UIT, (2010). Monitoring the WSIS Targets. A mid-term review, Target 9 (content) en *World Telecommunication/ICT Development Report 2010 (pp175-192)*. Ginebra, UIT – Consultable igualmente en http://www.itu.int/pub/D-IND-WTDR-2010 [diciembre 2014].

Unión Latina - http://unilat.org/ (consultado en mayo 2015).

Union latine (2010): *Présence, poids et valeur des langues romanes dans la société de la connaissance, actes de la journée d'étude du 30 avril 2010* (sous la direction de Daniel Prado), Paris. Unión Latina

W3Techs - http://w3techs.com/technologies/overview/content_language/all [diciembre 2014].

Wikipedia, lenguas del mundo - https://es.wikipedia.org/wiki/Familia_de_lenguas [diciembre 2014].