A lingua galega en Internet após dúas décadas

Xavier Gómez Guinovart
Universidade de Vigo

Resumo:

Neste artigo examino a evolución da presenza do galego en Internet a partir da comparación da situación actual coa debuxada hai dúas décadas en Gómez Guinovart (2003). En concreto, o obxectivo deste estudo hase centrar na avaliación do tamaño da web en galego en relación co das linguas do seu contorno cultural (inglés, español, francés, italiano, portugués, catalán, alemán, asturiano e mais éuscaro) e na análise comparativa desta estimación cos datos obtidos en 2003. A metodoloxía utilizada para cuantificar o espazo ocupado na web polos distintos idiomas parte da utilizada en diversas medicións pola Unesco entre 1996 e 2005, consistente en empregar motores de busca e un conxunto de conceptos léxicos para medir a presenza proporcional destes conceptos nas súas diversas equivalencias lingüísticas.

Palabras chave:

Diversidade lingüística; Internet; web; galego; política lingüística.

The Galician language on the internet after two decades

Abstract:

In this article I examine the evolution of the presence of Galician on the Internet by comparing the current situation with that of two decades ago (Gómez Guinovart 2003). Specifically, the aim of this study is the evaluation of the size of the Galician web in relation to that of the languages of its cultural environment (English, Spanish, French, Italian, Portuguese, Catalan, German, Asturian and Basque) and the comparative analysis of this estimate with the data obtained in 2003. The methodology used to quantify the space occupied on the web by the different languages is based on that used in various measurements by Unesco between 1996 and 2005, which make use of search engines and a set of lexical concepts to measure the proportional presence of these concepts in their various linguistic equivalents.

Key words:

Linguistic diversity; Internet; web; Galician; language policy.

1. Introdución

Hai xa case vinte anos publiquei un achegamento á extensión da lingua galega en Internet (Gómez Guinovart, 2003) que, pola súa orixinalidade, suscitou algún interese académico no seu momento. Os datos que se discutían naquel artigo estaban baseados nunha estimación do tamaño da web en galego realizada a partir do índice do buscador AlltheWeb, un sistema noruegués de recuperación da información en Internet que gozou dunha certa popularidade entre 1999 e 2003 polas súas posibilidades de busca, máis avanzadas das que ofrecía Google naquela altura.

Neste novo artigo hei de tentar explorar a situación actual na que se atopa o galego na web dos nosos días e actualizar os datos fornecidos sobre este particular no devandito traballo. Máis concretamente, o obxectivo deste estudo que aquí se presenta será a avaliación do tamaño da web en galego en relación co das linguas do seu contorno cultural e a comparación desta estimación coa realizada hai dúas décadas.

Aínda que o foco desta análise estea centrado na web, é evidente que esta non é a única plataforma da Internet onde se pode tentar medir a extensión atinxida por unha lingua. Sen dúbida, é posíbel obter outras perspectivas interesantes para o estudo da diversidade lingüística no ciberespazo a partir da análise da actuación lingüística nos blogs (Dopazo & Lombao, 2006; Canabal & Gago & López & Isasi & Limia & Pereira, 2008), nas redes sociais como Twitter ou Facebook (Honeycutt & Cunliffe, 2010; Caulfield 2013), na Wikipedia (Van Dijk, 2009) ou en YouTube (Cunliffe & Ap Dyfrig, 2013), por mencionarmos unicamente algúns dos sectores de Internet que mereceron o maior grao de atención por parte da comunidade investigadora nos últimos anos. Destas áreas, por agora só contamos con datos certos sobre os usos lingüísticos na Wikipedia, grazas ás estatísticas publicadas pola propia fundación Wikimedia[1], responsábel desta monumental enciclopedia colaborativa plurilingüe. Algúns destes datos estatísticos sobre as linguas na Wikipedia serán usados máis adiante na sección dedicada á comparación dos resultados deste estudo con outros indicadores lingüísticos.

2. Metodoloxía

A tarefa de estimar o espazo ocupado por unha determinada lingua na web non é doada (Gerrand, 2007; Pimienta & Prado, 2015). En principio, ábrense tres posibilidades de pescuda, baseadas en tres indicadores: o número de páxinas web escritas nesa lingua, a cantidade de palabras incluídas nesas páxinas web e o número de sitios web que fornecen esas páxinas.

1 https://wikimediafoundation.org/

A primeira opción, baseada no cómputo das páxinas web nun idioma, constitúe a metodoloxía usada con diferentes variacións técnicas en diversos traballos académicos nos últimos anos (Grefenstette & Nioche, 2000; Mas i Hernàndez, 2005; Pimienta, 2005; Pimienta & Prado & Blanco, 2009). Neste tipo de estudos, empréganse os resultados obtidos para certas consultas, en buscadores como Google ou Bing, como índice da cantidade de páxinas web nunha determinada lingua. Porén, a utilización destes motores de busca como ferramenta de investigación non está exenta de dificultades. En primeiro lugar, porque non todas as páxinas están indexadas polos motores de busca: as páxinas que residen nas denominadas *web profunda* e *web escura* nunca han aparecer nos resultados das consultas nos buscadores convencionais utilizados nestes estudos. En segundo lugar, porque as cantidades de resultados indicadas por Google ou Bing como resposta a unha consulta non son moi fiábeis e, en moitas ocasións, son claramente erróneas. E, en terceiro lugar, atendendo ao noso particular interese investigador centrado na nosa lingua, porque os dous motores de busca con maior capacidade de consulta (Google e Bing) non permiten por agora limitar os resultados da busca ás páxinas escritas en galego.

A segunda posibilidade de pescuda, baseada na cantidade de palabras incluídas nas páxinas web nunha lingua, resulta tecnicamente moi difícil de encarreirar na investigación universitaria e, no seu caso, só podería ser asumida por unha entidade propietaria dun motor de busca na web a partir dos datos do seu índice.

Finalmente, a terceira opción considerada, baseada no cómputo de sitios web que fornecen páxinas nunha lingua, resulta metodoloxicamente dubidosa, pois a rede está inzada de sitios que serven páxinas en máis dun idioma. Con todo, esa é a metodoloxía empregada por W3Techs, unha prestixiosa consultora en Internet que ofrece estatísticas actualizadas das linguas usadas nos sitios web, cos resultados que revisaremos más adiante na comparación dos resultados deste estudo con outros indicadores.

A metodoloxía usada no presente traballo parte da utilizada en diversas medicións pola Unesco entre 1996 e 2005, consistente en empregar motores de busca e un conxunto de conceptos léxicos para medir a presenza proporcional destes conceptos nas súas diversas equivalencias lingüísticas en inglés, español, francés, italiano, portugués, romanés, catalán e alemán (Pimienta & Prado & Blanco, 2009). Os conceptos léxicos interlingüisticamente equivalentes usados nestas medicións efectuadas pola Unesco correspóndense coa familia flexiva do lexema ou lexemas representantes do concepto en cada lingua. Por exemplo, para o concepto de "queixo", as formas para as que se documentaba a extensión da súa presenza nos buscadores eran *cheese*, *cheeses* para o inglés; *queso*, *quesos* para o español; *fromage, fromages* para o francés; *formaggio*, *formaggi* para o italiano; *queijo*, *queijos* para o portugués, *brânza*, *branza*, *brânze*, *branze*, *brânza*, *brânzele*, *branzele*, *brânzei*, *branzei*, *brânzelor, branzelor, brânzeturi*, *branzeturi*, *brânzeturile*, *branzeturile*, *brânzeturilor*, *branzeturilor* para o romanés; *formatge*, *formatges* para o catalán; e *kaese*, *kaesen*, *kase*, *kasen, käse, käsen* para o alemán.

Os criterios considerados na seleccións dos conceptos léxicos utilizados para estas estimacións non son en absoluto triviais e poden resumirse nos seguintes: (1) neutralidade cultural: rexéitanse os termos con significación cultural e, por tanto, maior frecuencia nalgún idioma; (2) homografía interlingüística: evítanse os termos que coinciden con palabras homógrafas noutros idiomas, xa que adulteran os resultados das buscas; (3) préstamos homógrafos: evítanse os termos nun idioma que se escriben igual noutros idiomas; (4) homografía por abreviatura: evítanse os termos que coinciden ortograficamente con abreviaturas noutros idiomas; (5) homografía con nome propio: evítanse os termos que coinciden ortograficamente con nomes propios noutros idiomas; (6) homografía por erro ortográfico: evítanse os termos que, escritos con algún erro ortográfico común, coinciden ortograficamente con palabras doutros idiomas; (7) polisemia: evítanse os termos polisémicos con significados que se expresan dun modo diferente nos outros idiomas; (8) morfosintaxe non equivalente: evítanse os termos que teñen distintos significados correspondentes a distintas categorías que se expresan dun xeito diferente nas outras linguas; (9) diversidade dialectal: evítanse os termos que posúen diversidade dialectal dentro dun ámbito lingüístico; e (10) diversidade ortográfica: evítanse os termos que manifestan diversidade ortográfica dentro dun ámbito lingüístico.

Nos seus informes realizados para a Unesco, os investigadores Pimienta & Prado & Blanco utilizan os resultados cuantitativos ofrecidos polos buscadores de Google e Yahoo/Altavista para 57 conceptos léxicos para produciren unha estimación da presenza relativa na web de cada lingua analizada en comparación coa do inglés. Como veremos, este método non permite determinar un valor absoluto en número de páxinas nin para toda a web nin para as linguas consideradas, senón unicamente unha estimación porcentual da presenza de cada lingua en relación ao peso na web do inglés, de uso innegabelmente maioritario neste ámbito.

Para este estudo decidimos adaptar o abano de linguas analizadas pola Unesco ao noso contorno cultural, abranguendo na nosa análise comparativa co galego todas as linguas examinadas por este organismo agás o romanés, coa adición do asturiano e do éuscaro. A presenza na web destas 10 linguas (galego, inglés, español, francés, italiano, portugués, catalán, alemán, asturiano e mais éuscaro) foi avaliada mediante os resultados obtidos en Google da consulta de 15 conceptos léxicos correspondentes aos seguintes elementos léxicos en galego: *hoxe*, *decembro*, *xuño*, *venres*, *febreiro*, *luns*, *xoves*, *xeonllo*, *coitelo*, *doenza*, *inmunidade*, *inestabilidade*, *homosexualidade*, *inmortalidade* e *heterosexualidade*. A elección de Google para o estudo baséase tanto na amplitude do seu índice, considerado o de maior extensión nos últimos anos[2], a moita distancia do seu competidor máis achegado (Microsoft

2 Véxanse, por exemplo, as estatísticas ao respecto fornecidas en https://www.worldwidewebsize.com/

Bing), coma na dispoñibilidade dunha API de consulta gratuíta[3] (con usos limitados) que, a diferenza da de Microsoft, non exixe a cesión dos datos da tarxeta de crédito para a súa utilización.

Inglés	thursday thursdays	knee knees	instability instabilities
Español	jueves	rodilla rodillas	inestabilidad inestabilidades
Francés	jeudi jeudis	genou genoux	instabilité instabilités
Alemán	donnerstag donnerstage donnerstagen donnerstages donnerstags	knie knien knies	instabilitaet instabilitaeten instabilität instabilitäten unbestaendigkeit unbestaendigkeiten unbeständigkeit unbeständigkeiten
Portugués	quinta-feira quintas-feiras	joelho joelhos	instabilidade instabilidades
Italiano	giovedì	ginocchio ginocchia ginocchi	instabilità
Catalán	dijous	genoll genolls	inestabilitat inestabilitats
Éuscaro	ostegunean ostegun osteguna osteguneko ostegunetan ostegunetik ostegunera	belaunean belaun belaunak belauna belaunen belaunetan belaunetaraino belaunez belaunetik belaunaren belaunek belaunetako	ezegonkortasuna ezegonkortasun ezegonkortasunak ezegonkortasunaren ezegonkortasunea
Galego	xoves quinta feira quintas feiras	xeonllo xeonllos	inestabilidade inestabilidades
Asturiano	xueves	rodiya rodiyes rodiella rodielles	inestabilidá inestabilidaes

Táboa 1. Exemplo de conceptos analizados

3 https://developers.google.com/custom-search

A Táboa 1 inclúe, a modo de ilustración, tres dos conceptos léxicos analizados, incluíndo os conxuntos de palabras asociados con cada concepto para cada un dos dez idiomas do estudo. Como se observa nos exemplos, nas linguas sen flexión nominal de caso (galego, inglés, español, francés, italiano, portugués, catalán e asturiano) téñense en consideración o singular e o plural dos elementos léxicos considerados sinónimos do concepto[4]. Para o alemán, por outro lado, tense en conta o paradigma completo da flexión nominal (xénero, número e caso). No caso do éuscaro, selecciónanse no corpus ETC (Egungo Testuen Corpusa)[5] de textos do éuscaro contemporáneo as verbas que agrupadas representen máis do 90% dos exemplos documentados de uso. Así, para o lema *ostegun* "xoves", cun total de 30 formas flexivas diferentes documentadas no corpus, selecciónanse as formas *ostegunean* (51,72% dos casos), *ostegun* (15,69%), *osteguna* (9,19%), *osteguneko* (5,85%), *ostegunetan* (3,73%), *ostegunetik* (3,18%) e *ostegunera* (2,78%), sete formas flexivas do lema *ostegun* que sumadas representan máis do 90% das ocorrencias deste lema no corpus[6]. Cómpre observar que a declinación nominal do éuscaro pode alcanzar as 38 formas para un lexema nominal debido ao seu uso de morfemas flexivos para indicar a determinación (determinado/indeterminado), número (singular/plural) e caso (até 14 casos) da forma nominal flexionada.

En relación coas formas léxicas representantes dos conceptos léxicos nas dez linguas, unha limitación desta metodoloxía radica na dificultade de cumprir por completo os criterios de selección estabelecidos para a súa constitución, unha dificultade que aumenta de maneira directamente proporcional ao número de linguas que se queiran considerar no estudo.

Así, entre os exemplos recollidos na Táboa 1, a palabra asturiana *rodiya* conculca o criterio de evitar a homografía interlingüística na selección, xa que este termo coincide formalmente coa palabra usada en inglés para denominar unha comunidade de Sri Lanka e mais a súa lingua[7]. Certo que o inglés *rodiya* non é un vocábulo moi frecuente nese idioma, mais a súa frecuencia absoluta na web é claramente superior á da forma homógrafa asturiana, polo que o total de resultados no buscador para esta última forma é moi superior ao correspondente á súa presenza real. Outro caso máis de homografía interlingüística nos exemplos recollidos na Táboa 1 obsérvase na forma léxica *inestabilidades*, coincidente en galego e en español. Non resulta difícil detectar que a homografía interlingüística é un problema que resulta moi difícil de evitar cando se inclúe un certo número de linguas no estudo.

4 No caso do italiano, inclúense os dous plurais complementarios de *ginocchio (i ginocchi, le ginocchia)*.

5 https://www.ehu.eus/etc/

6 https://www.ehu.eus/etc/?bila=ostegun

7 https://en.wikipedia.org/wiki/Rodiya

Outra limitación desta metodoloxía provén da inconsistencia na cantidade de resultados ofrecidos polo buscador de Google, xa que o seu número é variábel para unha mesma consulta repetida ás veces con só un segundo de diferenza. Deste xeito, a consulta do termo *hoxe* buscado dez veces ao longo de dez segundos produce os seguintes dez resultados consecutivos: 3.140.000, 3.350.000, 3.260.000, 3.350.000, 3.350.000, 3.260.000, 3.140.000, 3.140.000, 3.140.000 e 3.570.000[8]. Como se pode apreciar, entre o máximo e o mínimo calculado (3.570.000 e 3.140.000) existe unha variación de 430.000 documentos de diferenza, o que representa unha variabilidade superior ao 10% no valor dos resultados. Estas diferenzas aumentan cando os termos buscados posúen unha alta frecuencia. Así, a consulta en Google repetida dez veces para o termo *today* fornece os seguintes resultados consecutivos nun lapso de vinte segundos: 8.750.000.000, 9.350.000.000, 8.460.000.000, 7.140.000.000, 8.580.000.000, 8.790.000.000, 7.760.000.000, 8.270.000.000, 9.160.000.000 e 9.600.000.000. Nese caso, a diferenza entre os resultados máximo e mínimo (9.600.000.000 e 7.140.000.000) ascende a 2.460.000.000 documentos, o que representa case un 30% de variabilidade. As causas desta variabilidade nos resultados son incertas, debido ao segredo que Google mantén sobre o funcionamento do seu buscador[9], aínda que posibelmente habería que atribuíla ao feito de que Google distribúa as peticións de consulta entre os seus numerosos centros operativos, ofrecendo cada un deles a súa propia resposta a partir dos seus datos. No noso experimento, realizamos todas as pescudas nun prazo curto de tempo e mantivemos os resultados ofrecidos polo buscador na primeira consulta efectuada para todos os termos seleccionados, asumindo como inevitábel neste contexto tecnolóxico un factor de mutabilidade nos resultados.

Aínda outra dificultade máis para a obtención de resultados axeitados mediante esta técnica provén da imprecisión dos resultados obtidos con Google na busca de secuencias de palabras. Este buscador ofrece, de maneira oficial, a posibilidade de procurar unha cadea exacta de palabras escribindo a secuencia buscada entre aspas[10]. Porén, este tipo de buscas en Google non sempre funciona dun modo axeitado[11]. Así, a procura do termo galego *quinta feira* (formulada na cela de consulta de Google como *"quinta feira"* entre aspas) inclúe entre os resultados todas as páxinas que conteñen o termo *quinta-feira* (isto é, co termo en ortografía portuguesa, escrito cun trazo entre as dúas palabras formantes), facendo imposíbel a

8 Consulta realizada o 18.02.21 mediante a API JSON de busca de Google usando o parámetro *exactTerms* na solicitude HTTP para indicar o termo exacto (*hoxe*) que deben conter todos os documentos incluídos nos resultados da busca (https://developers.google.com/custom-search/v1/reference/rest/v1/cse/list/).

9 En https://www.google.com/search/howsearchworks/algorithms/ pode consultarse a escasa información oficial que Google ofrece sobre o funcionamento dos seus algoritmos de busca.

10 https://support.google.com/websearch/answer/2466433

11 http://www.searchengineshowdown.com/features/google/inconsistent.shtml

segregación dos resultados. De feito, a busca formulada en Google como *"quinta-feira"* produce un número resultados semellante á busca formulada como *"quinta feira"*, sendo bastante evidente que o número de ocorrencias na web de *quinta feira* é moito menor que o de *quinta-feira*. No experimento que presentamos aquí, estas inconsistencias só afectan aos resultados fornecidos polo buscador para as formas litúrxicas dos días da semana en galego e portugués, por seren estas as únicas formas pluriléxicas incluídas no estudo. Estas incidencias trátanse de solucionar manualmente e de forma individual para axustar as medicións á realidade, dentro das nosas posibilidades.

Por último, unha limitación adicional do método de traballo adoptado nesta investigación radica na imposibilidade de determinar, mediante esta metodoloxía, un valor absoluto, en número de páxinas web, para a extensión na web das linguas consideradas, xa que esta técnica unicamente fornece porcentaxes da extensión na web de cada lingua en relación á do inglés. Por esta mesma razón, esta metodoloxía de investigación tampouco permite determinar un valor absoluto, en número de páxinas, para toda a web.

Con todo, a respecto deste punto, hai que salientar que actualmente non hai un consenso académico sobre cal é a cantidade de páxinas web indexadas por Google, nin tampouco sobre cal sería o mellor método para pescudar ese dato. Consonte as estimacións realizadas diariamente na Universidade de Tilburg (Van Den Bosch & Bogers & De Kunder, 2016), o tamaño da web de Google correspondente ás datas da consulta do noso experimento sería de 53'8 mil millóns de páxinas[12]. Con seguranza, esta cifra é bastante menor do número de páxinas web indexadas por Google e contrasta fortemente coas cifras divulgadas pola propia empresa pois, segundo o que expón Google na súa web, "the Google Search index contains hundreds of billions of webpages"[13]. Por suposto, centos de miles de millóns non é unha cifra moi exacta. A última vez que Google facilitou unha cifra concreta[14] foi en novembro de 2016: "Search starts with the web. It's made up of over 130 trillion individual pages and it's constantly growing"[15]. A variabilidade das estimacións sobre o tamaño da web de Google e a súa falta de certeza representan unha limitación adicional para o estudo, xa que nos impiden converter dun xeito fiábel os valores porcentuais obtidos por cada lingua en valores absolutos que ilustren dun modo máis claro a súa implantación na web.

12 Tamaño da web de Google tomado de https://www.worldwidewebsize.com/ o 11.01.2021.

13 https://www.google.com/search/howsearchworks/crawling-indexing/ (consultada en 01.03.2021).

14 http://web.archive.org/web/20180724152828/https://www.google.com/insidesearch/howsearchworks/thestory/ (consultada en 01/03/21.

15 https://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378 (consultada en 01.03.2021).

3. Resultados

Presentamos na Táboa 2 os resultados finais obtidos neste estudo, incluíndo na segunda columna para cada lingua o número total de páxinas web nas que aparecen os 15 conceptos analizados, segundo os datos fornecidos polo buscador Google[16]; e, na terceira columna, a porcentaxe de páxinas web acadada por cada lingua en relación á extensión da lingua inglesa.

	Páxinas web nos resultados	Porcentaxe a respecto do inglés
EN	24.355.920.800	100
ES	4.572.550.850	18,774
FR	2.442.139.950	10,027
DE	2.274.900.444	9,340
PT	1.449.436.400	5,951
IT	1.101.064.000	4,521
CA	96.577.257	0,397
EU	42.042.981	0,173
GL	30.512.050	0,125
AS	608.400	0,002

Táboa 2. Resultados por lingua

A Táboa 2 reflicte unha clasificación das linguas ordenada pola súa extensión en Internet calculada en número de páxinas web mediante a técnica explicada, baseada na busca en Google dun conxunto de expresións correspondentes a quince conceptos. A interpretación dos datos desta táboa sería como segue: por cada 100 páxinas en inglés hai 19 páxinas en español, 10 en francés, 9 en alemán e 6 en portugués. Para termos unha páxina en éuscaro, precísanse máis de 500 páxinas en inglés, case 1.000 para termos unha páxina en galego e 50.000 para unha en asturiano. Veremos, a continuación, que existe unha certa correlación entre estes resultados e outros datos lingüísticos relacionados obtidos doutras fontes, como o número de falantes por lingua, a cantidade de artigos por lingua na Wikipedia e a porcentaxe de sitios web que fornecen páxinas en cada idioma.

3.1. Comparación con outros datos lingüísticos

Na Táboa 3 preséntase unha comparación entre os resultados por lingua obtidos no noso experimento e os datos sobre o seu respectivo número de falantes tirados da

16 Consulta realizada os días 11.01.2021 e 12.01.2021 mediante a API JSON de busca de Google usando o parámetro *exactTerms*.

décimo sétima edición do atlas lingüístico *Ethnologue* (Lewis & Simons & Fennig, 2013), incluíndo para cada lingua, na segunda columna, a súa cantidade de falantes; na terceira, a súa posición relativa ao número de falantes e, na cuarta, a desviación da súa posición na web en relación á súa posición por falantes. A táboa débese interpretar do seguinte modo: o francés, con 68 millóns de falantes, ocupa o quinto lugar na clasificación das dez linguas polo número de falantes e ascende dous postos na clasificación das linguas polo número de páxinas web, ocupando a terceira posición nesta escala.

	Falantes	Posición por falantes	Desviación da posición na web
EN	334.800.758	2	+1
ES	405.638.110	1	–1
FR	68.458.600	5	+2
DE	83.812.810	4	0
PT	202.468.100	3	–2
IT	61.068.677	6	0
CA	7.220.420	7	0
EU	657.872	9	+1
GL	3.185.000	8	–1
AS	110.000	10	0

Táboa 3. Comparación con clasificación por falantes

Como se pode observar na Táboa 3, a maior desviación entre a clasificación por número de falantes e por cantidade de páxinas web dáse nas linguas francesa e portuguesa, que intercambian as súas posicións na web en favor do francés. Tamén supera na web o inglés ao español e o éuscaro ao galego, intercambiando as súas respectivas posicións, aínda que en xeral se observa unha certa correspondencia entre os dous rangos.

A Táboa 4 permite a comparación entre os resultados deste estudo e o número de artigos na Wikipedia por lingua tirado das estatísticas publicadas pola fundación Wikimedia[17]. A Táboa 4 inclúe, na segunda columna, o número de artigos na Wikipedia para todas as linguas consideradas; na terceira columna, a posición de cada lingua na clasificación por número de páxinas web obtida neste experimento e, na terceira columna, a diferenza entre a clasificación por páxinas web e a clasificación por número de artigos na Wikipedia.

17 Datos tirados de https://meta.wikimedia.org/wiki/List_of_Wikipedias o día 19.02.2021.

	Artigos na Wikipedia	Posición na web	Desviación da posición na Wikipedia
EN	6.252.687,00	1	0
ES	2.538.920,00	4	+2
FR	2.300.822,00	3	0
DE	1.675.227,00	6	+2
PT	1.661.807,00	2	-3
IT	1.056.859,00	5	–1
CA	671.703,00	7	0
EU	370.696,00	8	0
GL	171.382,00	9	0
AS	108.130,00	10	0

Táboa 4. Comparación con clasificación por artigos na Wikipedia

Na Táboa 4 obsérvase que a posición do galego constatada no experimento con respecto á súa extensión na web coincide coa situación que ocupa o galego na Wikipedia en relación co número de artigos redactados na nosa lingua. Comparando as posicións relativas das linguas nas clasificacións por artigos na Wikipedia e por páxinas web, pódese concluír que, en xeral, existe unha correlación bastante directa entre as dúas medicións, isto é, case todas as linguas (seis de dez) ocupan a mesma posición nas dúas clasificacións. Dúas linguas (alemán e italiano) están algo mellor na Wikipedia que na web, o portugués está un pouco mellor na web que na Wikipedia, e o español está bastante mellor na web que na Wikipedia.

	Porcentaxe sitos web	Posición por sitios web	Desviación cos resultados
EN	60,5	1	0
ES	3,8	2	0
FR	2,7	3	0
DE	2,3	4	0
PT	0,9	5	0
IT	0,8	6	0
CA	0,04	7	0
EU	0,006	8	0
GL	0,003	9	0
AS	< 0,0002	10	0

Táboa 5. Comparación con clasificación por sitios web

Unha última constatación indirecta da consistencia dos datos obtidos neste estudo provén da comparación destes resultados cos datos fornecidos por W3Techs sobre a porcentaxe de sitios web que serven páxinas en cada idioma[18], recollidos na Táboa 5. Nesta comparación, os resultados deste estudo e os que presenta a consultora W3Techs sobre a implantación das linguas nos sitios web son absolutamente parellos. No caso da lingua asturiana, a consultora non dá a porcentaxe específica de sitios web neste idioma por ser inferior a 0,0002.

A alta correlación entre os datos deste traballo e os relativos ao número de falantes, á cantidade de artigos na Wikipedia e á porcentaxe de sitios web que fornecen datos en cada lingua apuntan á consistencia dos resultados obtidos e permiten, ao mesmo tempo, interpretar con coherencia o sentido das desviacións observadas.

3.2. Evolución dos datos observados

Como diciamos anteriormente, o método empregado non permite determinar un valor absoluto en número de páxinas nin para toda a web nin para as linguas consideradas, senón unicamente unha porcentaxe da presenza de cada lingua en relación á do inglés. Porén, se comparamos estes datos cos obtidos en 2003 co buscador AlltheWeb (Gómez Guinovart, 2003), recollidos na Táboa 6, imos poder observar a evolución das linguas na súa implantación na web nos últimos vinte anos. A Táboa 6 inclúe, na segunda columna, a cantidade de páxinas estimada para cada lingua segundo os datos fornecidos en 2003 por AlltheWeb; na terceira columna, a porcentaxe de páxinas web acadada en 2003 por cada lingua en relación ao número de páxinas web en lingua inglesa, de novo conforme os datos de AlltheWeb; na cuarta columna, a porcentaxe de páxinas web obtida por cada lingua en relación ás da lingua inglesa consonte os datos fornecidos actualmente por Google; na quinta columna, a diferenza porcentual entre os dous datos anteriores para cada lingua, isto é, entre a porcentaxe de páxinas web de cada lingua con respecto ás do inglés en 2003 e en 2021 calculadas, respectivamente, a partir dos datos de AlltheWeb e Google; e, na sexta columna, o factor de crecemento na web de cada lingua, calculado como a relación entre as porcentaxes estimadas para cada lingua en 2021 e en 2003.

18 Datos tirados de https://w3techs.com/technologies/overview/content_language o 01.03.2021.

	2003 (núm. páxinas)	2003 (%)	2021 (%)	Diferenza	Crecemento
EN	442.670.131	100,000	100,00	n. a.	n. a.
ES	16.429.978	3,712	18,774	15,062	5,058
FR	24.631.376	5,564	10,027	4,463	1,802
DE	51.281.028	11,584	9,340	-2,244	0,806
PT	12.520.979	2,829	5,951	3,123	2,104
IT	15.137.909	3,420	4,521	1,101	1,322
CA	681.768	0,154	0,397	0,243	2,575
EU	80.271	0,018	0,173	0,154	9,519
GL	98.998	0,022	0,125	0,103	5,602
AS	sen datos	sen datos	0,002	sen datos	sen datos

Táboa 6. Evolución dos datos 2003-2021

Para explicar a interpretación desta táboa comparativa, tomarei o exemplo do francés: en 2003, a extensión en número de páxinas web do francés representaba o 5,564% da extensión na web do inglés, mentres que en 2021 representa o 10,027%. Porcentualmente, a diferenza é de 4,463 puntos, o que supón que a extensión do francés do 2003 ao 2021 aumentou de tamaño 1,802 veces, isto é, practicamente duplicouse con respecto á do inglés.

Considerando este último parámetro, o factor de crecemento na web de cada lingua, o galego é a segunda lingua que máis medrou en relación á súa extensión inicial nos últimos vinte anos con respecto ao inglés, cun índice de crecemento de 5,602 só superado polo do éuscaro (9,519) e seguido polo do español (5,058). A única lingua que manifesta unha certa perda de representación en relación ao inglés é o alemán (o seu índice de 0,806, inferior a 1, indica decrecemento), probabelmente por partir dunha implantación precoz na web moi superior a das outras linguas en 2003, con excepción do inglés.

4. Consideracións finais

Analizando as estimacións obtidas neste estudo sobre a extensión na web da nosa lingua nos últimos vinte anos, podemos observar tanto indicios da súa estabilidade, e mesmo dun certo crecemento, como indicios do seu estancamento en relación á ocupación do lugar na web que lle corresponde por número de falantes.

Por unha banda, consonte os datos observábeis, a presenza do galego na web segue a ser menor da que lle corresponde por número de falantes. Neste senso, a comparación cos datos do éuscaro é reveladora: cun número de falantes cinco veces menor, supera claramente os índices do galego no número de páxinas web, na cantidade de artigos publicados na Wikipedia e no número de servidores web que fornecen páxinas no idioma. Resulta importante observar que este *sorpasso* tivo lugar nos últimos vinte anos, pois o número de páxinas web en éuscaro documentado en 2003 en AlltheWeb resultaba lixeiramente inferior ao do galego. Naquela altura, AlltheWeb contiña no seu índice da arañeira un 0'014% das páxinas en galego (98.998 páxinas) e só un 0'011% (80.271) en éuscaro. As causas deste crecemento desigual na extensión na web do éuscaro e do galego nos tempos recentes son complexas, con certeza, e a economía non é un factor alleo a esta desigualdade, mais tamén parece evidente que o desenvolvemento na web do éuscaro ten os seus alicerces nunha política lingüística nacional non agresiva contra a lingua propia e nunha militancia lingüística salientábel.

Por outra banda, pódese constatar unha certa estabilidade no espazo ocupado polo galego na web, mesmo con tendencia á súa extensión neste ámbito. Sen dúbida, é esperanzador para o futuro da lingua observar que, de acordo coas estimacións máis recentes aquí documentadas, o galego é a segunda lingua que máis medrou na web a respecto do inglés nas últimas dúas décadas, cun índice de crecemento de 5,602 superior ao do español (5,058) e só superado polo do éuscaro (9,519). Con todo, o uso do galego na web segue necesitado dunha planificación lingüística pública axeitada que lle permita ampliar dun xeito sostido o seu espazo comunicativo nesta plataforma e non depender unicamente de iniciativas particulares e militantes. Estas iniciativas son necesarias e merecedoras de recoñecemento, aínda máis necesarias en tanto non se deixe de lexislar en contra do galego, mais resultan insuficientes para garantir unha presenza sostíbel da nosa lingua no ámbito da web. Cómpre seguir pulando en prol do idioma, tamén na web, tamén en Internet. O que está en xogo é a preservación dixital da nosa lingua e, por tanto, da nosa cultura e da nosa identidade como pobo.

Hai vinte anos remataba o meu artigo (Gómez Guinovart, 2003) cunha cita de Daniel Pimienta que, considerando os indicadores actuais, segue gozando de plena vixencia para a os intereses da nosa lingua: "o importante é transformarse de espectadores en actores, e participar no destino do noso futuro cultural e lingüístico cos medios que a tecnoloxía pon ao noso alcance. Se non entramos nós no xogo, outros decidirán o noso futuro" (Pimienta, 1999)[19]. A chamada á militancia lingüística é o último recurso, sen dúbida, mais é o único que nos queda logo de tantos anos de políticas lingüicidas das administracións autonómica e estatal.

19 A tradución do francés ao galego da cita é responsabilidade do autor deste traballo.

Como sinala Freixeiro Mato (2014), na súa análise do proceso de substitución lingüística en que está inmerso o galego, "a súa salvación ou morte vai depender da política lingüística que na Galiza desenvolveren os poderes públicos e en último extremo da vontade da comunidade de falantes". En consecuencia, como falantes de galego, ocupemos a rede, galeguicemos a arañeira e non permitamos a extinción dixital da nosa lingua.

Referencias bibliográficas

Canabal, Silvia, & Gago, Manuel, & López, Xosé, & Isasi, Antonio, & Limia, Moisés, & Pereira, José (2006). *A blogosfera en galego: demografía, usos e contidos*. Santiago de Compostela: Grupo de Novos Medios do Departamento de Ciencias da Comunicación da Universidade de Santiago de Compostela.

Caulfield, John (2013). *A social network analysis of Irish language use in social media.* Cardiff: Cardiff University. Dispoñíbel en http://orca.cf.ac.uk/53228/ (consultado en 29.03.2021).

Cunliffe, Daniel, & Ap Dyfrig, Rhodri (2013). "The Welsh Language on YouTube: Initial Observations". En Jones, Elin Haf Gruffydd, & Uribe-Jongbloed, Enrique (eds.), *Social Media and Minority Languages: Convergence and the Creative Industries*, 130-145. Bristol: Multilingual Matters.

Dopazo, Lara, & Lombao, David (2006). "A normalización lingüística do galego através dos blogs: o caso de Blogaliza", *Prisma.com, 3*, 214-229. Dispoñíbel en https://ojs.letras.up.pt/index.php/prismacom/article/view/2119/1952 (consultado en 29.03.2021).

Freixeiro Mato, Xosé Ramón (2014). "Lingua oral, calidade da lingua e futuro do galego". En Sánchez Rei, Xosé Manuel (ed.), *Modelos de lingua e compromiso*, 13-84. A Coruña: Baía.

Gerrand, Peter (2007). "Estimating Linguistic Diversity on the Internet: A Taxonomy to Avoid Pitfalls and Paradoxes", *Journal of Computer-Mediated Communication, 12/4*, 1298-1321. Dispoñíbel en https://doi.org/10.1111/j.1083-6101.2007.00374.x (consultado en 29.03.2021).

Gómez Guinovart, Xavier (2003). "A lingua galega en Internet". En Bringas López, Ana, & Martín Lucas, Belén (eds.), *Nacionalismo e globalización: lingua, cultura e identidade*, 71-88. Vigo: Servicio de Publicacións da Universidade de Vigo. Dispoñíbel en http://hdl.handle.net/11093/1930 (consultado en 29.03.2021).

Grefenstette, Gregory, & Nioche, Julien (2000). "Estimation of English and non-English Language Use on the WWW". En *Proceedings of the 6th International Conference RIAO*, 237-246. Dispoñíbel en https://arxiv.org/abs/cs/0006032v1 (consultado en 29.03.2021).

Honeycutt, Courtenay, & Cunliffe, Daniel (2010). "The Use of the Welsh Language on Facebook: An Initial Investigation", *Information, Communication and Society, 13/2*, 226–248. Dispoñíbel en https://doi.org/10.1080/13691180902914628 (consultado en 29.03.2021).

Lewis, M. Paul, & Simons, Gary F., & Fennig, Charles D. (eds.) (2013) [1951]. *Ethnologue: Languages of the World*. Dallas: SIL International.

Mas i Hernàndez, Jordi (2005). *La salut del català a Internet el 2005*. Dispoñíbel en https://www.softcatala.org/noticies/salut-catala-internet-2005/ (consultado en 29.03.2021).

Pimienta, Daniel (1999). *Y-a-t-il un espace dans l'Internet pour les langues et les cultures non-américaines? Le cas du monde latin*. Dispoñíbel en http://www.funredes.org/funredes/html/francais/publications/infoethics98.html (consultado en 29.03.2021).

Pimienta, Daniel (2005). "Linguistic Diversity in Cyberspace: Models for Development and Measurement". En Paolillo, John & Pimienta, Daniel, & Prado, Daniel (eds.), *Measuring Linguistic Diversity on the Internet*, 13-34. Paris: UNESCO. Dispoñíbel en https://unesdoc.unesco.org/ark:/48223/pf0000142186 (consultado en 29.03.2021).

Pimienta, Daniel, & Prado, Daniel, & Blanco, Álvaro (2009). *Twelve years of measuring linguistic diversity in the Internet: balance and perspectives*. Paris: UNESCO. Dispoñíbel en https://unesdoc.unesco.org/ark:/48223/pf0000187016 (consultado en 29.03.2021).

Pimienta, Daniel, & Prado, Daniel (2015). "Exploring the Status of Languages of France on the Internet: Methods and Reflection of Possible Approaches for Other Groups of Languages". En *Proceedings of the 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*, 139-171. Dispoñíbel en http://www.ifapcom.ru/files/2015/khanty/yak_mling_2015.pdf (consultado en 29.03.2021).

Van Den Bosch, Antal, & Bogers, Toine, & De Kunder, Maurice (2016). "Estimating search engine index size variability: a 9-year longitudinal study", *Scientometrics, 107/2*, 839-856. Dispoñíbel en https://doi.org/10.1007/s11192-016-1863-z (consultado en 29.03.2021).

Van Dijk, Ziko (2009). "Wikipedia and lesser-resourced languages", *Language Problems and Language Planning*, *33/3*, 234 - 250. Dispoñíbel en https://doi.org/10.1075/lplp.33.3.03van (consultado en 29.03.2021).