Recurso: Indicadores sobre a Presença da Línguas na Internet

Daniel Pimienta
Observatório da Diversidade Linguística e Cultural na Internet
http://funredes.org/lc
Link do recurso: http://funredes.org/lc2022

Abstrato

Indicadores confiáveis e mantidos do espaço das línguas na Internet são necessários para apoiar políticas públicas adequadas e estudos linguísticos bem informados. As fontes atuais são escassas e muitas vezes fortemente aviesados. O modelo de produção de indicadores sobre a presença de línguas na Internet, lançado pelo Observatório em 2017, atingiu um nível sensível de maturidade e seus produtos de dados são compartilhados na licença CC-BY-SA 4.0. Atinge agora 329 línguas (falantes L1 > um milhão) e todos os vieses associados ao modelo foram controlados a um limite aceitável, dando confiança aos dados, dentro de um intervalo de confiança estimado de +-20%. Alguns dos indicadores (principalmente a porcentagem de falantes L1+L2 conectados à Internet por língua e derivados) contam com o Ethnologue Global Dataset #24 para dados demolinguísticos e ITU, preenchido pelo Banco Mundial, para o percentual de pessoas conectadas à Internet por país. O restante dos indicadores se baseia nas fontes anteriores, além de uma grande combinação de centenas de fontes diferentes para dados relacionados ao conteúdo da Web por língua. Este pôster de pesquisa tem como foco a descrição dos novos recursos linguísticos criados. As considerações metodológicas são expostas apenas brevemente e serão desenvolvidas em outro artigo.

Palavras-chave: Recurso Linguístico, Línguas, Internet, Indicadores, Multilinguismo

1. Introdução

O Observatório da Diversidade Linguística e Cultural na Internet[1] vem trabalhando com métodos alternativos para medir indicadores da presença de línguas na Internet desde 1996. O método padrão para calcular a porcentagem de conteúdo da Web por língua é logicamente aplicar um algoritmo de reconhecimento de língua a todas as páginas da Web existentes e contar. A enorme extensão da Web torna essa abordagem impraticável, exceto para direcionar subconjuntos menores, como foi feito eficientemente pelo Language Observatory Project, antes que o projeto desaparecesse (Mikami, 2005). As tentativas de usar essa abordagem aplicando-a a um alvo com um número limitado de páginas da Web, supostamente representando fielmente o todo, são propensas a enormes vieses, como mostrado para o método definido por Alis Technologies em 1997[2]e reutilizado em 1999 (Lavoie, 1999) e 2003 (O'Neil, 2003) pela OCLC. Oito mil sites foram selecionados aleatoriamente por números de IP e as conclusões foram derivadas de uma medição única, em vez de uma série repetitiva tratada estatisticamente como uma variável aleatória.

Desde 2011, a W3Techs[3], de fato um excelente e confiável provedor de estatísticas para tecnologias da Web, fornece resultados atualizados diariamente para conteúdo da Web por língua, aplicando um algoritmo de reconhecimento de língua à página inicial dos 10 milhões de sites classificados como os mais visitados pelo Alexa.com[4]. O método é análogo ao utilizado para as outras 25 tecnologias Web pesquisadas por esta empresa, apresentando resultados extremamente interessantes. No entanto, as línguas são um tipo de tecnologia da Web bem diferente das bibliotecas Java Script ou servidores da Web e processar as línguas do conteúdo da Web da mesma maneira pode levar a erros enormes. A questão começa por focar apenas as *home pages* da seleção de sites: se você planeja computar conteúdo da web você precisa focar páginas da web para evitar dar o mesmo peso a um site de dez páginas comparado a um site de dez mil páginas. Além disso, as páginas iniciais de sites que não são em inglês geralmente incluem palavras em inglês (seja por uma vontade de apresentar o site em inglês, seja porque poucas palavras em inglês, como *copyright*, *abstract* ou botões de navegação, estão presentes). Esta é uma causa de erro para o algoritmo. A maior parte do erro está em outro lugar de qualquer maneira: é causado pela **falta de consideração ao multilinguismo** que faz com que o algoritmo conte como sites em inglês, sites que oferecem décimos de opção de língua em suas interfaces. Muitas vezes o site configura a opção de língua automaticamente, de acordo com a preferência do usuário, prática cada vez mais comum, principalmente para os principais sites do mercado global (Facebook.com é apenas um exemplo) e o algoritmo conta uma língua por página inicial, Inglês nesses casos. Não é à toa que, desde 2011, a porcentagem de inglês na Web é mantida estável e até crescente por W3Techs, apesar do fato de que há evidências de que a Internet mudou drasticamente na última década, com o chinês se tornando a primeira língua em termos de usuários, e a maioria das línguas asiáticas e árabe estão

[1] http://funredes.org/lc
[2] https://web.archive.org/web/20010730164601/http://alis.isoc.org/palmares.en.html
[3] http://W3Techs.com
[4] A coleta de tráfego da Web e sites de análise pertencentes à corporação Amazon, prestes a ser retirado do mercado.

crescendo. A Web é hoje provavelmente mais multilíngue que a humanidade. De acordo com os últimos dados do Ethnologue, a proporção de falantes L1+L2 sobre falantes L1 é 10 361 716 756 / 7 231 699 136 = 1,43. Ninguém deve se surpreender, então, que mais de 50% dos sites exibem páginas em mais de uma língua única. Não prestando a devida atenção ao multilinguismo ser um viés inaceitável para tais estudos, o W3Techs poderia, sem alterar sua atual seleção de sites e programa principal, corrigir seus vieses com alguns retrabalhos, como:

- Analise as opções de língua oferecidas na página inicial e conte cada opção, bem como a versão em inglês.
- Encontre um método para obter uma estimativa aproximada do número de páginas do site e multiplique cada versão linguística por esse número para contar as páginas da Web em vez dos sites.
- Quando o algoritmo informa mais de uma língua na página inicial, por precaução, não conte o site como inglês, mas sim a segunda língua.

Os novos resultados serão drasticamente diferentes…

O problema preocupante é que, devido à singularidade da fonte, à qualidade comprovada do restante de suas pesquisas, seu histórico de longo prazo e marketing eficiente, uma grande porcentagem da comunidade de pesquisa linguística (e formuladores de políticas públicas) está tomando os dados de W3Techs como entradas confiáveis. Infelizmente, boas teorias alimentadas por números errados dificilmente podem fornecer resultados corretos.

O exemplo mais sintomático da situação é dado pelo agregador de estatísticas Statista[5]que intitula seu anúncio de 2022 sobre línguas na Internet[6]com uma afirmação que soa como um fato concreto: *o inglês é a língua universal da Internet*, suportado pelos dados da W3techs, onde os conteúdos da web em inglês representam 63,7% do total, enquanto o chinês apenas 1,3%.

Ao mesmo tempo, o Observatório da Diversidade Linguística e Cultural na Internet calcula o inglês e o chinês na mesma porcentagem juntos, cerca de 20%, enquanto o hindi, com seus 224 milhões de internautas, chega a 3,8% (contra os 0,1% medidos pelo W3Techs).) e conclui seu último anúncio com essa frase: *A transição da Internet entre o domínio das línguas europeias, inglês na liderança, para as línguas asiáticas e árabe, chinês na liderança, está bem avançada e o vencedor é o multilinguismo, mas as línguas africanas demoram a ocupar seu lugar.*

Uma, pelo menos, das duas fontes deve estar extremamente errada e os pesquisadores devem ter cautela e verificar os vieses de um método antes de tirar conclusões de seus dados produzidos…

2. Os métodos alternativos

Em 1998-2007, o método alternativo do Observatório, que forneceu séries coerentes durante uma década, limitava-se ao inglês, alemão e às 5 línguas latinas (francês, italiano, espanhol, português e romeno). Ele usou motores de busca para contar um vocabulário comparável[7] para cada língua (Pimienta, 2009). A partir de 2007, a "evolução do marketing" dos motores de busca tornou o método obsoleto à medida que os seus relatórios de ocorrência se tornaram pouco fiáveis.

Hoje, são computadas 329 línguas, aquelas com falantes de L1 acima de um milhão, seguindo o Ethnologue, limitação adotada para evitar vieses muito fortes da hipótese de trabalho da abordagem*: todos os falantes da língua no mesmo país são computados com a mesma porcentagem de pessoas conectadas à Internet, o valor nacional fornecido pela UIT/Banco Mundial*. Esta hipótese proíbe a comparação de línguas dentro de um país, é pouco aplicável a línguas com baixo número de falantes e tende a influenciar positivamente as línguas de imigração em países em desenvolvimento (que podem ser menos conectadas do que a média) e a influenciar negativamente as línguas europeias em países em desenvolvimento (que tendem a ser mais bem conectadas do que a média).

O método atual é uma aproximação indireta dos conteúdos, baseada na observação experimental de que a razão entre a porcentagem mundial de conteúdos e a porcentagem mundial de falantes conectados sempre se manteve entre 0,5 e 1,5 (para línguas com existência digital plena).

Sugere-se algum tipo de lei econômica natural, que vincularia, para cada língua, a **oferta** (conteúdos e aplicativos

---

[55] http://statista.com Nesse sentido, não perderei a oportunidade de questionar a ética de dois fenômenos emergentes que podem ser correlacionados. 1) Muitos pesquisadores preguiçosos citam o Statista como fonte de dados em vez da própria fonte. 2) A Statista oferece alguns dados em acesso gratuito, mas a identificação da fonte desses dados só é acessível a clientes pagos. Vamos simplificar então e citar o Google como a mãe de todas as fontes ou, mais simples ainda, citar a Internet como a matriz de todas as fontes! 😊

[6] https://www.statista.com/chart/26884/languages-on-the-internet/

[7]Um conjunto de palavras para cada idioma, selecionadas com muitos cuidados linguísticos, cujas ocorrências eram relatadas pelos motores de busca e permitiam, por contagem, os resultados.

da web) à **demanda** (falantes conectados à Internet). Quando o número de pessoas conectadas aumenta, o número de páginas da web aumenta logicamente em conjunto, mais ou menos na mesma proporção. Isso acontece porque governos, empresas, instituições educativas, etc., e algumas pessoas criam conteúdos para responder a essa demanda.

Além disso, pesquisas e estudos têm relatado consistentemente que os usuários médios da Internet preferem usar sua língua materna e também aproveitam a oportunidade para usar, como segunda opção, sua(s) segunda(s) língua(s)[8].

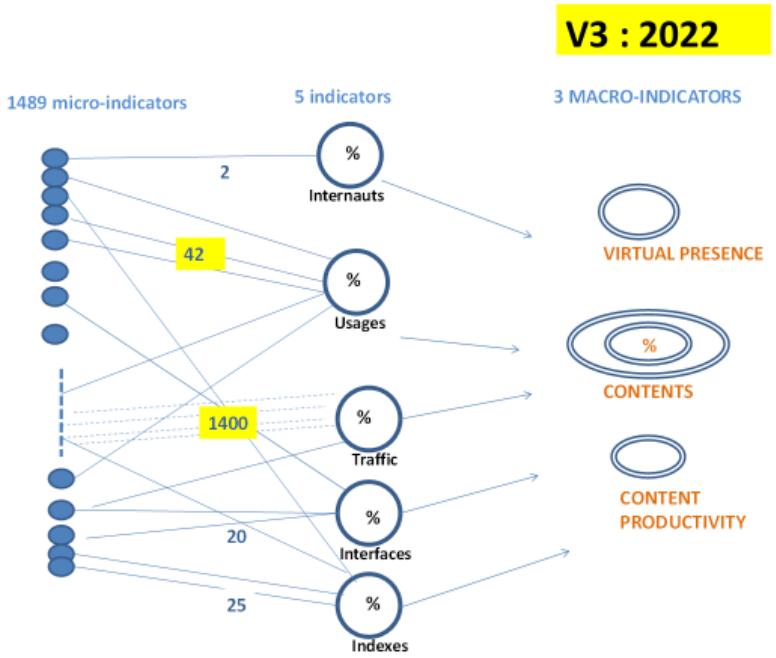
Assim, dependendo de cada língua, há algum tipo de modulação da referida razão, para torná-la acima ou abaixo de um. Isso significaria que algumas línguas têm mais produção de conteúdo do que outros, dependendo de um conjunto de fatores relacionados às línguas em seu contexto de país, como:

- Obviamente, a quantidade relativa de **falantes de L2**, pois algumas pessoas produzem, por exemplo, por razões econômicas, conteúdos em língua diferente da sua língua materna,

Mas também:

- A proporção do **tráfego** da Internet dependendo da tarifa do país, contexto cultural ou educacional.
- O número de **assinaturas** de redes sociais e outros aplicativos da Internet.
- **O suporte tecnológico digital** da língua e sua presença em interfaces de aplicativos e programas de tradução que facilitariam ou não a produção.
- O nível de submersão do país onde o falante vive em termos de instalações da **Sociedade da Informação** (comércio eletrônico, aplicativos governamentais para pagamento de impostos e assim por diante).

Então, se fosse possível coletar vários indicadores sobre cada uma das características mencionadas, seria possível aproximar a flutuação da modulação dos conteúdos da web em torno de um e deduzir de alguma forma a proporção dos conteúdos. Este é o núcleo do método e está sintetizado no diagrama a seguir que mostra todos os indicadores que são processados para cada língua e a quantidade correspondente de fontes que o modelo está usando. A primeira e a segunda versão da metodologia estão totalmente documentadas, veja para um lead (Pimienta, 2019). A descrição detalhada da versão 3 está a caminho.



Figura

1: Diagrama para criação de indicadores

[8] Veja, por exemplo, o relatório de pesquisa da União Europeia em https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556ou, para o caso desafiador da Índia, este relatório: https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

Este diagrama evoluiu, da versão 1 para a versão 3, ao longo da árdua tarefa de perseguir vieses, em termos de número de fontes e também em termos de indicadores. A computação do modelo estabelecido bastante complexo depende amplamente de uma variedade de operações de ponderação para realizar a tarefa, com, na maioria das vezes, o vetor de porcentagem de pessoas conectadas por país, que é o núcleo matemático do processo. A fonte de indicadores por língua disponível é escassa; a maioria dos indicadores é obtida por país e, como a maioria cobre apenas um subconjunto de países. A fonte de dados é extrapolada para todos os países, ponderando com os dados principais, e a transformação dos dados por país em dados por língua é obtida ponderando-os com os dados demolinguísticos (quantidade de falantes de cada língua em cada país).

3. Indicadores produzidos pelo modelo

Para cada uma das 329 línguas processados, o modelo está produzindo os seguintes indicadores por língua (observe que todas as porcentagens mundiais são baseadas em números L1+L2 e representam a parcela correspondente para cada língua).

Indicadores intermediários:
Internautas: falantes conectados à Internet
Usos
Tráfego
Interfaces e programas de tradução: em termos de porcentagem mundial dos números correspondentes de aplicativos e programas de tradução suportados
Índices: em termos de porcentagem mundial da classificação dos países nos parâmetros da Sociedade da Informação

Saídas do modelo (também chamado macroindicadores):
Falantes conectados: porcentagem do total de falantes L1+L2 do mundo daqueles conectados à Internet
Conteúdo: percentual de conteúdo da Web (calculado como a média dos 5 indicadores intermediários)
Produtividade de conteúdo: relação conteúdo/internautas
Coeficiente de presença virtual: relação conteúdo/quota mundial de palestrantes

Indicadores mais avançados
Ciber-geografia das línguas: uma repartição de saídas do modelo resumidas por famílias de línguas (europeu, asiático, árabe, americano, africano)
Indicador de Globalização Cibernética

$$CGI(L) = (L1 + L2)/L1(L) \times S(L) \times C(L)$$

Onde:

L1+L2/L1(L) é a razão do multilinguismo da língua L
S(L) é a porcentagem de países do mundo que possuem falantes da língua L
C(L) é a % de falantes da língua L conectados à Internet.

Este é um indicador das vantagens estratégicas de uma língua no ciberespaço.

Além disso, para algumas línguas, foi exibida a lista de países que detêm as maiores porcentagens de falantes conectados.

Os arquivos Excel com os resultados finais podem ser baixados em: http://funredes.org/lc2022

Está em projeto uma base de dados de acesso aos resultados, com possibilidade de consulta por nome de língua ou código iso.

4. Exemplos de indicadores produzidos

A seguir são apresentados alguns exemplos de dados, limitados aos melhores resultados, para a maioria dos casos. Os mesmos dados estão disponíveis para qualquer das 329 línguas processados.

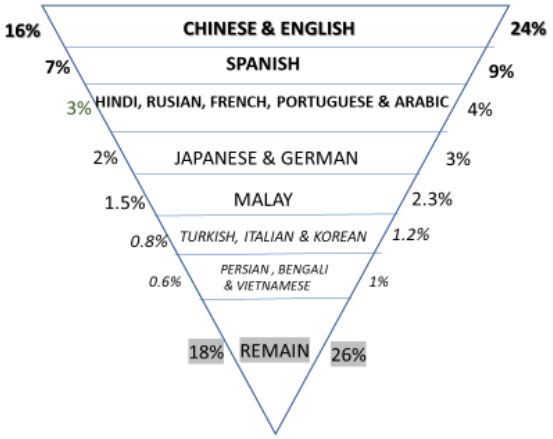


Figura 2: Porcentagem de janelas de conteúdo para as principais línguas

ANGUAGEM	CONECTADO CAIXAS DE SOM
norueguês	**96,89%**
dinamarquês	**96,42%**
sueco	**93,94%**
catalão	**92,88%**
japonês	**92,63%**
finlandês	**92,07%**
alemão. suíço	**91,55%**
limburguês	**91,42%**
flamengo ocidental	**91,30%**
holandês	**91,14%**
galego	**91,07%**
saxão superior	**89,81%**
estônia	**89,26%**
alemão padrão	**89,17%**
letão	**89,04%**
bávaro	**88,24%**

Tabla 1: Principais línguas em falantes conectados

A pirâmide invertida deve ser lida como uma expressão do intervalo de confiança: a porcentagem de conteúdo da Web em chinês está entre 16% e 24%, as demais línguas juntas representam entre 18% e 26% do total.

				População	População		Presença	Produtividade
			Internautas	Mundial	conectada	CONTEÚDO	Virtuais	Conteúdo
	ISO	LÍNGUAS	L1+L2	L1+L2	L1+L2	L1+L2	L1+L2	L1+L2
1	zho	*chinês*	18,46%	14,72%	71,38%	21,60%	1,47	1,17
2	eng	inglês	14,83%	13,01%	64,86%	19,60%	1,51	1,32
3	spa	espanhol	6,79%	5,24%	73,72%	7,85%	1,50	1,16
4	hin	hindi	4,19%	5,80%	41,16%	3,76%	0,65	0,90
5	rus	russo	3,51%	2,49%	80,32%	3,76%	1,51	1,07
6	fra	francês	2,98%	2,58%	65,80%	3,33%	1,29	1,12
7	por	português	2,99%	2,49%	68,43%	3,13%	1,26	1,05
8	ara	*árabe*	3,97%	3,53%	63,99%	3,09%	0,87	0,78
9	jpn	japonês	1,99%	1,22%	92,63%	2,66%	2,18	1,34
10	deu	alemão, padrão	2,04%	1,30%	89,17%	2,37%	1,82	1,16
11	msa	*malaio*	2,36%	2,36%	56,93%	1,96%	0,83	0,83
12	tur	turco	1,17%	0,85%	78,05%	1,14%	1,35	0,98
13	ita	italiano	0,87%	0,66%	75,83%	1,00%	1,53	1,14
14	kor	coreano	0,90%	0,79%	65,16%	0,98%	1,24	1,09
15	fas	*persa*	1,08%	0,81%	75,91%	0,88%	1,09	0,82
16	ben	bengali	1,11%	2,58%	24,55%	0,88%	0,34	0,79
17	vie	vietnamita	0,92%	0,74%	70,96%	0,85%	1,15	0,92
18	urd	urdu	0,95%	2,22%	24,38%	0,66%	0,30	0,70
19	tha	tailandês	0,80%	0,59%	77,95%	0,65%	1,12	0,82
20	pol	polonês	0,60%	0,39%	87,09%	0,63%	1,59	1,04
21	mar	marathi	0,69%	0,96%	41,06%	0,58%	0,60	0,83
22	tel	télugu	0,68%	0,92%	41,69%	0,56%	0,60	0,82
23	tam	tâmil	0,61%	0,82%	42,15%	0,51%	0,62	0,83
24	jav	javanês	0,62%	0,66%	53,76%	0,44%	0,66	0,70
25	nld	holandês	0,38%	0,24%	91,14%	0,41%	1,73	1,08
26	guj	gujarati	0,44%	0,60%	41,47%	0,36%	0,61	0,83
27	ukr	ucraniano	0,40%	0,32%	71,02%	0,35%	1,09	0,88
28	kan	kannada	0,41%	0,57%	41,11%	0,33%	0,59	0,82
29	ron	romena	0,32%	0,23%	79,57%	0,30%	1,29	0,93
30	aze	*azerbaijão*	0,33%	0,23%	81,54%	0,28%	1,21	0,85
		PERMANECER	22,60%	30,10%		15,13%		
		TOTAL	100,00%	100,00%		100,00%		

Tabla 2: Principais indicadores para as 30 principais línguas em porcentagem de conteúdo

Deve ser lido assim: o inglês representa 13% da população mundial L1+L2 e 14,8% da população conectada à Internet; 64,7% dos falantes de inglês L1+L2 estão conectados à Internet; 19,6% dos conteúdos da Web estão em inglês; o coeficiente de presença virtual do inglês é 1,5, o que significa que os conteúdos em inglês estão sobre representados em um fator superior a 50%; a produtividade de conteúdo do inglês é de 1,32, maior depois do japonês.

As línguas macro são indicadas em cursiva.

LÍNGUA	PRESENÇA VIRTUAL
Japonês	**2.18**
Norueguês	**1,88**
Alemão, padrão	**1,82**
sueco	**1,82**
dinamarquês	**1,78**
holandês	**1,73**
finlandês	**1,69**
catalão	**1,68**
alemão, suíço	**1,63**
polonês	**1,59**
italiano	**1,53**
estônia	**1,51**
russo	**1,51**
inglês	**1,51**
hebraico	**1,50**
grego	**1,50**
espanhol	**1,50**
chinês	**1,47**
letão	**1,46**
galego	**1,46**

Tabla1: Principais línguas na presença virtual

LINGUA	PRODUTIVIDADE CONTEÚDO
japonês	**1,34**
inglês	**1,32**
chinês	**1,17**
alemão, padrão	**1,16**
espanhol	**1,16**
italiano	**1,14**
francês	**1.12**
norueguês	**1.10**
sueco	**1.10**
coreano	**1,09**
holandês	**1,08**
russo	**1,07**
grego	**1,07**
cabo-verdiano	**1,05**
dinamarquês	**1,05**
português	**1,05**
finlandês	**1,04**
polonês	**1,04**
catalão	**1,03**
alemão, suíço	**1,02**
hebraico	**1,00**

Tabla 2: Principais línguas na presença virtual

LÍNGUAS DE (*)	ÁFRICA	AMÉRICAS	MUNDO ÁRABE	ÁSIA	EUROPA	PACÍFICO (**)
% de internautas	29,8%	56,7%	64,0%	49,3%	82,6%	
% de conteúdo	2,89%	0,22%	3,09%	44,77%	45,39%	
Presença Virtual	0,28	0,68	0,87	0,65	1,39	
Produtividade Conteúdos	0,51	0,68	0,78	0,72	0,95	
População L1+L2 %	9,15%	0,31%	3,53%	48,21%	30,91%	
População Conectada %	5,18%	0,32%	3,89%	44,60%	39,51%	
NÚMERO DE LÍNGUAS	138	8	1	135	47	0

Tabla 3: Ciber-geografia das línguas

(*) Deve ser entendido como línguas nativos. Por exemplo, as 8 línguas indígenas das Américas com mais de um milhão de falantes L1 incluídas no modelo são: aimará, guarani, crioulo haitiano, hunsrik, crioulo jamaicano, q'eqchi', kiche e quíchua.
(**) Nenhuma língua do Pacífico está incluída, pois nenhum possui mais de 1 milhão de falantes.

LÍNGUA	CGI	CGI%
inglês	1,61	14,24%
francês	1,09	9,66%
alemão	0,42	3,75%
russo	0,31	2,76%
espanhol	0,27	2,40%
árabe	0,18	1,56%
malaio	0,17	1,51%
italiano	0,17	1,50%
chinês	0,16	1,46%
português	0,15	1,37%
tailandês	0,15	1,37%
romani	0,15	1,35%
turco	0,15	1,34%
grego	0,15	1,31%
ucraniano	0,15	1,31%
polonês	0,13	1,15%
persa	0,12	1,10%
romeno	0,12	1,06%
hindi	0,12	1,04%

Tabla 4: Indicador de Globalização Cibernética

A segunda coluna é calculada dividindo o valor CGI pelo total de CGIs para as línguas processadas. Ele é mencionado como forma de medir, por exemplo, o peso relativo das duas primeiras posições, próximo a 25% do total.

CHINÊS	L1+L2	%CONN.	CONECTADOS	% DE CONN.
TOTAL	**1 525 335 340**	**71,38%**	**1 088 735 519**	**100%**
China	**1 448 870 000**	**70,64%**	**1 023 512 815**	**94,01%**
China–Taiwan	**37 320 000**	**88,82%**	**33 148 541**	**3,04%**
China–Hong Kong	**10 942 800**	**92,41%**	**10 112 585**	**0,93%**
Malásia	**7 838 700**	**89,56%**	**7 019 949**	**0,64%**
Cingapura	**4.026.000**	**75,88%**	**3 054 766**	**0,28%**
Estados Unidos	**2 894 390**	**88,50%**	**2 561 503**	**0,24%**
Vietnã	**2.500.000**	**70,64%**	**1 766 054**	**0,16%**
Indonésia	**2.054.000**	**53,73%**	**1 103 542**	**0,10%**
Tailândia	**1.729.000**	**77,84%**	**1 345 918**	**0,12%**
Canadá	**1 212 600**	**97,00%**	**1 176 222**	**0,11%**
Filipinas	**1 010 280**	**43,03%**	**434 689**	**0,04%**
DESCANSO	**4 937 570**	**71,04%**	**3 507 738**	**0,32%**

Tabla 5: Repartição de falantes de chinês conectados pelos principais países

HINDI	L1+L2	%CONN.	CONECTADOS	% DE CONN.
TOTAL	**600 800 970**	**41,15%**	**247 258 401**	**100%**
Índia	**596 000 000**	**41,00%**	**244 360 000**	**98,87%**
Kuwait	**700 000**	**98,60%**	**690 200**	**0,28%**
Estados Unidos	**643 000**	**88,50%**	**569 048**	**0,23%**
Nepal	**1 307 600**	**25,00%**	**326 900**	**0,13%**
África do Sul	**463.000**	**68,00%**	**314 840**	**0,13%**
Arábia Saudita	**171.000**	**97,86%**	**167 345**	**0,07%**
Austrália	**160.000**	**86,54%**	**138 472**	**0,06%**
Canadá	**111.000**	**97,00%**	**107 670**	**0,04%**
Iémen	**316 000**	**30,00%**	**94 800**	**0,04%**
DESCANSO	**929 370**	**52,63%**	**489 127**	**0,20%**

Tabla 6: Repartição de falantes Hindi conectados pelos principais países

5. Referências Bibliográficas

Ethnologue Global Dataset (2022). https://www.ethnologue.com/product/ethnologue-global-dataset-0

Lavoie B.F., O'Neill E. T. (1999). How "World Wide" is the Web? *Annual review of OCLC Research*, https://web.archive.org/web/20031006155123/http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=00000003496

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

O'Neill E.T., Lavoie B.F., Bennett R. (2003). Trend in the Evolution of the Public Web: 1998 – 2002. *D-Lib Magazine*, 9.4 http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

OIF (2022). Le français dans le monde, Gallimard, ISBN : 9782072976865. Synthèse en ligne: https://francophonie.org/sites/default/files/2022-03/Synthèse_La_langue_française_dans_le_monde_2022.pdf

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

Pimienta D. (2019). Indicators of Languages in the Internet, in Proceedings of International Conference Language Technologies for All (LT4All), 4-6 December 2019, UNESCO, Paris; PP 315-319 https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.79.pdf

6. Agradecimentos

Os estudos da versão 3 foram financiados pela Organisation Internationale de la Francophonie e os resultados alimentaram o Capítulo Internet da (OIF, 2022).

A ideia de usar várias fontes de dados por país e transformá-los em dados por língua foi concebida por Daniel Prado em 2012.