O método por trás da produção sem precedentes de indicadores de presença linguística na Internet

Daniel Pimienta (*), Álvaro Blanco (*), Gilvan Müller de Oliveira (**)

(*) Observatório da diversidade linguística e cultural na Internet
https://obdilci.org/

(**) Cátedra UNESCO em Políticas Linguísticas para o Multilinguismo

Tradução do artigo publicado em inglês: *The method behind the unprecedented production of indicators of the presence of languages in the Internet.* Frontiers Research Metrics & Analytics - Section Research Methods, Volume 8 – 2023. doi: 10.3389/frma.2023.1149347.

RESUMO

São necessários indicadores fiáveis e atualizados da presença de línguas na Internet para gerir eficazmente as políticas linguísticas, para prever o mercado do comércio eletrónico ou para apoiar novas pesquisas no domínio do suporte linguístico digital. Este artigo apresenta uma descrição abrangente dos elementos metodológicos envolvidos na produção de um conjunto inédito de indicadores da presença na Internet de 329 línguas com mais de um milhão de falantes de L1. É dada particular ênfase ao tratamento de todos os vieses envolvidos no processo, provenientes quer do método, quer das diferentes fontes utilizadas no processo de modelação. Também são discutidos vieses relacionados com outras fontes que fornecem dados semelhantes e, em particular, é mostrado como a falta de consideração do elevado nível de multilinguismo da Web leva a uma enorme sobrestimação da presença da Inglês na Web. A lista detalhada das fontes é apresentada nos diversos apêndices. Pela primeira vez na história da Internet, a produção de indicadores sobre a presença virtual de um grande conjunto de línguas poderá permitir avanços nos domínios da economia das línguas, da cibergeografia das línguas e das políticas linguísticas para o multilinguismo.

PALAVRAS-CHAVE: Linguas, Web, Internet, Indicadores, Metodologia, Viés, Webmetrics

OBRIGADO

O trabalho que conduziu à versão 3 (e à versão 1) do modelo descrito foi realizado graças ao financiamento da Organisation de la Francophonie. A base do método baseia-se em parte nas ideias apresentadas por Daniel Prado em 2012, em particular na ideia, aplicada à Web, de utilizar estatísticas por país, cruzadas com dados demolinguísticos, para obter dados por língua. Agradecimentos ao governo brasileiro e ao Instituto Internacional da Língua Portuguesa que viabilizaram a versão 2 do modelo, um passo essencial para a versão 3.

Conteúdo

TABELAS E FIGURAS

1. INTRODUÇÃO

A medição do espaço de representação das línguas na Internet ainda não entusiasma as multidões, mas as questões, a nível linguístico, cultural, socioeconómico e geopolítico, estão longe de serem neutras.

Quanto à situação das línguas no mundo, entre as cerca de 7.000 línguas ainda existentes, cerca de 40% estão em perigo[1] e a intensidade da sua presença na Internet pode ser um indicador preditivo significativo. Para definir políticas públicas eficazes para as línguas, medir a situação atual e a sua evolução é um pré-requisito, nomeadamente no que diz respeito à capacidade de avaliar o impacto dessas políticas.

Nos estágios iniciais da Internet, alguns pesquisadores investigaram um novo campo chamado cibergeografia, que é o estudo da natureza espacial das redes de comunicações de computadores.[2] A aquisição de indicadores da presença de um maior número de línguas na Internet permite-nos propor o conceito de cibergeografia das línguas como uma noção relacionada (Pimienta, Oliveira, 2022).

Embora a Internet não seja um território homogéneo do ponto de vista do seu funcionamento e governação (O'Hara, Hall, 2018), podemos tratá-lo como um ciberterritório reticular multilíngue, analisando a distribuição e a interação entre as línguas em um espaço geral. Numa segunda perspectiva, porém, cada língua é um território que orienta o adensamento das relações, nomeadamente políticas e económicas. Cada território linguístico é ao mesmo tempo um mercado, com capacidades específicas de produção e consumo.

Esta visão territorial permite incluir na discussão outro conceito relevante: a geopolítica das línguas e o multilinguismo. A geopolítica é composta principalmente por três fatores: território, que implica localização; a população, neste caso, os falantes conectados de cada língua; e o efeito de alavanca, que aqui é o equipamento digital de cada língua, a sua massa de conteúdos e as suas políticas de promoção, ou seja, a sua capacidade de receber investimentos (Flint, 2021). Deste ponto de vista, os ciberterritórios linguísticos são mercados em disputa política e económica (Bauböck, 2015).

Vários economistas analisaram o valor económico das línguas sob diferentes perspectivas (Grin, Vaillancourt, 1997; Gazzola, 2015). Mas apesar dos instrumentos disponíveis mostrarem que as línguas são fundamentais para todas as categorias da economia de serviços descritas pela OMC[3], responsável por uma parte crescente do PIB dos países com capitalismo avançado, os governos e os investidores têm sido lentos a desenvolver perspetivas mais contemporâneas sobre a gestão linguística.

[1]De acordo com Etnólogo (https://www.ethnologue.com) o número exato de línguas vivas é 7.168, enquanto outras fontes calculam que existiram cerca de 30.000 línguas (https://www.uh.edu/engines/epi2723.htm).
[2]https://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/about.html
[3]A Organização Mundial do Comércio (OMC) oferece quatro modos de troca de serviços:(a) do território de um Membro para o território de qualquer outro Membro (Modo 1 - Comércio Transfronteiriço); b) no território de um Membro ao consumidor de serviços de qualquer outro Membro (Modo 2 - Consumo no estrangeiro); (c) por prestador de serviços de um Membro, através de presença comercial, no território de qualquer outro Membro (Modo 3 - Presença Comercial); e 4) por prestador de serviços de um Membro, através da presença de pessoas físicas de um Membro no território de qualquer outro Membro (Modo 4 - Presença de pessoas físicas).

Em 2020, só o comércio eletrónico representou 20% do total das vendas globais no retalho[4], e as plataformas devem comunicar na língua dos seus clientes para manter a sua competitividade no mercado (ver várias fontes no Anexo 8). Quem conseguir penetrar em diferentes mercados linguísticos aumentará os seus lucros, o que leva as grandes empresas a investirem em estratégias multilingues (Oliveira, 2010). Está em curso um processo de mercantilização linguística e os dados sobre a presença de línguas na Internet são essenciais para a tomada de decisões nesta área (Heller, 2010).

Desde 2011, os decisores políticos e os investigadores linguísticos têm tido que confiar exclusivamente em duas fontes disponíveis, ambas da área do marketing empresarial, para avaliar o impacto das suas políticas ou apoiar as suas teorias .

- ✓ W3Techs oferece porcentagens de conteúdo web por língua[5], para os 40 principais línguas, com atualização diária, e também mantém histórico de porcentagem[6].

- ✓ InternetWorldStats relata porcentagens de falantes conectados à Internet para os 10 principais línguas[7], com atualização anual.

A análise do método W3Techs revela vieses graves que resultam da não consideração do multilinguismo significativo que prevalece na Web (ver 4.2 vieses da W3Tech). Os cálculos do InternetWorldStats baseiam-se na combinação da percentagem de pessoas online por país, um valor fiável que é publicado anualmente e disponibilizado pela União Internacional de Telecomunicações (UIT)[8], o órgão das Nações Unidas que publica estatísticas de telecomunicações, e dados demolinguísticos para falantes de L1 (primeira língua) e L2 (segunda língua) por país. As fontes existentes sobre dados demolinguísticos relatam grandes diferenças, particularmente ao nível dos dígitos L2; entre eles, o Ethnologue é geralmente considerado a fonte mais confiável; no entanto, esta fonte é proprietária e não gratuita[9].

Desde março de 2022, o Observatório da Diversidade Linguística e Cultural na Internet (doravante Observatório) oferece estes dois indicadores, bem como indicadores adicionais significativos, para as 329 línguas que compõem uma população de falantes L1 superior a 1 milhão (ver resultados em Pimienta 2022), com planos de atualizações anuais[10]. Este é o culminar de um longo processo de purificação dos vieses de um método definido em 2017[11]e que, em última análise, fornece resultados com um limite de confiabilidade aceitável.

O Observatório não é um novato neste campo: realizou uma série de medições pioneiras de conteúdos web em inglês, alemão e línguas latinas (francês, italiano, português, espanhol e romeno), entre 1997 e 2007 (Pimienta, Prado e Blanco 2009). O método aproveitou o número total de ocorrências de palavras ou frases em páginas web, que foi reportado por motores de busca que rastreiam uma percentagem significativa do espaço web. O Observatório foi forçado a desistir depois de 2007, quando os motores de busca deixaram de fornecer números fiáveis e a proporção de páginas web indexadas foi significativamente reduzida.

---

[4] https://www.digitalcommerce360.com/article/global-ecommerce-sales/
[5] https://w3techs.com/technologies/overview/content_language
[6] https://w3techs.com/technologies/history_overview/content_language/ms/y
[7] https://www.internetworldstats.com/stats7.htm
[8] https://itu.int
[9] https://www.ethnologue.com/data-consulting
[10] https://obdilci.org/lc2022
[11] O método é descrito em https://obdilci.org/lc2017/Alernative%20Languages%20Internet.docx.

O novo método, desenvolvido em 2017, que permitiu desenhar um conjunto de indicadores para as 139 línguas com mais de 5 milhões de falantes de L1, inaugurou uma nova abordagem, definida em 2012 e aplicada a línguas únicas, principalmente francês (Pimienta , 2014) e espanhol (Pimienta, Prado, 2016). Esta abordagem centrou-se em gerir um conjunto, o maior possível, de fontes de dados dispersas sobre línguas ou países, tendo algum tipo de relação com a Internet. Esta relação pode ser direta (por exemplo, distribuição por país de assinantes de uma rede social específica ou línguas suportados em serviços de tradução online) ou indireta (por exemplo, classificação na área de comércio eletrônico ou número médio de celulares por pessoa em cada país)[12]. A escassez de dados relativos às línguas utilizadas na Internet foi compensada pela utilização de números relativos aos países mais numerosos, e estes foram transformados em números por língua por ponderação com os dados demolinguísticos. Os dados coletados foram organizados em diferentes categorias: conteúdo, tráfego, usos, índices[13]e interfaces[14]. Em 2017, ao proporcionar consistência matemática e utilizar técnicas estatísticas para extrapolar dados faltantes, o método foi generalizado para vários línguas, além do francês ou espanhol. Um modelo foi projetado para transformar todas as fontes em indicadores significativos para os 139 línguas com mais de 5 milhões de falantes de L1.

Posteriormente e desde 2017, o trabalho tem sido dedicado principalmente ao combate aos vários vieses específicos do método ou das fontes de dados. Em 2021, isto resultou na versão 2, com a mesma estrutura, mas com certos enviesamentos significativos controlados, em particular, pela utilização da base de dados Ethnologue Global 24 (março de 2021) para dados demolinguísticos. Posteriormente, a cobertura linguística foi expandida para 329 línguas com mais de um milhão de falantes L1. A luta contínua contra o viés prosseguiu e conduziu, em março de 2022, a uma redefinição definitiva da abordagem e da confiança na obtenção de um nível razoável de controlo do viés, com a capacidade de produzir números fiáveis, dentro de um intervalo de confiança de mais ou menos 20%, uma estimativa empírica que não é apoiada por nenhum cálculo estatístico.

Por que é tão importante identificar os vieses e, sempre que possível, tentar mitigá-los ou, se não for possível, avaliar o seu impacto nos resultados obtidos a partir desses vieses intransponíveis? Em qualquer actividade de investigação, o método científico exige uma utilização cuidadosa dos dados e das estatísticas porque podem surgir vieses e, se forem pesados, podem desacreditar totalmente os resultados obtidos. Embora seja uma abordagem conhecida na saúde, onde são realizados um grande número de estudos estatísticos, seja para avaliar o efeito de um tratamento, seja para medir a prevalência de uma determinada doença em uma população específica, a amostragem deve ser cuidadosamente selecionada e o método deve basear-se em bases sólidas (por exemplo, com procedimentos duplo-cegos, em que nem os participantes nem os investigadores sabem que tratamento o participante recebeu). Esta preocupação com o viés deve aplicar-se igualmente a todas as áreas de investigação.

O campo da medição da língua na Internet situa-se na intersecção de duas áreas onde os vieses são notáveis: a demolinguística (demografia linguística) e a Web. Em ambas as áreas, não existe um forte consenso sobre os dados e podem surgir grandes diferenças, dependendo das fontes, em números como o número de falantes de uma determinada língua residentes num determinado país ou o número total de páginas web.

[12]https://en.wikipedia.org/wiki/List_of_Wikipedias
[13]O índice refere-se a classificações em diferentes parâmetros associados ao progresso na sociedade da informação.
[14]A presença de idiomas como opção de interface em uma lista de aplicações incluindo tradução online, como aproximação de uma métrica até então inexistente para o nível de suporte tecnológico de idiomas.

O viés pode ocorrer de diversas formas, específicas das fontes de dados utilizadas, inerentes ao método utilizado, à seleção feita para amostragem, à hipótese de cálculo ou à hipótese que sustenta certas simplificações necessárias. Se é responsabilidade primária do produtor de dados abordar sistematicamente possíveis vieses e documentar aqueles que permanecem, também é responsabilidade do pesquisador que utiliza esses dados identificar as fontes e verificar sua credibilidade, encontrar a descrição do método e analisar possíveis vieses, tudo antes de tirar conclusões com base nesses dados. O raciocínio correto sobre dados falsos dificilmente produzirá conclusões confiáveis! A facilidade oferecida hoje pelos motores de busca para identificar fontes de dados públicos na Web não elimina a necessidade de verificação dessas fontes!

O método teórico padrão para medir o espaço linguístico na Web é navegar em todas as páginas Web na Internet, aplicar um algoritmo de reconhecimento de língua a cada uma delas e contar o(s) língua(s) de cada página, prestando atenção para que uma única página poderia conter mais de um língua. Por fim, dividindo o número obtido para cada língua pelo número total de páginas rastreadas, obtém-se o percentual. Antes deste processo, obviamente devem ser analisados possíveis vieses do algoritmo de reconhecimento de língua.

De acordo com a Netcraft[15], hoje existem mais de 1,2 bilhão de sites, dos quais 200 milhões estão ativos. Uma fonte[16]estima o número total de páginas web em cerca de 50 mil milhões, das quais menos de 10% são indexadas por motores de busca. Neste contexto, direcionar websites em vez de páginas web é uma simplificação utilizada pela maioria dos estudos, o que implica novos riscos de enviesamento a ter em conta, ainda mais se o reconhecimento da língua se aplicar exclusivamente na página inicial de cada site, que muitas vezes contém elementos em inglês, mesmo para sites que não falam inglês. Porém, explorar todo o universo de sites existentes é uma opção que mesmo os motores de busca poderosos não conseguem dar conta; é então necessária uma maior simplificação, na prática para selecionar uma amostragem reduzida que, esperançosamente, seja representativa de toda a web. Este é outro risco de viés que uma rápida verificação do histórico de tentativas irá destacar.

Antes de 2007, o número de iniciativas para tentar medir a percentagem de presença linguística na Web era limitado; abaixo é realizada uma exploração rápida, para uma análise mais aprofundada consultar (Pimienta, Prado, Blanco 2009).

Três das primeiras tentativas, no período 1995-1999, utilizaram a abordagem padrão (Team Babel, uma iniciativa conjunta da Alis Technologies e da Internet Society[17]) e outros dois. (Grefenstette, Noche, 2000) e o Observatório[18]usaram abordagens diferentes. (Grefenstette, Noche, 2000) utilizou uma técnica para estimar o tamanho de um corpus específico de um língua a partir da frequência de palavras comuns nesse corpus e aplicou-a à Web. O Observatório comparou o número de ocorrências de vocabulário equivalente nas diferentes línguas estudadas (dados fornecidos pelos motores de busca).

A equipe do Babel configurou sua amostragem da web para ser analisada por uma técnica de randomização de números IP que, em última análise, consistiu em pouco mais de 3.000 sites em cuja página inicial foi realizado o reconhecimento de língua. Houve muitas causas de viés,

[15] https://news.netcraft.com/archives/category/web-server-survey
[16] https://www.worldwidewebsize.com
[17] https://web.archive.org/web/20011201133152/http://alis.isoc.org/palmares.en.html
[18]https://obdilci.org/lc2005/francais/L1.html

mas o maior problema é que apenas uma amostragem e, portanto, apenas uma medição, foi realizada. Em termos estatísticos, a ausência de uma série de medições invalida os resultados porque uma única amostragem de 3.000 websites num universo, na altura, de um milhão, está completamente fora do âmbito. A abordagem válida deveria ter sido replicar a operação, digamos pelo menos cem vezes, e calcular a média, a variância e outros atributos estatísticos da distribuição resultante. O facto é, no entanto, que esta abordagem padrão foi reutilizada duas vezes, (Lavoie, O'Neil, 1999) e (O'Neil et ale, 2003), e transmitiu aos meios de comunicação a ideia errada de que 80% do conteúdo da web estava em Inglês, sem alterações durante o período 1996-2003.

No mesmo período, o Observatório melhorou o seu método com a colaboração de linguistas de uma instituição parceira, com base em vocabulários equivalentes em diferentes línguas, evitando ao máximo potenciais vieses. O Observatório apresentou resultados que mostram que o inglês estava em declínio constante, de 80% do conteúdo da web em 1996 para 50% em 2007. Esta abordagem, embora limitada às línguas latinas, inglês e alemão, produziu uma série de medidas consistentes durante o período; no entanto, sua dependência de contagens confiáveis de acessos de mecanismos de pesquisa desencadeou seu desaparecimento em 2007.

Duas outras iniciativas ocorreram durante o período, ambas utilizando a abordagem padrão: o projeto 'Observatório da Língua' - LOP (Mikami et al, 2005) e um projeto do Instituto de Estatística da Catalunha - IDESCAT (Monras, 2006). O projeto LOP, um consórcio acadêmico com parceiros unindo forças nos dois principais requisitos, web crawling de alta capacidade e algoritmo moderno de reconhecimento de língua, apresentou todos os atributos para se tornar a melhor solução para abordar o tema, aliando o rigor dos pesquisadores e o poder de capacidade de exploração. Começou a se concentrar nas línguas dos países asiáticos menos populosos e se expandiu gradualmente. Foi estabelecida uma colaboração com o Observatório, sob a égide da Rede Global para a Diversidade Linguística - MAAYA[19], quando o LOP produziu dados para os países latino-americanos, mas infelizmente este projeto, coordenado pela Universidade de Nagaoka, terminou logo após o terremoto e o tsunami que atingiram o Japão em 2011.

Quanto ao projeto IDESCAT, que se concentrou especificamente na língua catalã, a sua vida foi curta. Este período de actividade académica em torno do tema foi seguido praticamente de um abandono desta área às empresas de marketing, com, como consequência, o reinado de metodologias não completamente transparentes e não revisadas pelos pares e, ao mesmo tempo, excelentes estratégias de marketing. permitindo grande impacto público.

Depois de 2017, além das iniciativas do Observatório, um consórcio de universidades gregas (Giannakoulopoulos et al., 2020) utilizou a abordagem padrão para avaliar a presença do inglês em sites sob domínios de primeiro nível de países da União Europeia (ccTLDs). A sua amostragem inclui pouco mais de 100.000 sites e o seu método prestou especial atenção ao multilinguismo dos sites, verificando sistematicamente o língua de todos os links internos da página inicial. A partir dos seus dados de produção, é possível calcular um valor de 28% de versões em inglês de websites para todos os sites da União Europeia (incluindo Reino Unido, Irlanda e Malta) ou 13% para países europeus que não falam inglês (Pimienta, 2023).

A W3Techs aplicou, até maio de 2022, o seu algoritmo de reconhecimento de língua, diariamente, a uma lista dos 20 milhões de websites mais visitados, fornecida pela Alexa.com, um serviço comercial de análise de tráfego web. A partir de maio de 2022, quando o serviço

[19] https://web.archive.org/web/20190904002849/http://maaya.org/?lang=fr

Alexa foi descontinuado, ele passou a constar da lista de milhões de sites mais visitados da Tranco[20]um serviço sem fins lucrativos, orientado para a investigação e que se apresenta como "robusto contra a manipulação".

O algoritmo W3Techs é aplicado na página inicial de cada um dos sites da lista Tranco e conta um único língua para cada um deles, ignorando seu potencial multilinguismo. A falta de alternativa há muito tempo, e também a merecida reputação da empresa pelo seu principal serviço, as pesquisas Web Technologies, tornaram esta fonte extremamente popular e muitas vezes uma referência, até mesmo para a comunidade de pesquisa. Ao contrário das outras 26 tecnologias web estudadas pela empresa, como JavaScript, línguagens de marcação ou data centers, as línguagens são tecnologias web um tanto especiais, com a propriedade de que várias dessas tecnologias podem ser associadas a uma página Web única ou a um site. único site. O multilinguismo é uma propriedade da Web que requer atenção especial para fornecer resultados imparciais. Esta propriedade está no centro do método apresentado a seguir.

2. O MÉTODO

2.1 VISÃO GERAL

Esta é uma aproximação indireta do conteúdo da Web por língua, baseada na observação experimental feita sistematicamente, desde o início do Observatório, de que a relação entre a percentagem global de conteúdo e a percentagem global de falantes conectados (definida como produtividade de conteúdo) sempre permaneceu dentro da janela [0,5 --- 2], para línguas com existência digital.

Esta observação sugere a existência de uma espécie de lei económica natural, que ligaria, para cada língua, a oferta (conteúdo web e aplicações numa determinada língua) à procura (falantes conectados desta língua) à Internet). Quando aumenta o número de pessoas conectadas, o número de páginas web aumenta naturalmente e ao mesmo tempo, mais ou menos na mesma proporção. Isto acontece porque governos, empresas, instituições de ensino, etc., e alguns indivíduos, estão a criar conteúdos e aplicações para satisfazer esta procura.

É importante notar, em apoio à afirmação anterior, que inquéritos e estudos sobre o comportamento dos utilizadores da Internet reportam sistematicamente que estes preferem utilizar a sua língua materna, quando estão disponíveis conteúdos, nomeadamente para o comércio electrónico, e, além disso, desejam utilizar a(s) sua(s) segunda(s) língua(s) (ver, no Anexo 8, uma seleção de fontes sobre este assunto).

Assim, dependendo de cada contexto linguístico, há uma espécie de modulação da referida relação, para torná-la, mais ou menos, maior ou menor que um. Algumas línguas apresentam melhor produtividade de conteúdo do que outras, dependendo de um conjunto de fatores específicos da língua ou ligados ao contexto dos diferentes países onde uma determinada proporção de falantes dessa língua se conecta à Internet. Os seguintes fatores foram identificados:

Específico da língua:

[20] https://tranco-list.eu

- Obviamente, o número de falantes de L2, porque algumas pessoas produzem, por exemplo por razões económicas, conteúdos numa língua diferente da sua língua materna.
- O suporte tecnológico da língua para o ciberespaço, estimado através de sua presença em interfaces de aplicativos e programas de tradução, o que facilitaria ou não a produção de conteúdo.

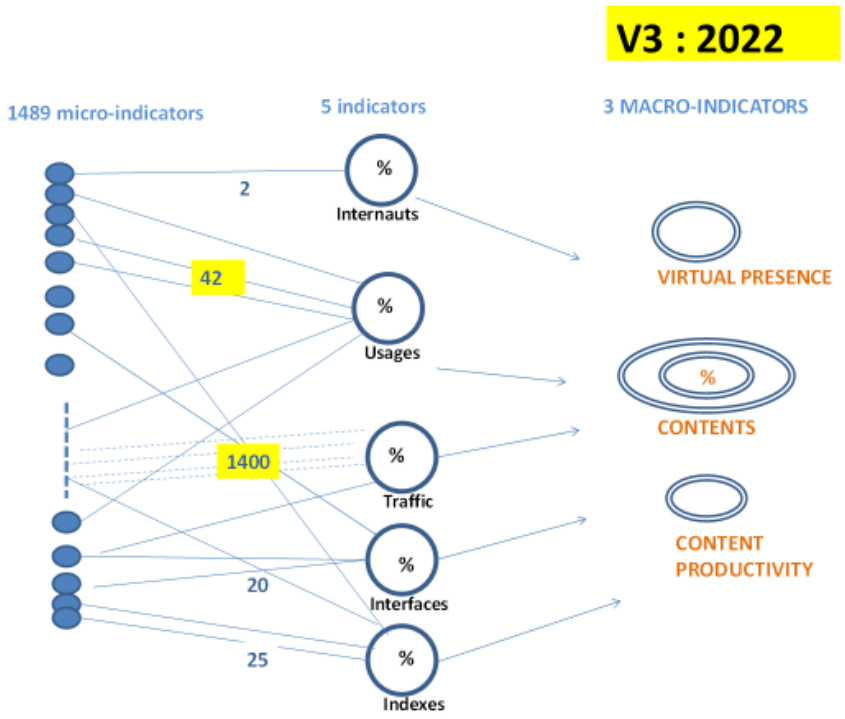
Mas também, dependendo de cada país onde existam falantes L1 ou L2 desta língua:

- A quantidade de tráfego da Internet, dependendo da taxa do país, do contexto cultural ou educacional.
- O número de assinaturas de redes sociais e outros aplicativos da Internet.
- O nível de progresso do país em termos de serviços da sociedade da informação (comércio electrónico, aplicações governamentais para pagamento de impostos, etc.).

Portanto, se fosse possível recolher dados significativos suficientes sobre cada um dos factores mencionados para criar indicadores correspondentes, aproximaríamos o valor do rácio de produtividade de conteúdos e, da proporção de falantes conectados, deduziríamos a proporção de conteúdos.

Este é o coração do método e está resumido no diagrama a seguir que mostra todos os indicadores que são processados para cada língua e a quantidade correspondente de fontes que o modelo utiliza.

Figura 1: Das fontes aos produtos



2.2 DESCRIÇÃO DAS ENTRADAS DO MODELO

As entradas do modelo são divididas em 5 tipos de fontes: usuários da Internet, usos, tráfego, interfaces e índices.

2.2.1 Internautas

Esta é a porcentagem de falantes L1+L2 conectados à Internet para cada língua. A transformação dos dados de origem, expressos por país, em dados necessários, expressos por língua, é realizada por ponderação:

CL(j) é a porcentagem de falantes conectados para o língua j.

$$CL(j) = \sum_{j=1}^{j=P} LP(i, j) \times PC(i) / \sum_{j=1}^{j=P} LP(i, j)$$

Ou :
P é o número total de países
LP (i, j) = O número de falantes L1+L2 da língua j no país i.
PC (i) = A porcentagem de pessoas conectadas para o país i

O produto da matriz CL = LP + . x PC, em notação APL[21], ou = SumProduct (LP;PC), em notação Excel, é uma operação de ponderação que produz a partir de um vetor do tamanho do número de países, um novo vetor, desta vez do tamanho do número de línguas.

A validade deste cálculo baseia-se no pressuposto implícito de que, dentro do mesmo país, todos os grupos linguísticos partilham a mesma percentagem de pessoas ligadas. Este é um dos vieses fundadores do método, discutido no capítulo Viés.

O vetor CL(j) é um elemento chave do modelo que será utilizado, novamente em operações de ponderação, com diferentes fontes, para calcular a modulação de cada indicador.

A fonte da matriz LP é Etnólogo; o modelo utiliza o Global Dataset #24 de março de 2021. As fontes da matriz PC são a União Internacional de Telecomunicações (UIT) e o Banco Mundial; a UIT, a fonte histórica destes dados[22], baseia-se em fontes governamentais e, quando estas não estão disponíveis, nas suas próprias estimativas. Como a UIT deixou de fornecer as suas próprias estimativas em 2017, a fonte é complementada por dados do Banco Mundial[23]que preenche esta lacuna em muitos casos. Quando não há dados recentes disponíveis, é realizada uma extrapolação de dados mais antigos.

2.2.2 Interfaces

Pesquisadores da rede MetaNet[24]fazem um bom trabalho na análise do suporte tecnológico para as línguas europeias, mas ainda não existe uma métrica para avaliar o suporte tecnológico para todas as línguas do mundo. Para aproximar este parâmetro, foi dada ênfase à presença de cada língua nas interfaces de um conjunto de aplicações populares da Internet e como um dos pares num conjunto de serviços de tradução online. Foram identificadas dezesseis fontes para as quais a lista de línguas suportados está acessível (ver apêndice 3).

[21]APL, "A Programming Language", que é ao mesmo tempo um formalismo matemático e sua implementação na forma de uma linguagem de programação, desenhada por Kenneth. Iverson. Para mais detalhes vejahttps://fr.Wikipedia.org/wiki/APL_(idioma).
[22]https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2021/December/PercentIndividualsUsingInternet.xlsx
[23]Fonte :https://data.worldbank.org/indicator/IT.NET.USER.ZS
[24]http://www.meta-net.eu

2.2.3 Índices

O tema aqui é avaliar os países com base no seu progresso de acordo com os critérios da sociedade da informação. A ponderação com dados demolinguísticos transformará esses dados numa classificação por língua. Na versão 1, foi utilizada uma lista de 4 fontes. A partir da versão 2 foi realizada uma busca sistemática e foram identificadas 27 fontes, tornando a seleção quase exaustiva (ver anexo 4).

2.2.4 Usos

Foram identificados cinco subindicadores e utilizadas as fontes correspondentes:

- Assinantes de mídia social: foram exploradas 36 fontes, cada uma ligada a redes sociais com mais de 100 milhões de assinantes. Para as principais redes sociais ocidentais, foram identificadas fontes sobre o número de assinantes por país; para as restantes redes sociais, principalmente da Ásia, foram obtidos dados parciais de tráfego por país, utilizando SimilarWeb[25], extrapolado para o resto dos países, proporcionalmente à percentagem de pessoas ligadas por país.
- Comércio eletrônico: foi usada apenas uma fonte que faz o trabalho perfeitamente. Este é o indicador T-index do Imminent Translated Research Center[26]. Este indicador classifica os países com base no seu potencial de vendas online, estimando assim a quota de mercado de cada país no comércio eletrónico global. O conjunto de percentagens por país é transformado pela ponderação dos falantes conectados por língua num conjunto de percentagens por língua[27].
- Transmissão de vídeo (streaming):O modelo utiliza apenas duas fontes neste momento: percentagem de assinantes da Netflix por país e penetração do YouTube por país.
- Conteúdo aberto:O modelo utiliza apenas uma fonte nesta fase: a percentagem por país da soma dos downloads do OpenOffice 2012/21.
- A infraestrutura:O modelo utiliza três dados principais do Banco Mundial que são fundidos em dois indicadores: % de assinantes de banda larga fixa por país e % de assinantes de telefonia fixa + móvel por país.

Os resultados finais foram ponderados, para refletir a confiança atual nos dados[28]e assim reduzir o viés, com os seguintes valores:

- Assinantes de mídia social: 0,3
- Comércio eletrônico: 0,3
- Transmissão de vídeo: 0,05
- Conteúdo aberto: 0,05
- Infraestrutura: 0,3

que são posteriormente transformados por ponderação em distribuição por língua.

A lista detalhada de fontes do indicador de usos pode ser encontrada no Apêndice 1.

---

[25]Um serviço de marketing que fornece uma proporção de tráfego por país para um grande conjunto de websites:https://www.similarweb.com/

[26]https://imminent.translated.com/t-index

[27]Observe que o Imminent também fornece as porcentagens completas por idioma, provavelmente fazendo algo semelhante. Existem pequenas diferenças entre o Imminent e os nossos cálculos, provavelmente devido a diferentes dados demolinguísticos. O modelo usa nossos cálculos em vez da fonte direta do Iminente porque o Iminente está limitado a 89 idiomas, enquanto a técnica de extrapolação permite alcançar todos os idiomas do estudo.

[28]Uma média simples não ponderada será utilizada na próxima versão, quando cada item tiver obtido as fontes necessárias para arquivá-lo.

2.2.5 Tráfego

Existem ferramentas (como o SimilarWeb, já mencionado) para obter uma estimativa da distribuição do tráfego por país para um site específico; Normalmente, essas ferramentas oferecem dados de sites classificados entre os milhões ou dez milhões de sites mais visitados. Os dois desafios são: 1) avaliar estas ferramentas e compreender o seu potencial enviesamento e 2) estabelecer uma seleção de locais com enviesamento mínimo, permanecendo dentro de um tamanho viável (digamos, cerca de 1000 locais). Muitas mudanças ocorreram da versão 1 para a versão 3 para superar vieses; eles são descritos em 2.4.5. A lista de sites utilizados para tráfego pode ser encontrada no Apêndice 5.

2.2.6 Conteúdo

O indicador de conteúdo foi um insumo para o modelo nas duas primeiras versões, pois a visão metodológica original era coletar o maior número possível de fontes. A fonte escolhida foi a Wikimedia, que coleta, para cada uma de suas aplicações[29], e para cada língua processado, estatísticas confiáveis e interessantes por língua, apesar de ser provavelmente a aplicação mais multilíngue da Web com suas versões em 327 línguas. A versão 3 fez com que este sinalizador fosse removido da lista de entradas. O capítulo 2.4.8 trata dos enviesamentos deste indicador e dá a justificação para esta decisão.

2.3 DESCRIÇÃO DAS SAÍDAS DO MODELO

O modelo fornece as seguintes saídas, para cada língua:

Falantes: a porcentagem global de falantes L1+L2

Falantes conectados: a porcentagem de falantes deste língua conectados à Internet

Usuários de internet: a porcentagem global de Falantes conectados

Conteúdo[30] : a porcentagem global de conteúdo

Presença virtual: a proporção entre conteúdo e Falantes.

O valor global (e médio) é 1: um valor maior que 1 significa uma presença virtual maior que a presença real e vice-versa.

*Produtividade de conteúdo:*o relatório de conteúdo sobre usuários da Internet

O valor global (e médio) é 1: um valor superior a 1 significa alta produtividade dos Falantes conectados.

Índice de ciberglobalização: ICM (l) = (L1 + L2) / L1(l) x P(l) x C(l)

Ou :

L1+L2/L1 (l) é a proporção de multilinguismo da língua l (obtida na fonte Ethnologue)

P(l) é a porcentagem de países no mundo que possuem falantes da língua l (fonte Ethnologue)

C(l) é a % de falantes da língua L conectados à Internet (calculado pelo modelo)

É um indicador das vantagens estratégicas de uma língua no ciberespaço[31].

[29]Wikipedia, Wikcionário, WikiBooks, WikiQuote, WikiVoyage, WikiSources, Wikimedia Commons, WikiSpecies, WikiNews, Wikiversity e WikiData.

[30]Nas duas primeiras versões, sendo o conteúdo um insumo, os indicadores de saída eram chamados de Potência, Capacidade e Gradiente, com exatamente a mesma definição de hoje Conteúdo, Presença Virtual e Produtividade de Conteúdo.

[31]Em termos percentuais, o inglês e o francês detêm juntos quase 25% do peso, seguidos, de longe, pelo alemão, russo, espanhol e árabe.

Além disso, ao agrupar os resultados por família linguística, a tabela anteriormente apresentada Cibergeografia das línguas foi produzida agrupando os indicadores por famílias linguísticas[32], produzindo uma perspectiva global interessante sobre a situação e as tendências.

2.4 ANÁLISE DE VIÉS

A tabela a seguir mostra a evolução dos vieses de V1 a V3 usando uma classificação subjetiva de 0 (vieses tão grandes que os dados não fazem sentido) a 20 (absolutamente nenhum viés), com 10 (vieses perceptíveis, mas aceitáveis) no meio.

Tablela 1: Avaliação de viés

AVALIAÇÃO DE VIÉS Marque de 20	V1 2017	V2 2021	V3 2022
MÉTODO BÁSICO	**17**	**17**	**17**
MÉTODO PARA L2	**13**	**19**	**19**
USUÁRIOS DE INTERNET	**19**	**16**	**19**
ÍNDICES	**15**	**18**	**18**
CONTEÚDO	**5**	**8**	
TRÁFEGO	**13**	**11**	**17**
INTERFACES	**19**	**19**	**19**
USOS	**12**	**12**	**16**

2.4.1 Método básico

O viés implícito no cerne do modelo é considerar que todas as línguas do mesmo país partilham a mesma taxa de conectividade à Internet (o valor nacional fornecido pela UIT). A realidade é obviamente diferente porque o conceito de exclusão digital também existe dentro de cada país.

Esta hipótese de trabalho causa um viés positivo para falantes de línguas não europeias que vivem em países desenvolvidos (que provavelmente estão menos conectados que a média) e reciprocamente um viés negativo para falantes de línguas europeias em países em desenvolvimento (que provavelmente estão mais conectados que a média). Sendo um fundamento do método, este pressuposto não pode ser alterado e as decisões tomadas para lidar com ele são:

- Não são possíveis comparações entre o desempenho de diferentes línguas dentro de um país.
- Como o risco de viés significativo aumenta inversamente com o tamanho da população de falantes, o estudo foi inicialmente limitado a línguas com mais de 5 milhões de falantes L1, depois estendido a línguas com mais de um milhão de falantes. Versões futuras podem tentar estender esse limite, mas provavelmente nunca abaixo de 100.000, pois os vieses podem se tornar inevitáveis.

2.4.2 Método para L2

Pela primeira vez, em 2021, o Ethnologue estendeu os seus dados demolinguísticos por país aos falantes de L2. Isso removeu um dos vieses mais significativos do método (em V1) que resultou na extrapolação de dados (por exemplo, porcentagem de falantes conectados) de L1 para L2, método que viesou resultados positivos para línguas com forte presença no desenvolvimento países, como inglês e francês. Na verdade, este processo atribuiu taxas de

[32]A definição de famílias linguísticas utilizada é a do Etnólogo.

ligação à Internet mais elevadas aos falantes de L2 nos países em desenvolvimento do que na realidade. A partir de V2, com a existência de dados demolinguísticos por país tanto para L2 como para L1, o modelo funciona diretamente a partir de populações L1+L2 e este viés de extrapolação desaparece; mas obviamente não o viés básico que se manifesta da mesma forma para L1, L2 e L1+L2.

Deve-se notar que as fontes demolinguísticas têm um viés maior para dados de L2 do que para dados de L1, porque não existe uma definição perfeita do nível de proficiência em um segundo língua necessário para ser contado como L2. Na verdade, as fontes de dados para L2 variam enormemente, especialmente para o inglês.[33].

2.4.3 Usuários da Internet
É, depois dos dados demolinguísticos, o segundo elemento principal do modelo e é importante garantir uma fonte fiável. Tal como mencionado no ponto 2.2.1, os dados da UIT e do Banco Mundial são combinados para obter os melhores e mais fiáveis dados atualizados.

2.4.4 Índices
Com a expansão das fontes em V2, atingindo a quase completude e seleção de instituições confiáveis (organizações internacionais e organizações não governamentais), o viés de seleção é mínimo e a confiança nos dados é máxima.

2.4.5 Tráfego
As ferramentas disponíveis para obter a distribuição do tráfego por país para um grande conjunto de sites (aqueles considerados os mais visitados) são: Alexa.com, SimilarWeb.com, Ahrefs.com e Semrush.com. Todos vêm de empresas de marketing, que não são totalmente transparentes quanto ao seu método. Por exemplo, Alexa, a mais antiga e famosa, embora tenha encerrado as operações em maio de 2022, produz a partir de um banner que os usuários podem baixar. Este banner, associado a um navegador da web, notifica Alexa sobre os sites visitados pelo usuário a partir desse navegador. Com a coleta de todos os dados enviados por todos os banners do mundo, Alexa constrói seus resultados, tanto em termos de ranking do site quanto de distribuição de tráfego por país. É óbvio que a distribuição geográfica dos banners pode ser uma indicação de prováveis vieses, mas infelizmente esta informação não é publicada.

O trabalho para superar os vieses deste indicador foi o mais demorado. Na versão 1 foi utilizado Alexa.com, com uma seleção de 450 sites. Foi estabelecido, comparando os dados de tráfego de países da Alexa com dados de assinantes de países, coletados de várias fontes, que a Alexa era positivamente tendenciosa em relação ao inglês e ao francês e fortemente tendenciosa em relação aos países asiáticos e ao Brasil. Para combater o inevitável viés de seleção, o processo do indicador não foi realizado por média simples mas sim por uma média reduzida com uns elevados 20%, tentando assim mitigar os vieses de seleção.

Os testes da versão 2 mostraram que Alexa parecia ter corrigido o viés asiático negativo, mas um novo viés parecia agora afetar os países europeus. Testes adicionais levaram à descoberta

[33]Os dados do Ethnologue para o inglês são de 1,348 bilhão de falantes L1 + L2 (L1 = 370 milhões, L2 = 978 milhões), enquanto outras fontes sugerem 1,18 bilhão (https://en.wikipedia.org/wiki/List_of_countries_by_English-Speaking_population) ou 1,5 bilhão (https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/fonte real não citada). Em 2008, David Crystal expressou a possibilidade de este número tender para 2 mil milhões (https://www.cambridge.org/core/journals/english-today/article/two-thousand-million/68BFD87E5C867F7C3C47FD0749C7D417).

de um erro em que o país líder em termos de tráfego por vezes não estava listado e esta pode ser a razão para o enviesamento observado nos resultados, uma vez que isto acontece nomeadamente com os países europeus. Optou-se então por utilizar Alexa apenas quando a soma dos percentuais oferecidos fosse superior a 70%, uma forma simples de eliminar esses casos errôneos. Ahrefs e Semrush foram tentados, mas rejeitados devido a uma forte tendência para o inglês e para um deles uma percentagem total por país muitas vezes superior a 100%. SimilarWeb forneceu resultados relativamente próximos do Alexa.com, após a correção mencionada e optou-se pela versão 3 utilizar ambas as ferramentas utilizando a metade da soma dos resultados.

Após numerosos testes e experimentos realizados, concluiu-se que o viés de seleção era definitivamente um problema sério que precisava ser resolvido de forma mais drástica do que com a média reduzida. A versão 3 abordou esta situação com uma nova abordagem que tornou possível gerir uma seleção de mais de 1000 websites onde o viés foi reduzido por todos os meios possíveis[34].

Para atingir o objetivo de seleção livre de vieses, decidiu-se finalmente estabelecer uma seleção dos websites mais visitados em cada país, com um número de websites proporcional ao tráfego global do país. Por razões práticas, o algoritmo não foi definido para atingir todos os países, mas foi limitado aos 55 países que detinham os primeiros lugares em termos de conteúdo para as línguas faladas nesses países.

Tablela 2: Lista de países processados para seleção de sites nacionais

Afeganistão	Argélia	Alemanha	Angola	ArábiaArábia
Argentina	Austrália	Bangladesh	Bélgica	Brasil
Bulgária	Camboja	China	Hong Kong	Taiwan
Colômbia	Coreia do Sul	Egito	Emirados Árabes Unidos	
Espanha	Estados Unidos	França	Índia	Indonésia
Irã	Iraque	Itália	Japão	Cazaquistão
Kuwait	Lituânia	Malásia	México	Marrocos
Moçambique	Nepal	Nigéria	Uzbequistão	Paquistão
Países Baixos	Filipinas	Polônia	Portugal	Romênia
Reino Unido	Rússia	Cingapura	Sudão	Sri Lanka
Tanzânia	Tailândia	Turquia	Ucrânia	Vietnã

A regra foi definida para selecionar, para cada país, pelo menos os três primeiros sites mais visitados,com entre eles, pelo menos, o primeiro domínio local (ccTld[35], como .fr para França). Esta regra foi definida para evitar que a seleção fique muito concentrada nos sites globais mais visitados (geralmente .com). Obviamente, isso não impediu que os sites mais globais (como Google.com ou facebook.com) aparecessem na seleção de muitos países e foi realizada uma ponderação para respeitar este fenómeno.

Para realizar a seleção foram utilizadas todas as quatro ferramentas (Ahrefs, Alexa, Semrush e SimilarWeb) embora em algumas ocasiões, devido à falta de dados em países com populações pequenas, tivemos que coletar os dados de outras fontes.

[34]Esta decisão obrigou a abandonar um resultado interessante, mas estatisticamente fraco, das versões anteriores que consistia em agrupar os sites por tema e tirar para determinadas línguas conclusões provisórias sobre a sua força ou fraqueza em relação a esses temas. A questão de saber se estes resultados reflectiam mais o viés de selecção do que certas realidades temáticas da presença linguística na Internet permaneceu por resolver.
[35]Domínio de nível superior com código de país.

Um total de 1.421 sites foram selecionados automaticamente[36], dos quais 733 eram sites diferentes. O número de ocorrências de cada site na seleção de 1.421 foi mantido para posterior ponderação. Para cada país, o número de websites correspondente à sua participação no tráfego global da Internet também foi calculado e retido para posterior ponderação para controlar o viés de seleção.

Este método garantiu a seleção menos tendenciosa possível para a medição do tráfego e permitiu superar o enorme viés face aos países asiáticos que marcou desde o início este indicador. Definitivamente melhorou os resultados finais para chinês, hindi e árabe, bem como para outras línguas asiáticas.

Pode permanecer um viés que penaliza os países (e línguas associadas) onde o nível geral de literacia digital é mais elevado e para os quais existe, portanto, um tráfego significativo para sites com conteúdo científico ou literário, e em qualquer caso excluindo redes sociais e mundialmente famosas. sites. Infelizmente, este é o preço a pagar para obter resultados livres de grandes vieses. É claro que este viés marginal não favorecerá as línguas dos países desenvolvidos, ou seja, na maioria das vezes as línguas europeias.

Uma possível melhoria para a versão 4 poderia ser a inclusão de um novo indicador que estabelecesse a proporção por país de sites de domínio nacional em comparação com sites de domínio genérico; este indicador poderia ser um primeiro passo para medir o grau global de literacia digital por país e poderia até ser utilizado através de uma nova ponderação para compensar o enviesamento residual em questão. Entretanto, os resultados brutos do modelo poderão prejudicar ligeiramente o francês e o inglês e, pelo contrário, parecem agora favorecer ligeiramente os chineses.

2.4.6 Interfaces

Para cada língua é estabelecido um ranking de acordo com o número de vezes que existe a presença desse língua na lista de aplicações selecionadas (interface ou tradução online). A partir desta classificação, a operação de ponderação com o percentual de Falantes conectados produz o percentual “modulado” esperado. Obviamente, este indicador é bastante "agressivo" porque centenas de línguas estão completamente ausentes da lista e, portanto, recebem um valor de 0%, o que significa que não há absolutamente nenhum suporte tecnológico. Esta dura medida é, em qualquer caso, um reflexo da dura realidade deste campo onde demasiadas línguas têm um nível de suporte tecnológico digital próximo de zero, apesar dos esforços crescentes dos investigadores de tecnologia linguística.[37].

4.7 Usos

O elemento subscritores de redes sociais esteve na origem de um forte viés pró-Ocidente nas versões 1 e 2, devido à ausência de redes sociais não-ocidentais, tendo sido feito um esforço especial na versão 3 para completar as 11 fontes iniciais[38]com aplicações semelhantes do resto do mundo.

O critério escolhido para completar foi manter redes sociais com mais de 100 milhões de assinantes. Quando as fontes de dados por país forem identificadas (geralmente distribuição de

[36]O processo de seleção foi feito por programação computacional para evitar erros ou vieses indesejados.
[37]Veja as intensas conferências e workshops semestrais da comunidade de pesquisa do LREC desde 1998:http://www.lrec-conf.org.
[38]Facebook, Linkedin, Twitter, Instagram, Reddit.

assinantes por país), elas serão utilizadas; caso contrário, a distribuição dos assinantes por país foi estabelecida a partir do tráfego por país, dados obtidos pelo serviço SimilarWeb, e alargada a todos os países por extrapolação (ver 3.4).

A distribuição por país, após extrapolação de cada elemento, é ponderada em função do número de assinantes e finalmente transformada numa percentagem por língua através da ponderação com a matriz demolinguística.

Na versão 3, a complementação reduziu significativamente o viés contra países não-ocidentais e indiretamente contra línguas não-europeias. A lista completa das redes sociais processadas pode ser consultada no Anexo 1.

Para o elemento de comércio eletrônico, conforme mencionado anteriormente, a fonte é única, mas totalmente confiável.

Para streaming de vídeo, o modelo utiliza apenas duas fontes neste momento: a percentagem de assinantes da Netflix por país. Este subindicador necessita claramente de ser expandido na próxima versão do modelo com aplicações de streaming alternativas além do YouTube e Netflix, com um esforço especial para países não ocidentais. Até então, o elemento recebe um peso baixo.

Para conteúdo aberto, este subindicador também deverá ser ampliado na próxima versão do modelo com mais dados relacionados à abertura, principalmente na área de MOOCs. O item recebe um peso baixo no momento.

No que diz respeito às infra-estruturas, os dados do Banco Mundial sobre linhas fixas, móveis e banda larga por país são fiáveis e fornecem uma base sólida para o indicador. A soma das linhas fixas e móveis num único dado equilibra a situação entre os países desenvolvidos com elevada penetração de linhas fixas e os países em desenvolvimento com elevada penetração móvel.

Resta que nesta fase o indicador de utilizações é o que tem recebido menos atenção e deve ser melhorado para a próxima versão, embora o objectivo principal que era superar o viés ocidental pudesse ser alcançado. As redes sociais não ocidentais foram integradas na componente de redes sociais e isso produziu os efeitos esperados nos resultados, revelando a presença crescente de países e línguas asiáticas.

2.4.8 Conteúdo

Este indicador é aquele, juntamente com o tráfego e os usos, que tem recebido maior atenção no trabalho contra vieses. É também aquele cujos vieses, herdados da galáxia Wikimedia, tiveram maior influência nos resultados das duas primeiras versões, conferindo uma vantagem notável, para indicadores independentes da população de falantes, aos resultados das línguas principalmente presentes. na Wikimedia.

Os dois principais desafios da Wikimedia são, em primeiro lugar, que apesar dos seus notáveis esforços e sucesso em ser verdadeiramente global, ela sofre de um viés ocidental e, em segundo lugar, que línguas específicas têm investido pesadamente na participação na enciclopédia online e nas presenças presentes. extremamente desproporcional à realidade do número de Falantes

conectados[39], enquanto outras línguas viram seus resultados nas primeiras versões impulsionados por sua forte presença nos serviços da Wikimedia[40]. Além disso, alguns línguas aumentaram artificialmente o número de artigos, traduzindo-os de versões de outros línguas, mantendo uma taxa de atualização extremamente baixa.

Na versão 2, uma fórmula foi definida e utilizada como indicador, em vez do número de artigos da Wikipédia, para remover efetivamente a mencionada vantagem artificial:

W (i) = Artigos (i) x Edições (i) x Editores (i) x Profundidade (i) / L1+L2 (i) ²

Ou :
Artigos (i) = o número de artigos da Wikipedia para o língua i
Edições (i) = o número de edições de artigos para o língua i
Editores(i) = o número de editores para artigos no língua i
Profundidade (i) = um indicador da frequência de atualizações de artigos para o língua i[41]
L1+L2(i) = o número de falantes de primeira e segunda língua de i.

Todos os elementos da fórmula são fornecidos nas estatísticas da Wikipedia; para obter detalhes, consulte (Pimienta 2021).

Para a versão 3, foi dedicado um esforço profundo e sistemático para equilibrar os dados da Wikipédia com dados equivalentes para outras línguas. A tabela do Apêndice 2 lista as enciclopédias on-line processadas com os dados coletados, principalmente em número de artigos. A partir desta tabela, construiu-se o indicador de conteúdo com uma representação mais justa dos línguas ao acumular, por língua, os diferentes números de artigos. A conclusão deste esforço pesado, necessário, mas em última análise frustrante, é que algumas línguas (como o chinês ou o turco) investiram pesadamente em enciclopédias online, enquanto outras não parecem estar interessadas na questão.

Foi um verdadeiro dilema abandonar as maravilhosas estatísticas da Wikimedia; No entanto, a remoção do indicador de conteúdo como dados de entrada levou à renovação positiva da concepção da abordagem no sentido de um modelo coerente onde os vieses são controlados.

Em vez de chamar o poder de saída principal, ele foi renomeado diretamente como conteúdo e capacidade e gradiente, com a mesma operação aritmética tornou-se indicador de presença virtual e produtividade de conteúdos, conceitos muito mais naturais e compreensíveis. Além disso, todas as operações de ponderação desenvolvidas no modelo a partir da versão 1 foram agora refletidas de forma coerente na conceituação da abordagem, como uma modulação da produtividade do conteúdo. Ao mesmo tempo, as anomalias mencionadas nas duas primeiras versões do modelo, motivadas pelas particularidades da Wikimedia, desapareceram para dar lugar a resultados mais fiáveis e previsíveis.[42].

[39]É o caso do cebuano, do malgaxe e do tagalo.
[40]Como hebraico, sueco ou servo-croata.
[41]Veja a definição precisa emhttps://meta.wikimedia.org/wiki/Wikipedia_article_profundidade
[42]O melhor sintoma é que o japonês alcançou o primeiro lugar em presença virtual e produtividade de conteúdo, o que é consistente com a realidade do uso onipresente da Internet no Japão. Algumas das línguas que têm sido favorecidas pela sua forte presença na Wikipédia permanecem em posições elevadas, mas não nos primeiros lugares, o que ajuda a manter a afirmação de que as línguas dos países (ou regiões) com melhor desempenho na Informação os parâmetros da sociedade beneficiam de bons lugares nos indicadores de presença virtual ou produtividade de conteúdo.

2.5 MODELAGEM

2.5.1 Pré-processamento

A maior parte dos dados fornecidos pelo Ethnologue está na forma de uma matriz Excel de 11.500 linhas no seguinte formato: "ISO639[43], *Nome do língua, Nome do país, número de falantes L1, número de falantes L2 »*, bem como um grande número de parâmetros associados não utilizados para este método e que foram removidos.

Para obter o formato exigido pelo modelo (uma matriz com todos os países considerados em colunas e todos os línguas considerados em linhas), uma série de etapas foi implementada com o apoio de diferentes programas escritos como macros VBA[44]. Uma das etapas mais complexas foi mesclar todos os dados de línguagens pertencentes à mesma macrolíngua. Este processo envolveu 60 macrolínguagens compreendendo 434 línguas diferentes[45](ver detalhes no apêndice 6).

Depois de concluída esta etapa, o processo consistiu em reduzir toda a lista de línguas apenas àqueles suportados pelo modelo, somando cuidadosamente todos os números restantes por país em uma única linha para o restante dos línguas.

É importante compreender que a adoção dos dados do Ethnologue implica a aceitação das suas regras de apresentação, que se baseiam em considerações puramente linguísticas:

- Agrupamento de macrolíngua[46]
- Lista de países e nomes em ingslês correspondentes.

A lista de países processada pelo Ethnologue é mais longa do que a processada pela UIT para o fornecimento de percentagens de ligação à Internet por país: a UIT, como entidade das Nações Unidas, não separa, por exemplo, a Martinica da França. Neste caso, a regra da UIT é a que prevalece e a exigência tem sido reunir cuidadosamente os dados do Ethnologue para os 29 países não considerados pela UIT (para a lista completa, ver #39;Anexo 7) numa única coluna " ;Outros países"[47].

2.5.2 Gestão de fontes para microindicadores

Todo o processo de gestão das fontes de microindicadores é a tarefa mais pesada e difícil do projeto, com elevado consumo de recursos humanos. Muitas etapas são necessárias:

1. Para cada indicador, verifique se as fontes de 2017 ainda estão disponíveis e atualizadas, caso contrário pesquise na Internet outras fontes comparáveis.
2. Selecione novas fontes com base em sua confiabilidade e aplicabilidade ao processo[48].

[43]O código ISO de 3 caracteres atribuído a cada um dos 7.486 idiomas identificados.
[44]Virtual Basic Applications, uma linguagem usada para criar macros executáveis no Excel.
[45]Por exemplo, a macrolíngua árabe reúne 29 línguas, como o árabe egípcio ou o árabe marroquino.
[46]Um exemplo significativo é o caso da macrolíngua servo-croata, cuja definição inclui, em ordem alfabética, o bósnio, o croata, o montenegrino e o sérvio. Este agrupamento não atende de forma alguma a critérios geopolíticos e pode até ser considerado controverso deste ponto de vista. Além disso, algumas fontes separam claramente as línguas e os países em causa, o que conduz a um risco de erro nos resultados, mesmo que a entrada das fontes tenha sido transformada para ter em conta esta situação (o risco surge quando os dados não não deve ser adicionado, mas sim calculado como no indicador de profundidade da Wikipedia).
[47]Note-se que o Kosovo dispõe de dados fornecidos pela UIT, mas está ausente da lista de países Etnólogos: por esta razão, não aparece nos resultados.
[48]Pode acontecer que dados fiáveis estejam num formato que proíba a exploração automatizada.

3. Reúna as fontes selecionadas em um formato que permita uma entrada simplificada no modelo.
4. Introduza fontes validadas no modelo.
5. Avalie o viés da fonte.

No Anexo 5 é apresentada a lista completa de fontes, para cada indicador.

Para realizar o passo 4, os dados devem ser transformados em formato Excel, com os nomes dos países e línguas correspondentes aos do modelo e na mesma ordem sequencial.

No passo 3, todas as fontes são coletadas de uma URL específica (veja o Apêndice 5 para a lista completa de URLs) e a maioria das fontes são obtidas em formato HTML. Algumas fontes estão em formato PDF e um subconjunto limitado (principalmente as da UIT e do Banco Mundial) está em formato Excel, o escolhido para transformar todas as fontes. O processo de conversão de PDF para Excel pode ser relativamente simples na maioria dos casos quando as tabelas estão bem estruturadas, mas em alguns casos há uma incompatibilidade e alguns truques são necessários, como passar por um formato .doc intermediário.

O processo de transformação de HTML para Excel muitas vezes pode se tornar um pesadelo exigindo muita imaginação, inclusive em alguns casos a necessidade de buscar os dados dentro do código-fonte HTML e a partir daí, tentar construir uma tabela utilizando a função de conversão do Excel, após limpar o HTML em torno dos dados.

Num número crescente de casos, a fonte oferece acesso geográfico aos dados (mapas clicáveis) que, exceto quando o número de países ou línguas é limitado e a cópia manual não é muito complicada, impossibilita o processamento automatizado ou exige o terceirização do trabalho de coleta manual que é tedioso e exige muita concentração e disciplina para evitar erros.

O crédito deve ser dado a instituições (geralmente organizações internacionais ou ONGs) que fornecem os dados em formato legível por computador (a Wikimedia fornece, por exemplo, na sua versão em inglês, tabelas HTML que podem ser transformadas diretamente em formato Excel sem perda de estrutura) .

Obter uma cópia da fonte em Excel ou formato compatível (geralmente uma tabela de nomes de países ou línguas com valores ou porcentagens associados) não é o fim do processo. Com 215 países e 329 línguas para lidar e, em vez de usar um código ISO inequívoco, o uso comum de nomes literais que podem estar em diferentes línguas e grafias não padronizadas, a integração dos dados no modelo não pode ser feito à mão. Dois programas foram escritos para este processo, ambos exigindo ajuste recursivo[49]para acomodar grafias diferentes. As saídas do programa são arquivos Excel que podem ser usados diretamente para integrar os dados no modelo. Além da considerável economia de tempo deste método informatizado, garante dados livres de erros.

Observe também que o gerenciamento das macrolínguagens tornou esse processo ainda mais complexo, pois o agrupamento das línguagens deve ser realizado nos dados de origem antes do processamento pela macro. Para citar alguns exemplos, as ocorrências frequentes do árabe egípcio ou marroquino nas fontes foram combinadas na macrolíngua árabe e as do sérvio, bósnio, croata e montenegrino foram fundidas no servo-croata (sendo o número de casos semelhantes bastante alto). Para o processamento manual de grafias desconhecidas informadas

[49]O processo recursivo reconhece novas grafias e termina quando a verificação de erros não identifica mais grafias desconhecidas.

pelo programa (incorporação de grafias como sinônimos ou rejeição na outra categoria), foi utilizada como suporte a página descritiva do Ethnologue de cada código de língua.[50].

2.5.3 Estrutura e processo do modelo
O modelo é implementado em um arquivo Excel com 17 abas que são apresentadas a seguir, com o processo correspondente.

UIT: cópia da fonte ITU, modificada conforme pré-processamento.

SP:a matriz de Falantes L1+L2 por país.
Em linhas, os 329 línguas, ordenados por código ISO de 3 dígitos (ISO369), começando na linha 9 com a soma dos demais línguas não processados.
Nas colunas, os 215 países tratados, ordenados por código ISO de 2 dígitos, começando na coluna I com a soma dos restantes países não tratados.

As primeiras 8 linhas e colunas são reservadas para informações de controle:
Linhas de controle: código do país de 3 caracteres, código do país de 2 caracteres, nome do país, total de falantes L1+L2 no país, % de pessoas conectadas, número de pessoas conectadas, % de conexão global, total ou média (número de línguas falado por país), restantes línguas.
Colunas de controle: ISO639, nome do língua[51], total de Falantes L1+L2, % global de Falantes L1+L2, % global de Falantes L1+L2 conectados, número de países com Falantes, número de Falantes L1, proporção L1+L2/L1, países restantes.

Esta folha está protegida contra leitura porque contém informações proprietárias da Ethnologue que não podem ser tornadas públicas.

SP2: (para SP número 2) dados demolinguísticos secundários calculados a partir de SP.
Para os 329 línguas on-line e o restante dos línguas: % mundial de falantes L1+L2, número de falantes L1+L2, % mundial de falantes L1+L2 conectados, número de falantes L1+L2 conectados, % mundial de falantes L1 +L2 conectados, % de Falantes L1 conectados em todo o mundo, % de usuários de Internet L1+L2 em todo o mundo.

PT: (para "Porcentagem de Lingua") Matriz paralela a SP onde PL(i,j) = % de usuários de Internet do língua i do país j conectados, calculado a partir de SP e SP2. Esta é uma informação redundante usada para simplificar a operação de ponderação realizada na aba Wut.

Mil: (para "Lingua do Micro-Indicador") Contém a lista de línguas em linhas e o valor 0 ou 1 dependendo da ausência ou presença do língua em um dos 16 aplicativos usados para a interface do indicador #39;preenchida de fontes linguísticas.

Mic: (para "País Microindicador") Contém a lista de países em colunas e os dados de origem por país, sucessivamente para entradas de índice, usos e tráfego. Para a versão 3 existem 786 linhas.

Note-se que o pré-processamento é necessário para utilizações que visam integrar redes sociais não ocidentais; Isso é feito na guia MIcU.

[50] https://www.ethnologue.com/linguagem/srp
[51]Seguido de "macro" se for uma linguagem macro.

Observe que o pós-processamento é necessário para o tráfego, a fim de ponderá-lo com o número ideal de sites com base na porcentagem de pessoas conectadas por país; isso é feito no MIcT e no MIcT1.

As colunas de controle são as seguintes:
Colarinho. A: Indica o tipo de indicador a partir de uma busca pelo nome em MATRIX.
Colarinho. B: indica o nome do indicador
Colarinho. C: dependendo do tipo de dados, calcula a média ou total, ou um produto matricial com o número de pessoas conectadas por país das entradas em cada linha
Colarinho. D: indica o tipo de dados, seja uma percentagem global por país, ou uma quantidade por país, ou uma percentagem dentro de cada país
Colarinho. E: indica se a extrapolação é necessária e, em caso afirmativo, qual dos dois tipos de extrapolação
Coluna F: calcula o número de países com dados de origem
Colarinho. H: contém a URL da fonte exceto para tráfego onde indica o número de vezes que o site foi citado, para permitir a ponderação correspondente em Wut.

As linhas de controle são as seguintes:
A linha 8 indica para cada país o número de websites que foram medidos.
A linha 9 indica a proporção do número de sites que deveriam ter sido usados para respeitar a proporcionalidade das pessoas conectadas (o produto da linha 8 pela linha 9 para cada país representa o número exato de sites necessários para aquele país na suposição do total real sites (célula C7). Isso será usado como um fator de ponderação para obter uma representação justa das medições de tráfego no MIcT1 e MIcT antes de ponderar pelo número de ocorrências de sites feitas no Wut (isso foi adicionado na V3.c para corrigir um erro em V3 onde a ponderação foi feita em paralelo com a ponderação demolinguística, o que foi um erro com consequências muito marginais).

MIc: (para “Uso do país do microindicador”) A última guia adicionada para o novo processamento de uso V3. Inclui uma cópia completa das fontes de uso de MIc migradas, nos mesmos moldes, complementada pelo T-Index e a lista de novas redes sociais adicionadas para V3. Para esta nova lista, são definidas métricas de tráfego de país obtidas do SimilarWeb, seguidas do processo de extrapolação (ver EX). O resultado é uma linha nova e final chamada "Mídia Social Ponderada" que é obtido ponderando toda a lista de redes sociais com o correspondente total de assinantes, equilibrando, em última análise, as redes sociais ocidentais com as do resto do mundo.

MA: (Máscara de ausência) Aba paralela ao MIc contendo o valor 1 quando faltam dados para o casal (país, fonte). Usado para extrapolação.

MP: (Máscara de presença) Uma aba paralela ao MIc contendo o valor 1 quando existem dados para o casal (país, fonte). Usado para extrapolação.

EX: (Extrapolação) Uma aba paralela ao MIc onde o processo de extrapolação é realizado. Dois processos diferentes são usados dependendo do tipo de dados.

Quando os dados são expressos como uma percentagem global por país, o complemento de 100% é distribuído entre os países que não receberam dados, proporcionalmente à sua percentagem global de pessoas ligadas à Internet. Este é normalmente o caso da medição de tráfego em que as ferramentas utilizadas, Alexa e SimilarWeb, não cobrem todos os países.

Quando os dados são pontuados por país, utiliza-se a técnica do quartil: são colocados quatro valores quartis com base na porcentagem de pessoas conectadas no intervalo entre: 0%, 15%, 35%, 65%, 85% e 100%. Este é normalmente o caso dos dados de índices.

Observe que quando parece que um ou outro dos métodos não pode fornecer uma extrapolação significativa, a fonte de dados é excluída do modelo.

Nos raros casos em que todos os países são reportados pela fonte, nenhuma extrapolação é obviamente necessária, como é o caso dos dados do NapoleonCat sobre a percentagem de assinantes por rede social.

Observe que para o processo de utilização V3 a extrapolação para redes sociais não é realizada no EX e foi replicada no MIcU.,

Observe que a soma dos valores do T-Index para os países listados é de 99,78%, tão próximo de 100% que nenhuma extrapolação foi feita.

MicT1: ("Micro-Indicador de Tráfego1") A aba é paralela ao MIc e é preenchida apenas para indicadores de tráfego. Cada célula (país, site) contém o produto do tráfego de origem do MIc adicionado ao tráfego extrapolado do EX, multiplicado pelo fator de ponderação do país (linha 10 do MIc. A soma das porcentagens é calculada e colocada na coluna G para posterior normalização para 100% em MiCT.

MicU: a guia é paralela ao MIcT1 e é usada para normalizar os dados para 100% para cada site, dividindo cada célula pelo total. Os resultados serão utilizados no Wut para calcular a distribuição final do tráfego por língua.

WuT: ("Ponderação de Utilizações e Tráfego") Neste separador são processados os indicadores de utilização e tráfego. O processo consiste em ponderar os valores com a percentagem de falantes ligados por país, do PL, após aplicação de extrapolação. Para o indicador de tráfego a extrapolação já foi realizada no MIcT mas há uma ponderação adicional a realizar com os dados calculados no MIc (coluna H) para o número de ocorrências de cada website.

Wi: ("Índices de Ponderação" - índices de ponderação) Nesta planilha, a ponderação é realizada com os dados demolinguísticos de forma a obter dados por língua para os índices de indicadores, da coluna BA, seguida, a partir da coluna 10, de normalização para 100 %.

Pi1: ("Lingua do indicador de processo" – processa o indicador de língua) Nesta aba é realizada a ponderação com dados demolinguísticos, a fim de obter dados por língua, para indicadores por país, usos e tráfego. Para efeitos, é realizada uma ponderação adicional com o peso atribuído a cada componente deste indicador (ver 2.2.4). Para o tráfego, a ponderação adicional é feita com o número de ocorrências de cada site da amostra (ver 2.2.5)

RES:(Resultados) São calculados os resultados finais de cada indicador por língua (usos, tráfego, índices, interfaces).

FINAL: Esta aba apresenta os resultados finais com todos os parâmetros associados e oferece os resultados classificados por conteúdo, presença virtual, produtividade de conteúdo e Falantes conectados. Também apresenta as 20 primeiras posições de conteúdo e cria a cibergeografia do

resultado linguístico (ver Tabela 1). Uma cópia desta guia sem fórmula é tornada pública como um produto modelo (consultehttp://obdilci/Resultados).

Note-se que o acesso a estes resultados sob a forma de base de dados está planeado antes do final de 2022, com códigos ISO 639-2 como chave de acesso.

MATRIZ: A lista de todos os microindicadores utilizados no modelo para cada tipo (índices, interfaces, usos, tráfego).

3. RESULTADOS

O modelo desenvolvido produz, para cada língua, os seguintes indicadores:

A. Porcentagem de falantes L1+L2 no mundo
B. Porcentagem de falantes L1+L2 conectados
C. Porcentagem de falantes conectados L1+L2
D. Porcentagem de conteúdo da Internet
E. Indicador de presença virtual (definido como relação D/A)
F. Indicador de produtividade de conteúdo (definido como relação D/C)

A partir da agregação destes indicadores são realizadas construções mais elaboradas, como o quadro seguinte "Ciber-Geografia das Famílias Linguísticas", que dá uma perspectiva global da situação das diferentes famílias linguísticas.[52]e mostra que as línguas asiáticas estão prestes a ter precedência sobre as línguas europeias enquanto as línguas africanas estão numa situação difícil, devido à exclusão digital prevalecente, traduzida numa divisão linguística[53].

Tablela 3: Cibergeografia das famílias linguísticas

Linguas de	África	Américas	Mundo árabe	Ásia	Europa	Pacífico	Não incluso	TOTAL
Falantes L1 + L2	9,21%	0,31%	3,53%	48,24%	30,91%		7,81%	**100%**
Usuários de internet %	29,8%	56,7%	64,0%	49,3%	82,6%		47,06%	**56,91%**
% Usuários de internet	5,21%	0,32%	3,89%	44,63%	39,51%		6,36%	**100%**
Conteúdo	2,89%	0,22%	3,09%	44,77%	45,39%		3,64%	**100%**
Presença virtual	0,31	0,71	0,88	0,93	1,47		0,47	**1**
Produtividade de conteúdo	0,56	0,69	0,79	1,00	1,15		0,57	**1**
Número de línguas	138	8	1	135	47	0		**329**

Os resultados do modelo podem ser visualizados no modo CC-BY-SA-4.0, emhttps://obdilci.org/lc2022e pode ser lido em (Pimienta, 2022). Os resultados das medições anteriores e subsequentes estão acessíveis emhttps://obdilci.org/Results.

[52]As famílias de línguas incluem, para cada região, as línguas nativas daquela região. Inglês, francês e espanhol são línguas europeias e de acordo com a classificação Ethnologue que utilizamos, o russo é classificado como língua europeia enquanto o turco e o hebraico são línguas asiáticas.
[53]Menos de 30% dos falantes de línguas africanas estão ligados à Internet e conseguem uma presença virtual e uma produtividade de conteúdos muito baixas.

Para efeitos de verificação cruzada dos resultados, o modelo foi executado separadamente apenas com dados L1 e apenas com dados L2 (ver Apêndice 9 para os resultados correspondentes que representam um controlo indireto completamente positivo do método).

4. DISCUSSÃO

A observação da presença de línguas na Internet conheceu forte actividade no período 2000-2007 (Pimienta, 2009) mas, após este período, como referido na introdução, apenas duas opções permaneceram disponíveis para o grande público: InternetWorldStats e W3techs.

Ambos apresentam alguns destaques de suas respectivas metodologias, mas nenhum artigo científico revisado por pares abordou seus respectivos vieses; a sua presença de longa data, sem dados alternativos, garantiu-lhes um grande número de citações em vários artigos que requerem esses dados, muitas vezes sem a cautela necessária que exigiria a realidade dos seus vieses.

4.1 Viés do InternetWorldStats (IWS)

Os dados produzidos pelo IWS diferem ligeiramente dos do Observatório, principalmente porque as fontes dos dados demolinguísticos não são as mesmas e, especialmente para os dados L2, as diferenças entre as fontes podem ser enormes (ver nota 26). No entanto, existe outra diferença no gerenciamento de dados L2. O Observatório calcula as porcentagens L1 + L2 das línguas do mundo em relação ao número total de falantes L1 + L2. Ou seja, um valor 43% superior à população mundial, segundo fonte do Ethnologue.[54]. Já o IWS calcula as porcentagens para L1+L2 em relação à população mundial (um método denominado abordagem de soma zero[55]). A menos que haja um truque escondido em algum lugar nos cálculos, a abordagem de "soma zero" parece causar um erro grosseiro ao supervalorizar os 10 línguas mencionados, erro oculto na porcentagem dos demais línguas, que se tornará negativo em algum momento se o número de línguas for estendido até o ponto em que a soma dos falantes L1+L2 cruze o valor L1.

4.2 Viés da W3Techs

O método utilizado pela W3Techs envolve a aplicação de um algoritmo de reconhecimento de língua à página inicial de 10 milhões de websites que são selecionados por determinados serviços de análise de tráfego web (Alexa.com ou trancolist.eu, até ao final de 2022) como os mais visitados.

As diferenças entre os resultados da W3Techs e os do Observatório são enormes, muitas vezes numa proporção de 1 para 3, por vezes, como para o chinês e o hindi, numa proporção superior a 1 para 10). Pelo menos uma das duas fontes deve ser extremamente tendenciosa! A tabela a seguir descreve essas diferenças usando dados W3Techs de 24/08/22 e dados do Observatório de V3.1 em 2022/08.

[54]Nos dados de 2021, que utilizamos, o Ethnologue conta a população global (número total de falantes de L1) em 7.231.699.136 e o número total de falantes de L1+L2 em 10.361.716.756.

[55]Citação do site do IWS: Na verdade, muitas pessoas são bilíngues ou multilíngues, mas aqui atribuímos apenas um idioma por pessoa, de modo que todos os totais de idiomas somam o total da população mundial (abordagem de soma zero).

Tablela 4: Comparação de dados W3Techs vs Observatório

	W3TECH		OBSERVATÓRIO	
LINGUA	Classificação	Conteúdo[56]	Classificação	Conteúdo
Inglês	1	61,4%	1	19,92%
Russo	2	5,6%	4	3,86%
Espanhol	3	3,9%	3	8,09%
Turco	4	3,2%	12	1,15%
Alemão	5	3,1%	dez	2,38%
Francês	6	3,0%	6	3,43%
Persa	7	2,7%	16	0,89%
Chinês	9	1,7%	2	19,82%
Arabe	13	1,1%	8	3,14%
Hindi	35	0,1%	5	3,67%

As maiores diferenças estão nos percentuais de conteúdo para hindi e chinês, além da diferença em inglês (mais de 60% versus cerca de 20%).

Em agosto de 2022, o agregador de estatísticas Statista[57], apoiando-se em dados da W3Techs, afirma que "o inglês é a língua universal da Internet", enquanto o Observatório, ao mesmo tempo, afirma: "A transição da #39;Internet entre o domínio das línguas europeias, o inglês na liderança , em relação às línguas asiáticas e ao árabe, com o chinês na liderança, está bem avançado e o vencedor é o multilinguismo, mas as línguas africanas estão atrasadas para ocupar o seu lugar. Mais uma vez, estas duas afirmações não são compatíveis, pelo menos uma é falsa.

Poderíamos discutir o viés inglês dos algoritmos de reconhecimento de língua, o viés inglês da seleção dos 10 milhões de sites mais visitados[58]; mas estes são vieses marginais que não poderiam explicar diferenças tão significativas. O principal problema está na falta de consideração do multilinguismo, uma característica da Web que é ignorada pelo método W3Techs, embora a Web seja provavelmente ainda mais multilingue do que a humanidade.[59].

Como pano de fundo para esta discussão, é importante recordar o ponto levantado e documentado no Apêndice 8, nomeadamente que os utilizadores da Internet preferem usar a sua língua nativa na Internet como primeira opção e estão interessados em usar a(s) sua(s) segunda(s) língua(s) além .

O problema reside, portanto, na decisão de medir apenas as páginas iniciais e contar apenas um língua para cada uma. Muitos sites que não são em inglês podem ter resumos em inglês ou algumas palavras em inglês em suas páginas iniciais e provavelmente são contados como inglês. Muitos sites em inglês têm versões em vários outros línguas que também devem ser contabilizados (se, como é provável o caso, o algoritmo for definido em um ambiente de computação em inglês, o site será contado apenas como em inglês).

[56]Observe que a W3Techs oferece valores com apenas um dígito após a vírgula.
[57]https://www.statista.com/chart/26884/languages-on-the-internet/
[58]De acordo comhttps://news.netcraft.com/archives/category/web-server-survey, em maio de 2022 existiam 1,16 mil milhões de websites, dos quais 270 milhões estão ativos. A cobertura dos mais visitados é então inferior a 4% do total.
[59]Este seria o caso se os 270 milhões de sites ativos juntos oferecessem mais de 400 milhões de interfaces de idiomas diferentes, uma média de cerca de 1,5 idiomas por site.

A W3Techs produziria dados bastante diferentes (e esperançosamente mais próximos dos do Observatório) se as seguintes regras fossem adicionadas ao seu algoritmo:

- A contagem é feita em páginas da web, não em sites.
- O algoritmo verifica a existência de opções de língua na página inicial e conta cada língua oferecido como opção.
- Se não houver opções de língua, o algoritmo verifica a existência de um língua diferente do inglês na página inicial, se for o caso, ele conta este site neste língua em vez de inglês.
- O algoritmo avalia um número aproximado de páginas do site e multiplica cada número de línguas por esse número após dividir pelo número de opções de línguas.

Em Pimienta (2023) tenta-se desviar a figura da W3Techs para o conteúdo em inglês, avaliando a taxa de multilinguismo da amostragem Tranco utilizada pela W3Tech, e a partir disso, estabelecendo a correção nos resultados.

É uma equação simples: $P' = (P - Err) / Rm$, onde:

- ✓ P é a porcentagem dada pela W3Techs para conteúdo em inglês
- ✓ P' é a porcentagem distorcida para conteúdo em inglês
- ✓ Err é a porcentagem de sites contados com erro em inglês
- ✓ Rm é a taxa de multilinguismo da amostra.

A partir dos dados calculados, a janela da percentagem de conteúdo em inglês deslizaria dos 50% - 60%, anunciados pela W3Techs, para os 20% - 30% anunciados pelo Observatório ou pelo consórcio universitário grego que estudou os ccTLDs da UE.

O Observatório encorajou os colegas gregos a aplicar o seu algoritmo à lista de sites Tranco, com uma resposta promissora. Isto contribuiria definitivamente para este debate, uma vez que o seu método honra o multilinguismo da web. Esta é uma perspectiva otimista para os próximos meses para qualquer pessoa interessada neste tema.

5. CONCLUSÃO

Pela primeira vez na história da Internet, um método é capaz de oferecer uma variedade de indicadores significativos sobre a presença de 329 línguas na Web. O modelo fornece resultados consistentes com estudos anteriores realizados pelo Observatório, mas está em forte contradição com os resultados fornecidos pela fonte única que cobre o assunto desde 2011; Mostra, em particular, que o conteúdo em inglês na Web está hoje ao mesmo nível que o conteúdo em chinês, cerca de 20%, enquanto os meios de comunicação continuam a reportar conteúdo em inglês bem acima dos 50%.

O método utilizado para obtenção desses resultados é exposto de forma completa e transparente e seus vieses são discutidos abertamente para que a comunidade científica possa analisá-los.

Estes resultados reflectem simplesmente um passo lógico na evolução da Web, que passou de uma primeira fase centrada no inglês (1992-2000), para uma segunda fase centrada nas línguas europeias, com liderança inglesa (2000-2010), seguida por uma fase internacionalizada, com a ascensão das línguas asiáticas e árabes e ainda uma lacuna significativa que deixa as línguas africanas para trás, com uma Web cada dia mais multilingue (2010-2020). A próxima fase

(2020-2030) verá provavelmente uma rede mais homogénea em termos de representação linguística, esperando-se que a exclusão digital comece a ser superada em África, abrindo o espaço linguístico local de África. O enraizamento do multilinguismo na Web está em curso e poderá muito bem ultrapassar o da humanidade, se ainda não for o caso. No entanto, prevalecerão diferenças na produtividade dos conteúdos, mantendo-se algumas vantagens para determinadas línguas que tenham uma combinação de uma grande população L2 e presença num grande número de países (como o inglês e o francês).

A surpresa não deve advir dos dados produzidos pelo Observatório que são apenas o reflexo, na sua componente cibernética, da evolução natural do mundo; deveriam provir do facto de dados altamente tendenciosos terem sido admitidos na última década sem grande reacção por parte da comunidade científica.

Esperançosamente, a total transparência do método ajudará mais mentes científicas a desafiar os resultados fornecidos pelo mundo do marketing e a reposicionar este tópico onde deveria pertencer: a comunidade científica. Obviamente, isso inclui questionar o método descrito acima e detectar e discutir possíveis vieses que não foram detectados pelos autores. Deixe a abordagem científica ter precedência sobre o marketing!

REFERÊNCIAS

Baubock, R. (2015). O valor político das línguas. Crítico. Rev. Internacional Soc. Pol. Fil. 18, 212–223. doi:10.1080/13698230.2015.1023635

Flint, C. (2021). Introdução à Geopolítica. Milton Park: Routledge.

Gazzola, M. (2015). Il Valore Economico Delle Lingue (O valor econômico das línguas). Disponível online em: https://ssrn.com/abstract=2691086 (acessado em 22 de abril de 2023).

Giannakoulopoulos, A., Pergantis, M., Konstantinou, N., Lamprogeorgos, A., Limniati, L., Varlamis, I. (2020). Explorar o domínio da língua inglesa nos sites dos países da UE. Era. Internacional 12, 76. doi: 10.3390/fi12040076

Grefenstette, G. e Noche, J. (2000). Estimativa do uso de línguas inglesas e não inglesas na WWW. Rhone-Alpes: Xerox Research Centre Europe. Disponível on-line em:http://arxiv.org/ftp/cs/papers/0006/0006032.pdf

Grin, F. e Vaillancourt, F. (1997). A economia do multilinguismo: visão geral e quadro analítico. Anu. Rev. Apl. Linguista. 17, 43–65. doi: 10.1017/S0267190500003275

Heller, M. (2010). A mercantilização da língua. Ana. Rev. Antropol. 39, 101–114. doi: 10.1146/annurev.anthro.012809.104951

Lavoie, BF e O'Neill, ET (1999). Quão “mundial” é a Web? Revisão Anual da Pesquisa da OCLC.

Mikami, Y., Zavarsky, P., Rozan, MZA, Suzuki, I., Takahashi, M., Mak, T., et al. (2005). O projeto do observatório de línguas (LOP). In: Faixas e pôsteres de interesse especial da 14ª Conferência Internacional sobre World Wide Web. 990–991. Disponível on-line em:http://eprints.utm.my/id/eprint/3405/1/The_Language_Observatory_Project_%28LOP%29.pdf

Monrás, F., Medina, M., Cabré, S., Canto, P., Meléndez, V., Ripoll, E., et al. (2006). Estatística da presença do catalão na xarxa da Internet e das características das Webs catalãs, em Llengua i ús: Revista técnica de política linguística. Não. 37, 62–66. Disponível on-line em:https://raco.cat/index.php/LlenguaUs/article/view/128275

O'Hara, K. e Hall, W. (2018). Quatro Internets: A Geopolítica da Governança Digital.Waterloo: Centro para Inovação em Governança Internacional.

Oliveira, GM (2010). O lugar das línguas. América do Sul e os mercados linguísticos na nova economia. Brasil: Sinergias Brasil. 21–30.

O'Neill, ET, Lavoie, BF e Bennett, R. (2003). Tendências na evolução da PublicWeb: 1998 - 2002. Reston: D-Lib Magazine.

Pimienta, D. (2014). Francês na Internet, Relatório 2014 “A língua francesa no mundo”. Nathan: OIF. 501.

Pimienta, D. (2021). Internet e diversidade linguística: a cibergeografia das línguas com o maior número de falantes, LinguaPax Review 2021. Barcelona: Tecnologias linguísticas e diversidade linguística. 9–17. Disponível on-line em:https://www.línguapax.org/wp-content/uploads/2022/02/LinguapaxReview9-2021-low.pdf

Pimienta, D. (2022). Recurso: Indicadores sobre a presença de línguas na Internet em Anais da 1ª Reunião Anual do Grupo de Interesse Especial ELRA/ISCA sobre Línguas com Poucos Recursos, Marselha. Associação Europeia de Recursos Linguísticos. 83–91. Disponível on-line em:https://aclanthology.org/2022.sigul-1.11/

Pimienta, D. (2023). É verdade que mais de metade dos conteúdos da Web estão em inglês? Se o multilinguismo da Web receber a devida atenção, então não! Pré-impressão do ResearchGate. doi: 10.13140/RG.2.2.20767.43683

Pimienta, D. e Oliveira, GM (2022a). Cibergeografia das Línguas. Parte 2: O Fator Demográfico e o Crescimento das Línguas Asiáticas e do Árabe. Alberta: Revisão Internacional de Ética da Informação. 32. Disponível on-line em:https://informationethics.ca/index.php/irie/article/view/488

Pimienta, D. e Oliveira, GM (2022b). Cibergeografia das Línguas. Parte 1: Método, Resultados e Foco no Inglês. Alberta: Revisão Internacional de Ética da Informação. 32. Disponível on-line em:https://informationethics.ca/index.php/irie/article/view/491

Pimienta, D. e Prado, D. (2016). Medição da presença da língua espanhola na Internet: métodos e resultados. Revista Espanhola de Documentação Científica 39, e141. doi: 10.3989/redc.2016.3.1328

Pimienta, D., Prado, D. e Blanco, Á. (2009). Doze anos medindo a diversidade linguística na Internet: equilíbrio e perspectivas. Paris: Publicações da UNESCO para a Cúpula Mundial sobre a Sociedade da Informação. Disponível on-line em:http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Simons, GF, Thomas, AL e White, CK (2023). Avaliando o suporte à língua digital em escala global, nos Anais da 29ª Conferência Internacional sobre Lingüística Computacional. Gyeongju: Comitê Internacional de Lingüística Computacional. 4299–4305. Disponível on-line em:https://aclanthology.org/2022.coling-1.379.pdf

ANEXO 1: FONTES DO INDICADOR DE USOS

Tablela 5: Redes sociais selecionadas e número total de assinantes

REDE SOCIAL	TOTAL DE ASSINANTES (Milhão)
Whatsapp	2000
Conversamos	1225
TikTok	732
Douyin	600
Telegrama	600
QQ	595
Snapchat	528
Weibo	521
Zona Q	517
Kuaishou	481
Quora	300
Skype	300
Tieba	300
Viber	260
OMI	200
LINHA	169
arte	150
Gosto	150
Discórdia	140
Contração muscular	140
StackExchange	100
V.K.	650
Odnoklassniki	200
Douban	200
MOJ	160
JOSH	115
Compartilhar bate-papo	160
% de usuários do FACEBOOK por país (NapoleonCat 2021)	1455
%usuários do INSTAGRAM por país (NapoleonCat 2021)	1200
MESSENGER% de usuários por país (NapoleonCat 2021)	1300
LINKEDIN% de usuários por país (NapoleonCat 2021)	155
FACEBOOK Mundial % do IWS 2021	1455
Linkedin% de usuários por país (ApolloTech 2021)	155
% de usuários do Twitter por país (Statista 2021)	396
% de audiência do Pinterest (Statista 2021)	460
% de usuários do REDDIT por país (Statista 2021)	430

Tablela 6: Fontes de dados para redes sociais

REDES SOCIAIS	FONTE
% de usuários do FACEBOOK por país (NapoleonCat 2021)	https://napoleoncat.com/stats/
%usuários do INSTAGRAM por país (NapoleonCat 2021)	https://napoleoncat.com/stats/
MESSENGER% de usuários por país (NapoleonCat 2021)	https://napoleoncat.com/stats/
LINKEDIN% de usuários por país (NapoleonCat 2021)	https://napoleoncat.com/stats/
Linkedin% de usuários por país (ApolloTech 2021)	https://www.apollotechnical.com/linkedin-users-by-country/
% de usuários do Twitter por país (Statista 2021)	https://www.statista.com/statistiques/242606/ número de usuários ativos do Twitter em países selecionados/
FACEBOOK Mundial % do IWS 2021	https://www.internetworldstats.com/stats1.htm + stats2.htm+…stats6.htm
% de audiência do Facebook (Statista 2021)	https://www.statista.com/statistiques/268136/ 15 principais países com base no número de usuários do Facebook/
YouTube % de conectados no país (Statista 2021)	https://www.statista.com/statistics/1219589/youtube-penetration-worldwide-by-country/
% de assinantes da Netflix por país (CompariTech 2020)	https://www.comparitech.com/tv-streaming/netflix-subscribers/
% de audiência do Pinterest (Statista 2021)	https://www.statista.com/statistics/328106/pinterest-penetration-markets/
% de usuários do REDDIT por país (Statista 2021)	https://backlinko.com/reddit-users
Cumulativo. 2012/21% de downloads do OpenOffice por país	http://www.openoffice.org/stats/countries.html
# Servidores de Internet seguros	https://data.worldbank.org/indicator/IT.NET.SECR
% de assinantes de banda larga fixa no país (BM 2021)	https://data.worldbank.org/indicator/IT.NET.BBND.P2
% Tal. assinatura de telefone fixo + celular no país (BM 2021)	https://data.worldbank.org/indicator/IT.MLT.MAIN.P2 + https://data.worldbank.org/indicator/IT.CEL.SETS.P2

ANEXO 2: ENCICLOPÉDIAS ONLINE ANALISADAS

Tablela 7: Enciclopédias on-line

LINGUA	ENCICLOPÉDIA	NÚMERO DE ÍTENS (Milhões)	OUTRA INFORMAÇÃO
Diversos	Enciclopédia da vida (fim da vida)	0,75 (2010) 1,9 hoje	Linguas suportados: árabe, português brasileiro, inglês, finlandês, francês, macedônio, piemontês, chinês tradicional e turco Linguas da interface: os mesmos, além de alemão, espanhol, holandês, turco e ucraniano.
Diversos	thefreedictionary.com/ Gratuito com publicidade ou pago	sem estatísticas	Inglês, espanhol, alemão, francês, italiano, chinês, português, holandês, norueguês, grego, árabe, polonês, turco, russo, hebraico Não está claro se esta é uma versão paralela ou uma língua específica.
Diversos	fr.metapedia.org/ versão neonazista da Wikipédia	Marginal (5.000 artigos em inglês)	Checo, Dinamarquês, Alemão, Espanhol, Inglês, Húngaro, Holandês, Português, Romeno, Esloveno, Sueco, Estónio, Croata, Islandês, Norueguês, Macedónio
Chines	Baidu Baiké	24,5	194 milhões de alterações 7,5 milhões de editores
Chines	Baiké (Hudong)	18	5,8 milhões de editores (2013)
Chines	Sogou Baike	????[60]	
Arabe	Marefa	0,136636	2,4 milhões de páginas
Arabe	Mawdoo3	0,15	45 (2018)
Bengali e ingles	Bengala	0,0057	1.450 editores
Croata	enciklopedija.hr	0,067	Imprimir dados da versão
Croata	proleksis.lzmk.hr	0,062	
Dinamarques	Den Store Danske	0,161	1100 editores 1 milhão de usuários
Holandes	winklerprins. com	0,0115	por assinatura
Ingles	britannica.com		acesso gratuito limitado
Ingles	Everipédia Artigos copiados da Wikipedia	?	7.000 editores ativos (2019) Usuários 3M (2017) acesso aberto, mas também mercado blockchain
Ingles	Cidadania	0,017	As estatísticas pararam em 2014 perto de parar
Ingles	Conservapédia	0,0518	800 milhões de visualizações de páginas 1,5 milhão de alterações
Ingles	Scholarpédia	0,0018	Dados marginais
Ingles	Enciclopédia.com	0,3	Agregador formal de enciclopédia
Ingles	Enciclopédia da Colômbia		Agregado por Encyclopedia.com
Ingles	digitaluniverse.net		desligada
Frances	Larousse	0,317	
Alemão	babador retrô	0,3	
Hebraico Ingles	Hamichlol	0,28	Versão censurada da Wikipedia para um público hiper-religioso
Coreano	Doopédia	0,588	
Malaio sudanes javanes	Superpédia	0,02	
Italiano	Treccani	0,9	

[60]Sogou Baike é considerado pelo menos tão importante quanto Baidu Baké e o mesmo valor do número de itens foi assumido.

Malaiala	Sarvavijnanakosam	0,007	
Marathi	Viswakosh	0,016	
Noruegues bokmal e nynorsk	Loja Norske Leksikon	0,2 (2019)	3 milhões de usuários/mês leem 500.000 artigos
Polones	enciclopédia.interia.pl	0,12 (2006)	
Polones	enciclopédia.pwn.pl	0,08	
Russo	Grande Enciclopédia Russa	0,012 (2016)	
Russo	Krugosvet	0,012	
Espanhol	https://www.ecured.cu/ cubano	0,237	73.000.537 editores ativos
Espanhol	Enciclonet	0,185	
Espanhol	enciclopedia.us.es/	0,053	https://wikiapiary.com/wiki/
Sueco	ne.se/	0,26 (2005)	
Tamil	não ON-line		
Turco	Ekşi Sozluk	8 milhões de entradas em 2009[61]	400.000 usuários 110.000 editores 4 milhões de novas admissões/ano em 2013[62] Aberto para publicação, cada entrada é mantida após moderação.
Vietnamita	parece ter desaparecido		Acesse archive.org - https://bachkhoatoanthu.vass.gov.vn

[61] https://www.researchgate.net/publication/242100750_Web_Based_Authorship_in_the_Context_of_User_Generated_Content_An_Analysis_of_a_Turkish_Web_Site_Eksi_Sozluk

[62] https://www.researchgate.net/publication/271521393_SOCIAL_MEDIA_IN_A_DICTIONARY_FORMAT_ONLINE_COMMUNITY_OF_eksisozlukcom/figures?lo=1

ANEXO 3: FONTES DO INDICADOR DE INTERFACE

Tablela 8: Fontes para indicador de interface

Linguas de tradução do Bing Translator	https://www.bing.com/translator/
Linguas suportados pela Amazon Kindle Direct Publishing	https://kdp.amazon.com/en_US/help/topic/G200673300
Linguas suportados pela Cortana	https://en.wikipedia.org/wiki/Cortana
Linguas WordReference suportados	https://www.w.com
Linguas de tradução do WordLingo	http://www.worldlingo.com/en/línguas/
Linguas suportados pelo Facebook	https://www.facebook.com/língua.php
Linguas de anúncios in-stream do Facebook suportados	https://www.facebook.com/business/help/267128784014981
Linguas de tradutor gratuito suportados	http://www.free-translator.com
Linguas suportados pelo Google Play Console	https://support.google.com/googleplay/android-developer/table/4419860?hl=en
Linguas suportados pelo Google Cloud	https://cloud.google.com/translate/docs/languages?hl=en
Linguas suportados pelo Google Tradutor	https://en.wikipedia.org/wiki/Google_Translate
Linguas suportados pelo Google Scholar	https://scholar.google.com/scholar_settings?sciifh=1&hl=fr&as_sdt=0,5#1
Lingua suportado pelo Paralink Translator	http://paralink.com
Linguas suportados pelo oinline-Translator	https://www.online-translator.com/traduction
Linguas suportados pelo Reverso Translator	https://www.reverso.net/text_translation.aspx?lang=EN
Linguas suportados pela Free-Translations	https://www.freetranslations.org
Linguas suportados pelo Skype	https://support.skype.com/en/faq/FA34781/what-languages-are-supported-in-skype
Linguas suportados pelo Systran	https://support.systran.net/systranlinks/faq/

ANEXO 4: FONTES DO INDICADOR DO ÍNDICE

Tablela 9: Fontes para o indicador de índices

Índice de governo eletrônico	https://publicadministration.un.org/egovkb/Data-Center
Índice de Participação Eletrônica	https://publicadministration.un.org/egovkb/Data-Center
Índice de serviços online	https://publicadministration.un.org/egovkb/Data-Center
Índice de Capital Humano	https://publicadministration.un.org/egovkb/Data-Center
Índice de Infraestrutura de Telecomunicações	https://publicadministration.un.org/egovkb/Data-Center
Índice Global de Preparação Digital da Cisco 2019	https://www.cisco.com/c/dam/en_us/about/csr/reports/global-digital-readiness-index.pdf
Índice de preparação para IA do governo 2020	https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf
Pontuações de liberdade na Internet,	https://freedomhouse.org/countries/freedom-net/scores
Índice Global de Conectividade	https://www.huawei.com/minisite/gci/en/country-rankings.html
Índice Global de Cibersegurança 2018	https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf
Índice de comércio eletrônico B2C da UNCTAD, 2020	https://unctad.org/system/files/official-document/tn_unctad_ict4d17_en.pdf
O Índice Global de Dados Abertos	https://index.okfn.org/place/
Classificação Global de Competitividade Digital 2020	https://www.imd.org/globalassets/wcc/docs/release-2020/digital/digital_2020.pdf
Índice de preparação para tecnologias de fronteira	https://unctad.org/system/files/official-document/tir2020_en.pdf
Índice Global de Inovação	https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2020.pdf
Acesso ao conhecimento básico	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Acesso à informação e comunicações	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Acesso ao ensino superior	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Acesso à eletricidade (% da população)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Acesso à educação de qualidade (0=desigual; 4=igual)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Acesso à governança eletrônica (0=baixo; 1=alto)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Censura da mídia (0=frequente; 4=raro)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Liberdade de expressão (0=sem liberdade; 1=liberdade total)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Universidades ponderadas pela qualidade (pontos)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Documentos citáveis	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Mulheres com ensino superior	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Anos de graduação	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx

ANEXO 5: SELEÇÃO DE SITES PARA O INDICADOR DE TRÁFEGO

Tablela 10: Seleção de sites para o indicador de tráfego

LOCAL NA REDE INTERNET	NÚMERO DE VEZES
10086.cn	1
10jqka.com.cn	1
122.gov.cn	1
12306.cn	1
12371.cn	1
12377.cn	1
12388.gov.cn	1
1688.com	1
17ok.com	1
189.cn	1
1c-bitrix.ru	1
22.cn	1
24,kg	1
24h.com.vn	1
2gis.ru	1
300.cn	1
360.cn	1
4.cn	1
6.cn	1
66law.cn	1
999.md	1
abc.com.py	1
abc-comunicação.dz	1
abril.com.br	1
accuweather. com	2
activemind.de	1
actuniger.com	1
anúncio.iq	1
admin.ch	1
adminbuy.cn	1
Adobe.com	1
afrikmag.com	1
agenciaentrate.gov.it	1
ah.gov.cn	1
ahoraeg.com	1
ahram.org.eg	1
aktuality.sk	1
alakhbar.info	1
aliexpress.com	1
alipay. com	1
allegro.pl	1
allevents.in	1
almasryalyoum.com	1
alraziuni.edu.ye	1
alwakeelnews. com	1
alwatanvoice.com	1
amazon.ae	1
amazon.ca	1
amazon.cn	1
amazon.co.jp	1
amazon.co.uk	1
amazon.com	20
amazon.com.br	1
amazon.de	1
amazon.eg	1
amazon.es	1
amazon.fr	1
amazon.in	1
amazon.it	1
ameblo.jp	1
amritmahotsav.nic.in	1
andersnoren.se	1
anpc.gov.ro	1
anyxxx. com	1
ap.gov.in	1
aparat.com	2
apple.com	5
arabiaweather.com	1
argentina.gob.ar	1
aruba.it	1
autohome.com.cn	1
avaz.ba	1
babytree. com	1
baharain.bh	1
baidu. com	4
banda.us	1
bangladesh.gov.bd	1

bankmandiri.co.id	1
bayern.de	1
bb.com.br	1
bbc.co.uk	1
bbc. com	1
bbc.in	1
bcel.com.la	1
beian.gov.cn	1
beijing.gov.cn	1
Bélgica.be	1
belizebank. com	1
belonnanotservice.ga	2
bet365.com	1
bgeneral.com	1
bih.nic.in	1
Bing.com	3
biobiochile.cl	1
bitrix24.ru	1
bjx.com.cn	1
blogger. com	3
bnonline.fi.cr	1
boc.cn	1
Bongacams. com	3
borneobulletin.com.bn	1
bri.co.id	1
britannica.com	2
bshare.cn	1
bt.bt	1
bt.cn	1
bukalapak. com	1
bund.de	1
dia útil.ng	1
businessinsider.in	1
businesstoday.in	1
businessworld.in	1
cac.gov.cn	1
cafebazaar.ir	1
caixa.gov.br	1
cambridge.org	1
Canva.com	5
cao.ir	1
carreiras.sl	1
cas.cn	1
cbec.gov.in	1
cbic.gov.in	1

cbos.gov.sd	1
cbse.gov.in	1
cbse.nic.in	1
ccdi.gov.cn	1
ccgp.gov.cn	1
ccm.gov.cn	1
ce.cn	1
centrafrique-presse.over-blog.com	1
perseguição.com	1
chaturbate. com	2
china.cn	1
china.com.cn	1
chinadaily.com.cn	1
chinanews.com.cn	1
chinatax.gov.cn	1
chsi.com.cn	1
cib.com.cn	1
cmbc.com.cn	1
cmseasy.cn	1
cnil.fr	1
cninfo.com.cn	1
cnipa.gov.cn	1
cnnindonesia.com	1
cnpq.br	1
cnr.cn	1
cntv.cn	1
coinmarketcap. com	2
comoros-infos.net	1
conac.cn	1
consultor.ru	1
coremail.cn	1
correios.com.br	1
corriere.it	1
cupa.ng	1
tribunal.gov.cn	1
covid19.go.id	1
cowin.gov.in	1
cpdp.bg	1
cq.gov.cn	1
creditchina.gov.cn	1
cri.cn	1
cricbuzz. com	1
cro.ma	1
csdn.net	1
csrc.gov.cn	1

alfândega.gov.cn	1
cvc.nic.in	1
cyberpolice.cn	1
dahe.cn	1
postagem diária.ng	1
dakaractu.com	1
daraz.pk	1
dados.gov.in	1
proteção de dados.gov.cy	1
daum.net	1
defimedia.info	1
detik.com	1
dg-datenschutz.de	1
dicionário.com	1
digikala.com	1
digitalindia.gov.in	1
dinesh-ghimire.com.np	1
discord.com	1
ditaduraeconsenso.blogspot.com	1
dlszywz.cn	1
dns4.cn	1
docdro.id	1
dpboss.net	1
dr.dk	1
draugiem.lv	1
duckduckgo.com	1
dwz.cn	1
e.gov.kw	1
ebay.com	1
ebay.de	1
ebs.org.cn	1
eci.gov.in	1
education.gov.in	1
elcomercio.com	1
eldeber.com.bo	1
elnuevodia.com	1
elpais.com	1
Elsalvador.com	1
eluniverso.com	1
emansion.gov.lr	1
emprego.cg	1
ems.com.cn	1
enamad.ir	1
enimerotiko.gr	1
eol.cn	1

e-recht24.de	1
ernet.in	1
espn. com	1
espncricinfo.com	1
estadao.com.br	1
eta.gov.lk	1
ethiojobs.net	1
etnet.com.hk	1
facebook.com	80
facebook.com.br	1
fandom. com	3
fazenda.gov.br	1
fijivillage. com	1
upload de arquivo.com	1
findlaw.cn	1
firefox.com.cn	1
fiverr. com	1
flipkart.com	1
flydubai. com	1
www.fmprc.gov.cn	1
foco.cn	1
Siga isso	1
Force.com	1
grátis.fr	1
freebitco.in	1
freeindianporn2.com	1
freepik. com	1
fs.fed.us	1
ftc.go.kr	1
www.fujian.gov.cn	1
gansu.gov.cn	1
garantirprivacidade.it	1
gd.gov.cn	1
gênio	1
gesetze-im-internet.de	1
ganaweb.com	1
gismeteo.ru	1
globo.com	1
gmw.cn	1
gogo.mn	1
gome.com.cn	1
goo.ne.jp	1
google.com	1
google.ad	1
google.ae	1

google.at	1
google.az	1
google.be	1
google.bf	1
Google BG	1
google.ca	2
google.cd	1
google.cg	1
google.ch	1
Pesquise no Google	1
google.cl	1
google.cn	1
google.co.id	1
google.co.il	1
google.co.in	1
google.co.jp	1
google.co.ke	1
google.co.kr	1
google.com.ma	1
google.co.mz	1
google.co.nz	1
google.co.th	1
google.co.tz	1
google.co.ug	1
google.co.uk	1
google.co.uz	1
google.co.ve	1
google.co.za	1
google.co.zm	1
google.co.zw	1
google.com	146
google.com.af	1
google.com.ar	1
google.com.bd	1
google.com.bn	1
google.com.bo	1
google.com.br	1
google.com.bz	1
google.com.co	1
google.com.cu	1
google.com.do	1
google.com.eg	1
google.com.hk	2
google.com.jm	1
google.com.kw	1

google.com.lb	1
google.com.ly	1
google.com.mm	1
google.com.mt	1
google.com.mx	1
google.com.na	1
google.com.ng	1
google.com.ni	1
google.com.np	1
google.com.om	1
google.com.pa	1
google.com.pe	1
google.com.pg	1
google.com.ph	1
google.com.pk	1
google.com.pr	1
google.com.py	1
google.com.qa	1
google.com.sa	1
google.com.sb	1
google.com.sg	1
google.com.sl	1
google.com.sv	1
google.com.tj	1
google.com.tr	1
google.com.tw	1
google.com.ua	1
google.com.uy	1
google.com.vn	1
google.de	1
google.dj	1
google.dk	1
Google DZ	1
google.ee	1
google.es	2
Google França	3
google.ge	1
google.gr	1
google.gy	1
google.hn	1
google.ht	1
google.ie	1
google.iq	1
google.is	1
google.it	1

google.jo	1
google.kg	1
google.kz	1
google.la	1
google.lk	1
google.lt	1
google.lu	1
google.lv	1
google.md	1
google.me	1
google.mg	1
google.mk	1
google.ml	1
google.mn	1
google.mw	2
google.nl	1
google.não	1
google.pl	1
google.ps	1
google.pt	1
google.ro	1
google.rs	1
Google.ru	3
google.rw	1
google.se	1
google.si	1
google.sk	1
google.sn	1
google.so	1
google.sr	1
google.st	1
google.td	1
google.tg	1
google.tl	1
google.tm	1
google.tn	1
google.tt	1
gosuslugi.ru	1
gov.bw	1
gov.ls	1
govtrack.us	1
grade.id	1
grupobancolombia.com	1
gst.gov.in	1
gsxt.gov.cn	1

guardião.co.tt	1
guardião.ng	1
gujarat.gov.in	1
gxzf.gov.cn	1
gz.gov.cn	1
haberler. com	1
hainan.gov.cn	1
haosou.com	1
hatena.ne.jp	1
hd315.gov.cn	1
hdfcbank. com	1
healthline.com	1
heartland.us	1
henan.gov.cn	1
herald.co.zw	1
oi.é	1
hindustantimes.com	1
homedepot. com	1
hoster.kz	1
hotlog.ru	1
hotpepper.jp	1
hotstar.com	2
huanqiu.com	1
hubei.gov.cn	1
hunan.gov.cn	1
hurriyet.com.tr	1
ibps.in	1
ibw.cn	1
www.icbc.com.cn	1
icicibank.com	1
aqui.nós	1
ico.org.uk	1
idnes.cz	1
iitb.ac.in	1
iitkgp.ac.in	1
ijavhd.com	1
imagenshack.us	1
imdb. com	2
imjo.in	1
in.gr	1
imposto de renda.gov.in	1
incometaxindia.gov.in	1
rendataxindiaefiling.gov.in	1
índice.hr	1
index.hu	1

Índia.com	1
Índia.gov.in	1
indiamart.com	1
www.indianrailways.gov.in	1
www.indianvisaonline.gov.in	1
indiapost.gov.in	1
indiatimes.com	1
indiatoday.in	1
inflibnet.ac.in	1
instagram.com	47
inestrutura.com	1
intoday.in	1
iol.co.za	1
ionos.de	1
iplt20.com	1
irctc.co.in	1
irembo.gov.rw	1
irna.ir	1
é.fi	1
isna.ir	1
itau.com.br	1
jamaica-gleaner.com	1
japanpost.jp	1
jc001.cn	1
Jd.com	1
jiangsu.gov.cn	1
jiangxi.gov.cn	1
jiji.ng	1
jl.gov.cn	1
jne.co.id	1
jotform.us	1
jrj.com.cn	1
jumia.ci	1
jumia.com.ng	1
juraforum.de	1
justindianporn.me	1
kancloud.cn	1
kar.nic.in	1
karnataka.gov.in	1
kaskus.co.id	1
kemdikbud.go.id	1
kemenag.go.id	1
kemkes.go.id	1
kenh14.vn	1
kerala.gov.in	1

khaberni. com	1
knet.cn	1
knetreg.cn	1
kominfo.go.id	1
kompas. com	1
kriesi.at	1
kuaishang.cn	1
kuenselonline.com	1
kumparan.com	1
kupujemprodajem.com	1
lanouvelletribune.info	1
laodong.vn	1
laprensa.com.ni	1
laprensa.hn	1
lawtime.cn	1
lazada.co.id	1
líder.ir	1
lefigaro.fr	1
legifrance.gouv.fr	1
legítimo.ng	1
limão.fr	1
lex.uz	1
licindia.in	1
linha.me	2
linkado.in	1
linkedin. com	13
liputan6.com	1
lista.am	1
listindiario.com	1
live.com	19
liveinternet.ru	1
livroreclamacoes.pt	1
lnkd.in	1
ltn.com.tw	1
ltt.ly	1
m.in	1
macaodaily.com	1
mahaonline.gov.in	1
maharashtra.gov.in	1
mail.ru	2
mana.pf	1
mastercard.us	1
mayoclinic.org	2
medcol.mw	1
mediacongo.net	1

mercadolibre.cl	1
mercadolibre.com.co	1
mercadolibre.com.ve	1
mercadoliivre.com.br	1
merdeka. com	1
merriam-webster. com	1
meskerem.net	1
clima.nc	1
metruyenchu.com	1
mhlw.go.jp	1
microsoft.com	25
microsoftonline.com	4
miliyet.com.tr	1
mk.by	1
mof.gov.tl	1
moh.go.tz	1
moip.gov.mm	1
mol.gov.om	1
monetizze.com.br	1
msn.com	3
myshopify. com	5
namibian.com.na	1
namnak.com	1
naver. com	1
ncdc.gov.ng	1
nessma.tv	1
netafrique.net	1
netflix.com	13
nethouse.ru	1
nettruyengo.com	1
news24.com	1
niagahoster.co.id	1
noção.so	1
novinky.cz	1
nsw.gov.au	1
nzherald.co.nz	1
odnoklassniki.ru	1
office.com	8
ok.ru	3
okezone. com	1
onlinehome.us	1
laranja.fr	1
orientar.tm	1
otr.tg	1
Ouest-France.fr	1

oxu.az	1
ozon.ru	1
pagcor.ph	1
páginasjaunes.fr	1
paypal.com	1
pilha de pagamento.com	1
pikiran-rakyat.com	1
pinterest.com	11
pinterest.de	1
pinterest.es	1
pinterest.fr	1
pinterest.it	1
pixnet.net	1
planalto.gov.br	1
pornhub.com	4
portaldoconhecimento.gov.cv	1
postar.ir	1
postcourier.com.pg	1
postimees.ee	1
premierbet.co.ao	1
premierleague.com	1
prensa-latina.cu	1
prensalibre.com	1
presidência.gov.bi	1
presidente.ir	1
presidente.tj	1
baile.st	1
baile.ua	1
público.lu	1
pulso.ng	1
punchng. com	1
qq.com	2
r01.ru	1
rae.es	1
rakuten.co.jp	1
rambler.ru	1
reddit. com	9
reg.ru	1
republica.it	1
republica.co.id	1
ria.ru	1
www.rijksoverheid.nl	1
rt.com	1
rte.ie	1
rtvslo.si	1

s.id	1
sabay.com.kh	1
sacoronavirus.co.za	1
sahibinden. com	1
sakura.ne.jp	1
salesforce.com	1
salla.sa	1
sana.sy	1
saúde.gov.dz	1
saúde.gov.gn	1
sapo.pt	1
sapp.ir	1
saude.gov.br	1
Sielo.br	1
sekolahku.web.id	1
seneweb.com	1
serveriai.lt	1
service-public.fr	1
setn.com	1
seznam.cz	1
shopee.co.id	1
shopee.co.th	1
shopee.tw	1
shopee.vn	1
shop-pro.jp	1
singaporepools.com.sg	1
smarturl.it	1
sohu.com	2
solomonstarnews. com	1
soy502.com	1
Spiegel.de	1
stackoverflow.com	1
standardmedia.co.ke	1
estado.co.us	1
estado.fl.us	1
estado.il.us	1
estado.ma.us	1
estado.md.us	1
estado.mn.us	1
estado.nj.us	1
estado.nm.us	1
estado.nv.us	1
estado.ny.us	1
estado.oh.us	1
estado.ou.us	1

estado.pa.us	1
estado.tx.us	1
suara.com	1
sucursalelectronica.com	1
suribet.sr	1
sympala.com.br	1
syri.net	1
t.me	2
taobao. com	2
theguardian.com	1
thethao247.vn	1
tiktok. com	10
tempo.mk	1
times.co.sz	1
timesofmalta.com	1
timeweb.ru	1
tmall.com	1
tokopedia.com	1
t-online.de	1
tradingview.com	2
trendyol.com	1
tribunnews. com	1
tripadvisor.com.br	1
tripadvisor.fr	1
tripadvisor.it	1
turkiye.gov.tr	1
twitch.tv	5
twitter.com	32
ucoz.ru	1
uem.mz	1
ultimahora.com	1
uol.com.br	1
ura.go.ug	1
usp.br	1
vanguardngr. com	1
vg.não	1
vk. com	7
vkontakte.ru	1
vnexpress.net	1
walmart.com	1
wbs-law.de	1
clima.com	1
webmd.com	1
whatsapp.com	22
wikipedia.org	29

www.wikitionary.org	2
bagas silvestres.ru	1
assistente.id	1
www.gob.mx	1
www.gob.pe	1
www.gov.br	1
www.gov.pl	1
www.gov.uk	1
xhamster. com	1
xnxx.com	3
xosodaiphat.com	1
xvideos.com	6
yahoo.co.jp	1

yahoo.com	25
yandex.ru	5
yasour.org	1
yelp.com	1
ynet.co.il	1
youm7.com	1
YouTube	1
youtube. com	103
youtube.org	1
zalo.me	1
zambiaimmigration.gov.zm	1
zhizhuchi.cm	1
zoom.us	15

ANEXO 6: MACROLINGUAS

Conforme definido pelo Etnólogo.

Tablela 11: Lista de macrolínguas

CÓDIGO ISO	***MACRO LÍNGUAS***	***NÚMERO DE LINGUAS FUNDIDAS***
ara	**Arabe**	***29***
aym	**Aimara**	***2***
aze	**Azerbaijano**	***3***
bal	**Baluchi**	***3***
bik	**Bikol**	***8***
bnc	**Bontok**	***5***
bua	**Buriate**	***3***
chm	**Marido**	***2***
cre	**Gritar**	***6***
del	**Delaware**	***2***
den	**Eslavo (athapaskan)**	***2***
din	**Dinka**	***5***
doi	**Dogri**	***2***
est	**Estoniano**	***2***
fas	**Persa**	***2***
ful	**Fulfulde**	***9***
gba	**Gbaya**	***6***
gon	**Gondi**	***3***
grb	**Grebo**	***5***
grn	**Guarani**	***5***
hai	**Haida**	***2***
hbs	**Servo-croata**	***4***
hmn	**Hmong**	***25***
iku	**Inuktitut**	***2***
ipk	**Inupiatun**	***2***
jrb	**Judaico-arabe**	***5***
kau	**Kanuri**	***3***
kln	**Kalenjin**	***9***
kok	**Concani**	***2***
kom	**Komis**	***2***
kon	**Congo**	***3***
kpe	**Kpell**	**2**
kur	**Curdo**	***3***
lah	**Lahnda**	***7***
lav	**Letão**	***2***
luy	**Luia**	***14***
man	**Mandingo**	***6***
mlg	**Malgaxe**	***11***
mon	**Mongol**	***3***
msa	**Malaio**	***36***
mwr	**Marwari**	***6***
nep	**Nepales**	***2***
oji	**Ojibua**	***7***
ori	**Oria**	***2***
orm	**Gala**	***4***
pus	**Pashto**	***3***
que	**Quechua**	***42***
raj	**Rajastão**	**6**
rom	**Cigano**	***6***
sqi	**Albanes**	***4***
srd	**Sardenha**	***4***
swa	**Suaili**	***2***
syr	**Siriaco**	***2***
tmh	**Tamashek**	***4***
uzb	**Usbeque**	***2***
yid	**Iidiche**	***2***
zap	**Zapoteca**	***57***
zha	**Zhuang**	***16***
zho	**Chines**	***15***
zza	**Dimli**	***2***

ANEXO 7: LISTA DE PAÍSES OU TERRITÓRIOS SEM DADOS DA UIT

Tablela 12: Lista de países sem dados da UIT

Código ISO	NOME DO PAÍS	POPULAÇÃO
AX	**Ilha de Aland**	**27.652**
AS	**Samoa Americana**	**55.990**
IO	**Território Britânico do Oceano Índico**	**4.000**
QB	**Holanda caribenha**	**18.740**
CX	**Ilha do Natal**	**1.170**
CC	**Ilhas Cocos (Keeling)**	**630**
CK	**Ilhas Cook**	**15.000**
CW	**Curaçao**	**140.000**
GF	**Guiana Francesa**	**366.590**
GP	**Guadalupe**	**454.800**
GU	**Guam**	**139.550**
IM	**Ilha do homem**	**88.085**
QM	**Martinica**	**377 100**
NC	**Ilha Norfolk**	**1.500**
KP	***Coréia do Norte***	***25.579.000***
PM	**Ilhas Marianas do Norte**	**53.280**
PW	**Palau**	**17.550**
PN	**Pitcairn**	**36**
RE	**Reunião**	**751 580**
BL	**São Bartolomeu**	**7.850**
FM	**São Martinho**	**28.500**
PM	**São Pedro e Miquelon**	**6.340**
SX	**São Martinho**	**33.470**
CT	**Ilhas Turcas e Caicos**	**30 170**
GO	***Estado do Vaticano***	***330***
HE	***Saara Ocidental***	***544 150***
	TOTAL	**28.689.463**

Existem duas razões possíveis pelas quais o país ou território é excluído dos dados da UIT:

1) É um território cujos dados estão incluídos nos de outro país
2) Não há fonte ou estimativa para a percentagem de pessoas ligadas à Internet (em itálico na tabela).

ANEXO 8: FONTES SOBRE O COMPORTAMENTO LINGUÍSTICO DOS USUÁRIOS DA INTERNET

https://motsdici.be/wp-content/uploads/2019/04/Article-cant-read-wont-buy.pdf
Relatório Consultivo do Common Sense 2006 “Se não sei ler, não compro”.

https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556
Relatório de Inquérito da União Europeia de 2011 "Agenda Digital: Mais de metade dos utilizadores da Internet na UE utilizam uma língua estrangeira quando estão online"
Citação: "Embora 90% dos utilizadores da Internet na UE prefiram aceder a sítios Web na sua própria língua, 55% utilizam, pelo menos ocasionalmente, uma língua diferente da sua quando estão online, de acordo com um Eurobarómetro pan-europeu".

https://hbr.org/2012/08/speak-to-global-customers-in-t
Harvard Business Review 2012: “Converse com clientes internacionais em seu próprio língua”
Citação: "72,1% dos consumidores passam a maior parte ou todo o seu tempo em websites na sua própria língua"

https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-Languages-Defining-Indias-Internet.pdf
Estudo KPMG/Google 2017 "Línguas Indianas - Definindo a Internet Indiana"
Citação: “Espera-se que os usuários da Internet em línguas indianas representem quase 75% da base de usuários da Internet na Índia em 2021.”

https://insights.csa-research.com/reportaction/305013126/Marketing
https://csa-research.com/Blogs-Events/CSA-in-the-Media/Press-Releases/Consumers-Prefer-their-Own-Language
CSA Research Report 2020 "Se não sei ler, não comprarei - análise B2C das preferências linguísticas e comportamentos do consumidor em 29 países"
Citação: “Uma pesquisa com 8.709 consumidores em 29 países revela que 76% preferem comprar produtos com informações em seu próprio língua.”

https://octopustranslations.com/e-commerce-and-the-impact-of-language-on-consumer-behavior/
Octopus Translation Report 2021 E-commerce e o impacto da língua no comportamento do consumidor
Citação: “55% dos consumidores em todo o mundo compram online apenas em seu língua nativo.”

https://www.businesswire.com/news/home/20211026005375/en/Unbabel%E2%80%99s-2021-Global-Multilíngual-CX-Survey-Reveals-68-of-Consumers-Prefer-to-Speak- com-as-marcas-em-seu-língua-nativo
Relatório BusinessWire 2021 "A pesquisa Multilíngual Global CX 2021 da Unbabel revela que 68% dos consumidores preferem falar com as marcas em sua língua nativa"

https://www.prweb.com/releases/2014/04/prweb11725995.htm
2022.PRWeb Market Research “Uma pesquisa com 3.000 compradores on-line em 10 países revela que 60% raramente ou nunca compram em sites somente em inglês.”

ANEXO 9: RESULTADOS SEPARADOS PARA L1 E L2

Como método de verificação cruzada da validade do modelo, que se baseia em dados demolinguísticos L1+L2, foram realizadas duas execuções adicionais, uma apenas com dados L1 e outra apenas com dados L2.

Tablela 13: Modelo executado apenas com L1

		USUÁRIOS DE INTERNET	POPULAÇÃO	CAIXAS DE SOM	CONTEÚDO	PRESENÇA	PRODUTIVO.
		L1	L1	L1	L1	VIRTUAL	CONTEÚDO
1	chinês	22,34%	18,33%	71,18%	**25,55%**	**1,39**	**1.14**
2	Inglês	7,82%	5,12%	89,24%	**12,96%**	**2,53**	**1,66**
3	Espanhol	8,14%	6,52%	72,95%	**8,76%**	**1,34**	**1.08**
4	árabe	5,33%	4,80%	64,91%	**4,15%**	**0,86**	**0,78**
5	Português	3,91%	3,21%	70,99%	**3,91%**	**1.22**	**1,00**
6	japonês	2,77%	1,75%	92,63%	**3,47%**	**1,99**	**1,25**
7	russo	3,00%	2,13%	82,36%	**3,22%**	**1,51**	**1.07**
8	hindi	3,35%	4,73%	41,34%	**2,93%**	**0,62**	**0,88**
9	Francês	1,59%	1,10%	84,59%	**2,08%**	**1,89**	**1.31**
10	Alemão	1,62%	1,06%	89,51%	**1,96%**	**1,85**	**1.21**

Se considerarmos apenas os falantes de primeira língua, o francês estaria na posição 9 e, logicamente, o chinês apresentaria uma grande vantagem sobre o inglês, apesar da sua presença virtual muito grande e da produtividade do seu conteúdo. A presença virtual e a produtividade de conteúdo para o francês são muito altas, apesar deste nono lugar.

Tablela 14: Modelo executado apenas com L2

		USUÁRIOS DE INTERNET	POPULAÇÃO	CAIXAS DE SOM	CONTEÚDO	PRESENÇA	PRODUTIVO.
		L2	L2	L2	L2	VIRTUAL	CONTEÚDO
1	Inglês	32,53%	31,25%	55,64%	**37,91%**	**1.21**	**1.17**
2	chinês	8,68%	6,38%	72,65%	**10,68%**	**1,67**	**1.23**
3	Francês	6,47%	5,99%	57,81%	**6,90%**	**1,15**	**1.07**
4	hindi	6,32%	8,25%	40,93%	**5,96%**	**0,72**	**0,94**
5	Espanhol	3,37%	2,28%	78,82%	**5,47%**	**2,39**	**1,62**
6	russo	4,82%	3,33%	77,32%	**5,12%**	**1,54**	**1.06**
7	malaio	5,37%	5,21%	55,08%	**4,52%**	**0,87**	**0,84**
8	Alemão	3,10%	1,87%	88,72%	**3,61%**	**1,93**	**1.17**
9	tailandês	1,86%	1,28%	77,84%	**1,55%**	**1.21**	**0,83**
10	urdu	1,81%	5,15%	18,86%	**1,15%**	**0,22**	**0,63**
11	Português	0,68%	0,81%	44,81%	**0,89%**	**1.10**	**1,32**

Se considerarmos apenas os falantes de uma segunda língua, o inglês ocupa logicamente o primeiro lugar e o francês o terceiro, à frente do espanhol.

Como lembrete, aqui estão os resultados para L1+L2.

Tablela 15: Resultados do modelo para L1+L2

		USUÁRIOS DE INTERNET	POPULAÇÃO	CAIXAS DE SOM	CONTEÚDO	PRESENÇA	PRODUTIVO.
		L1+L2	L1+L2	L1+L2	L1+L2	VIRTUAL	CONTEÚDO
1	chinês	18,46%	14,72%	71,38%	**21,60%**	**1,47**	**1.17**
2	Inglês	14,83%	13,01%	64,86%	**19,60%**	**1,51**	**1,32**
3	Espanhol	6,79%	5,24%	73,72%	**7,85%**	**1,50**	**1.16**
4	hindi	4,19%	5,80%	41,16%	**3,76%**	**0,65**	**0,90**
5	russo	3,51%	2,49%	80,32%	**3,76%**	**1,51**	**1.07**
6	Francês	2,98%	2,58%	65,80%	**3,33%**	**1,29**	**1.12**
7	Português	2,99%	2,49%	68,43%	**3,13%**	**1,26**	**1.05**
8	árabe	3,97%	3,53%	63,99%	**3,09%**	**0,87**	**0,78**
9	japonês	1,99%	1,22%	92,63%	**2,66%**	**2.18**	**1,34**
10	Alemão	2,04%	1,30%	89,17%	**2,37%**	**1,82**	**1.16**

É realizada uma verificação de consistência entre os 3 resultados, devendo o terceiro seguir logicamente os dois primeiros.

Tablela 16: Controle dos resultados L1 e L2

	POP. Mundialmente	POP. Conectado	% pop. Conectar	Pop. Inglês	Pop. Conectar. Inglês	% pop. Conectar. Inglês	Ao controle
L1	**7.231.699.136**	**4.223.428.027**	**58,40%**	5,12%	7,82%	89,24%	89,24%
L2	**3.130.017.620**	**1.673 121.762**	**53,45%**	31,25%	32,53%	55,64%	55,64%
L1+L2	**10.361.716.756**	**5.896.549.789**	**56,91%**	13,01%	14,83%	64,86%	64,86%
Ao controle			56,91%	13,01%	14,83%	64,86%	

Em verde são realizadas as verificações: trata-se de calcular diretamente os mesmos valores e, portanto, verificar se os dois modelos L1 e L2 funcionaram corretamente: a prova está feita.

O segundo conjunto de verificações é mais complexo e não devem ser esperadas correspondências perfeitas (já que a modelagem não é um processo linear em relação aos dados demolinguísticos).

Tablela 17: Verificação dos resultados L1 e L2 (continuação)

	Inglês	chinês	Espanhol	Francês	hindi	Português	russo	Alemão
Conteúdo L1	12,96%	25,55%	8,76%	2,08%	2,93%	3,91%	3,22%	1,96%
Conteúdo L2	37,91%	10,68%	5,47%	6,90%	5,96%	0,89%	5,12%	3,61%
Conteúdo L1 + L2	19,60%	21,60%	7,85%	3,33%	3,76%	3,13%	3,76%	2,37%
Ao controle	20,04%	21,33%	7,83%	3,45%	3,79%	3,05%	3,76%	2,43%

As três primeiras linhas mostram os resultados dos três respectivos modelos. A linha de controle verde é calculada ponderando as respectivas porcentagens L1 e L2 em relação às respectivas populações conectadas. Assim, para o inglês, 20,04% é obtido pela seguinte fórmula: ((12,96 x 4 233 428 027) + (37,91 x 1 673 121 762)) / 5 896 549 789

É ao mesmo tempo notável e muito tranquilizador, no que diz respeito à validade do modelo, que os resultados obtidos pelos dois métodos (o modelo L1+L2 ou o pro rata dos resultados dos modelos L1 e L2 em relação às respectivas populações conectadas) são tão parentes.