**Uma abordagem alternativa para a produção de indicadores da presença de línguas na Internet**

**Daniel Pimienta**

**pimienta@funredes.org**

**Observatório de línguas e culturas na Internet**

**http://funredes.org/lc**

**Rede Global para Diversidade Linguística**

**http://maaya.org**

**Dezembro de 2020**

**(tradução de um articulo de 2017)**

**RESUMO**

Dada a dificuldade de obtenção de dados confiáveis ​​sobre a presença de línguas na Internet, é apresentada uma abordagem alternativa para produzir uma série de indicadores para as 140 línguas com mais de 5 milhões de falantes. O método é baseado na compilação de uma série de micro-indicadores que medem a presença de línguas ou países em diferentes espaços ou aplicações da Internet. Um cálculo de ponderação transforma porcentagens mundiais por país em porcentagens mundiais por língua. São definidos cinco indicadores relacionados à presença de línguas na Internet: usuários, tráfego, uso, conteúdo, índices sociais e interfaces. A partir daí, são deduzidos 4 macro-indicadores, para cada língua tratado: potência, capacidade, gradiente e produtividade do conteúdo. Os valores dos indicadores são calculados processando um conjunto de 369 micro-indicadores e os dos macro-indicadores são obtidos com operações simples a partir dos indicadores. O artigo lembra que importantes vieses marcaram a história da cibermetria linguística. Todos os possíveis vieses derivados do método, suas hipóteses explícitas ou implícitas e suas fontes são analisados ​​e, finalmente, é proposta uma estimativa que leva esses vieses em consideração. As comparações críticas e análises dos respectivos vieses são feitas com os poucos indicadores existentes (W3Techs e InternetWordStats). Conclui com recomendações para que a cibermetria linguística retorne onde deve permanecer: no mundo acadêmico e de pesquisa. Os créditos pertencem à Organização Internacional da Francofonia (OIF),

Palavras-chave: linguagens, Internet, indicadores, cibermetria linguística, vieses

Conteúdo

[1. Introdução 5](#_Toc58250152)

[2. Metodologia 6](#_Toc58250153)

[2.1 Indicadores da língua na Internet 6](#_Toc58250154)

[2.2 Micro-indicadores por língua ou país 8](#_Toc58250155)

[2.2.1 Por língua 8](#_Toc58250156)

[2.2.2 Por país 9](#_Toc58250157)

[2.3 Fontes de informação 9](#_Toc58250158)

[2.4.1 O caso da língua materna (L1) 10](#_Toc58250159)

[2.4.2 O caso do segundo língua (L2) 11](#_Toc58250160)

[2.5 Processo 12](#_Toc58250161)

[2.5.1 Extrapolação 13](#_Toc58250162)

[2.5.2 Processo para indicadores 14](#_Toc58250163)

[2.5.3 Processo para alcançar resultados diferenciados por assunto 17](#_Toc58250164)

[2.5.4 Tipos de ponderação usados 17](#_Toc58250165)

[3. Resultados 18](#_Toc58250166)

[4. Análise dos resultados 20](#_Toc58250167)

[4.1 Sobre a Wikipedia 20](#_Toc58250168)

[4.2 Comparação dos resultados com os do InternetWorldStats (IWS) 21](#_Toc58250169)

[4.3 Sensibilidade dos fatores 23](#_Toc58250170)

[5. Limitações metodológicas, análise de viés e controles realizados 24](#_Toc58250171)

[5.1 Cuidado com os viés! 24](#_Toc58250172)

[5.1 Limitações e vieses do método 25](#_Toc58250173)

[5.2 Línguas 26](#_Toc58250174)

[5.2.1 Selecionando a fonte para o cálculo de L1 26](#_Toc58250175)

[5.2.2 O caso de L2 26](#_Toc58250176)

[5.2.3 Reduzindo o número de línguas 28](#_Toc58250177)

[5.2.4 Verificando a invariância dos resultados com o número de línguas 28](#_Toc58250178)

[5.3 Países 28](#_Toc58250179)

[5.4 Fontes 29](#_Toc58250180)

[5.4.1 Conceitos básicos 29](#_Toc58250181)

[5.4.2 Exceções aos princípios básicos 29](#_Toc58250182)

[5.4.3 A questão das datas 30](#_Toc58250183)

[5.4.4 A questão do significado do transformação país > língua 30](#_Toc58250184)

[5.4.5 Limitações devido a fontes 31](#_Toc58250185)

[5.4.6 Potencial viés Alexa.com e W3Techs 32](#_Toc58250186)

[5.4.7 Correção de vieses W3Techs 36](#_Toc58250187)

[5.4.8 Limitações / vieses relacionados ao grau de localidade das fontes 37](#_Toc58250188)

[5.4.9 Sobre o princípio de ponderação 39](#_Toc58250189)

[6. Conclusões e perspectivas 40](#_Toc58250190)

[7. Bibliografia 41](#_Toc58250191)

[Anexo I. Lista de micro-indicadores 43](#_Toc58250192)

[Anexo II. Fontes selecionadas 45](#_Toc58250193)

[Anexo III: Valores escolhidos para L2 46](#_Toc58250194)

[Anexo IV: Sites totalmente locais 48](#_Toc58250195)

**LISTA DE TABELAS E FIGURAS**

**TABELAS**

[Tabela 1 : Descrição dos indicadores 15](#_Toc58251622)

[Tabela 2 :Descrição dos macro-indicadores 16](#_Toc58251623)

[Tabela 3: Os 3 tipos de ponderação usados 18](#_Toc58251624)

[Tabela 4: Indicadores para os 15 principais línguas em termos de poder 18](#_Toc58251625)

[Tabela 5 :Línguas ordenados por capacidade 19](#_Toc58251626)

[Tabela 6 : Línguas ordenados por porcentagem de pessoas conectadas 19](#_Toc58251627)

[Tabela 7 : Línguas ordenados por gradiente 20](#_Toc58251628)

[Tabela 8: Artigos da Wikipedia vs. presença na internet 20](#_Toc58251629)

[Tabela 9: Diferenças de hipóteses com IWS para L1 + L2 21](#_Toc58251630)

[Tabela 10: Dados da InternetWorldStats (junho de 2016) 22](#_Toc58251631)

[Tabela 11: Comparação com IWS após redução para 100% 22](#_Toc58251632)

[Tabela 12: Simulação com dados IWS 23](#_Toc58251633)

[Tabela 13: Distribuição de línguas na Web de acordo com Inktomi (2000): 24](#_Toc58251634)

[Tabela 14 : Países: afiliação de territórios 29](#_Toc58251635)

[Tabela 15: Relações tráfego / assinante 33](#_Toc58251636)

[Tabela 16: Classificação especulativa 37](#_Toc58251637)

[Tabela 17: Distribuição de sites com forte localidade 38](#_Toc58251638)

[Tabela 18: Simulação para interfaces 40](#_Toc58251639)

[Tabela 19: Simulação para índice 40](#_Toc58251640)

**FIGURAS**

[FIGURA 1: Indicadores da presença de línguas na Internet 8](#_Toc22209582)

# Introdução

No período 1998-2007, o autor e Daniel Prado colaboraram, a partir de suas respectivas instituições, a Fundação de Redes e Desenvolvimento (FUNREDES) e a União Latina, no desenho de métodos de medição de linguagens na Internet que pudessem fornecer indicadores confiáveis. e reproduzível (ver [6]). Ao mesmo tempo, outras iniciativas[[1]](#footnote-1) existiam para os mesmos fins. Desde 2007, mudanças profundas no comportamento dos motores de busca na Web, bem como no tamanho do espaço web indexado, tornaram os métodos obsoletos e isso criou um vácuo na produção de indicadores de linguagem na Internet[[2]](#footnote-2). Entre 2010 e 2012, sob a direção do autor, FUNREDES e MAAYA propuseram lançar um ambicioso projeto de pesquisa com o objetivo de preencher esta lacuna. UNESCO, OIF e União Latina aderiram a este projeto e um consórcio de prestigiosas instituições de pesquisa europeias foi formado para tentar obter financiamento do Programa-Quadro de Pesquisa da União Europeia (projeto DILINET[[3]](#footnote-3)) No entanto, este esforço não teve sucesso na obtenção de financiamento, apesar da persistência e qualidade das equipas de investigação participantes (duas tentativas foram feitas em 2012 e 2013, em resposta aos apelos para uma proposta europeia[[4]](#footnote-4) e uma última tentativa com parceiros de aborto do Catar em 2014).

Para preencher esse vazio, um novo método menos ambicioso, mais pragmático e muito mais barato foi proposto por Daniel Prado, em 2012, e abriu uma nova colaboração com o autor, sob o chapéu institucional do MAAYA e com o apoio da OIF. O método foi desenvolvido com base na observação do comportamento da língua francesa nos mais diversos espaços e aplicações da Internet. Dois primeiros estudos ajudaram a fornecer resultados em termos de classificação do francês na Internet. O segundo, realizado em 2013, integrou o capítulo na Internet do relatório "Le français dans le monde" de 2014 (ver [3]) e foi seguido por um estudo semelhante de espanhol na Internet (ver [10]). O último estudo da OIF, mais ambicioso, que inspira este artigo, conseguiu a aplicação de uma abordagem estatística, autorizado pelo aumento do número de fontes, o que permite obter resultados em termos de indicadores para um grande grupo de línguas, além do francês.

O método baseia-se na compilação de informações quantitativas sobre o uso de linguagens em diversas aplicações e espaços da Internet, constituindo uma série de microindicadores. A compilação e organização dos dados das fontes permite medir a presença das línguas na Internet e colocar os resultados em perspectiva para a construção de uma série de indicadores que medem a respetiva quota de cada língua na Internet, em termos dos internautas, o tráfego gerado, a utilização de serviços, conteúdos, interfaces de software e serviços de tradução no língua em questão e com base em uma série de índices que avaliam critérios relacionados à sociedade da informação. Uma síntese da presença das línguas na Internet é produzida na forma de uma série de macro-indicadores, que fazem a síntese de todos os indicadores linguísticos:

* *Poder* da linguagem na Internet,
* *Capacidade* de seus falantes,
* Dinamismo dos falantes conectados, chamado *gradiente*
* *Produtividade* de seus falantes conectados em termos de produção de conteúdo.

Agrupamentos temáticos de microindicadores podem diferenciar ainda mais o potencial de uma determinada língua em relação a esses tópicos e, assim, lançar alguma luz sobre seus pontos fortes e fracos na Internet[[5]](#footnote-5).

O referencial metodológico estabelecido é utilizar o maior número de fontes disponíveis para quantificar o papel das línguas na Internet, seja diretamente quando estão disponíveis figuras relacionadas às línguas, o que infelizmente é raro, ou indiretamente, utilizando figuras por país e transformá-los em figuras por linguagem com um método que é um pilar do modelo estabelecido[[6]](#footnote-6).

Essa transformação de figuras relacionadas aos países em figuras relacionadas às línguas é um elemento original e inédito e que abre a possibilidade de tratar a questão das línguas na Internet, em um contexto onde os dados relativos às línguas na Internet são extremamente escassos e de confiabilidade duvidosa (a Wikimedia é uma exceção admirável).

Esta abordagem é suportada por hipóteses implícitas que devem ser explicitadas e avaliadas, portanto, alguns cuidados devem ser tomados para garantir a consistência e confiabilidade dos resultados. A discussão sobre os limites do método e os controles que têm sido realizados para garantir a confiabilidade de um método que implica certa complexidade, tanto nos cálculos realizados quanto na compreensão dos conceitos resultantes, portanto, ocupa uma parte significativa de este artigo.

# Metodologia

## 2.1 Indicadores da língua na Internet

Espaços (por exemplo, smartphones ou governo eletrônico) e aplicativos (por exemplo, o mecanismo de busca do Google ou a rede social Facebook) da Internet serão explorados em busca de fontes de dados. Quando fontes quantitativas confiáveis ​​podem ser identificadas, os micro-indicadores correspondentes são definidos. Métodos estatísticos (diferentes tipos de médias ponderadas, médias simples ou truncadas) são então implantados para construir indicadores com a melhor credibilidade. Seis indicadores são identificados que medem a participação mundial**[[7]](#footnote-7)** de cada língua de acordo com vários elementos característicos da Internet:

* **Usuários:** (pessoas com acesso à Internet) que está relacionado à cota de falantes de cada língua que têm a possibilidade de se conectar à Internet. Um único micro-indicador (oferecido pela UIT) fornece as respostas necessárias e servirá como fonte fundamental para o método estabelecido.
* **Usos**: Refere-se a assinaturas de aplicativos por meio da conexão com a Internet. Onze microindicadores estão envolvidos na construção deste indicador.
* **Tráfego**: Indicação do tráfego gerado pelos usuários para os aplicativos. Trezentos e dezesseis microindicadores são usados ​​para construir este indicador.
* **Índices**: Refere-se à classificação dos países em vários aspectos da sociedade da informação. Cinco microindicadores são usados ​​atualmente para construir este indicador.
* **Conteúdo**: Refere-se ao conteúdo da Web para cada língua (no momento, ele coleta principalmente dados da galáxia da Wikimedia). Treze microindicadores fornecem dados para este indicador.
* **Interfaces e tradução**: refere-se à presença de línguas em interfaces de aplicativos ou como uma língua de tradução. Vinte e três micro-indicadores alimentam este indicador.

Quatro macro-indicadores da presença de línguas na Internet expressam a síntese dos indicadores anteriores:

* um macro-indicador do *poder* da linguagem na Internet, que mede a participação mundial da linguagem na Internet, uma média dos seis indicadores anteriores;
* um macro-indicador da *capacidade* da língua na Internet que é calculado como a divisão de poder com a porcentagem da população mundial nessa língua;
* um macro-indicador do *gradiente* de língua na Internet que é calculado como a divisão de poder pela porcentagem de pessoas conectadas à Internet nessa língua.
* um macro-indicador de *produtividade* em termos de criação de conteúdo, calculado como a divisão da porcentagem de conteúdo naquela língua pela porcentagem de usuários da Internet nessa mesma língua.

O indicador de poder é expresso como uma cota participação mundial. Os outros três indicadores são adimensionais e normalizados para 1. O conceito descrito é tão alto quanto o valor é maior que 1 e vice-versa.

O processamento estatístico basear-se-á principalmente na única fonte considerada fiável e essencial, disponibilizada pela UIT: a percentagem de pessoas ligadas à Internet por país. A leitura do artigo mostra que essa fonte está envolvida em um grande número de operações a serem realizadas, em particular a ponderação.

O diagrama a seguir mostra todos os indicadores que são processados ​​para cada língua.

FIGURA 1 : Indicadores da presença de línguas na Internet



Todos os micro-indicadores são apresentados no Anexo 1: Lista de micro-indicadores. Os indicadores e macro-indicadores são detalhados em tabela na seção 4.5.3 Tratamentos para resultados.

## 2.2 Micro-indicadores por língua ou país

Um total de 369 microindicadores representam a fonte de dados para esta edição do estudo. Alguns (36) informam diretamente sobre as línguas na Internet, os demais (333) oferecem dados por país.

### 2.2.1 Por língua

Os micro-indicadores por língua referem-se aos conteúdos (13) e interfaces (23) e são 36 no total para esta edição.

As fontes de conteúdo são expressas em termos de unidades por língua (por exemplo, número de artigos da Wikipédia por língua) ou porcentagem mundial (por exemplo, a porcentagem mundial de livros vendidos pela Amazon[[8]](#footnote-8)por língua). No primeiro caso, o valor é convertido para uma porcentagem mundial dividindo pelo total. O indicador de conteúdo é calculado como a média truncada em 20%[[9]](#footnote-9) dos 13 micro-indicadores de conteúdo: número de livros por língua na Amazon, dados da W3Techs[[10]](#footnote-10) e 11 micro-indicadores produzidos pela Wikimedia).

Os micro-indicadores de interface e tradução são medidos por um número binário que expressa a ausência (0) ou existência (1) da língua na interface do aplicativo ou a presença ou ausência da língua no aplicativo de tradução. Existem 23 aplicativos analisados. O indicador de interface mostra inicialmente o percentual de presença da língua em todos os aplicativos, depois esse valor é transformado em percentual mundial pelo método de ponderação com o percentual de internautas em cada língua.

### 2.2.2 Por país

Mais de 90% dos microindicadores desta edição do estudo foram obtidos de fontes que fornecem informações por país, que podem oferecer dados de diferentes maneiras:

* em quantidades (por exemplo, o número de usuários da Internet para cada país);
* em percentagens nacionais (por exemplo, a percentagem de visitas ao Facebook por país, ou a percentagem de tráfego de Internet móvel feita em cada país);
* em porcentagens em todo o mundo (por exemplo, a distribuição de tráfego mundial para o site Facebook.com por país);
* na forma de notação em uma escala predeterminada (no caso de índices que fornecem classificações para países de acordo com critérios, como Web Index[[11]](#footnote-11) disponibilização de índices de países da sociedade da informação com valores de 0 a 1 ou de 0 a 100, consoante o caso);

Para conseguir um processo homogêneo, todos os micro-indicadores de várias origens são transformados em uma porcentagem mundial dentro do modelo de trabalho.

A transformação dos dados por país em dados por língua é feita ponderando os dados por país com o número de falantes de cada língua em cada país.

As fontes podem fornecer valores para o presente (2016), ou mais frequentemente para uma data anterior, às vezes até projeções para o futuro.

As fontes de dados disponíveis raramente cobrem todos os países do mundo[[12]](#footnote-12) e muitas vezes apenas uma minoria. Portanto, a transformação de "dados por país" em "dados por língua" pode exigir algumas técnicas de extrapolação para complementar razoavelmente, a partir dos dados existentes na fonte, os dados dos países não relatados (ver capítulo 4.5.1 Extrapolação)

## 2.3 Fontes de informação

Para reunir a maior quantidade de dados, dezenas de diferentes fontes de informação (estatísticas, pesquisas, avaliações, índices, bases de dados, etc.) foram identificadas e avaliadas para cada fonte quanto à sua relevância, confiabilidade e consistência com o método. Algumas fontes que fornecem resultados para poucos países foram rejeitadas e outras em que as porcentagens por país perdem significado quando expressas pela língua também foram excluídas.

É relativamente fácil encontrar estatísticas por país sobre acesso à Internet e novas tecnologias (graças à UIT, às Nações Unidas ou ao Banco Mundial). No entanto, é muito mais difícil encontrar fontes que forneçam dados por língua, fora do universo da Wikimedia e da esfera do mundo aberto[[13]](#footnote-13). Ambos se caracterizam por uma genuína preocupação com a transparência, tanto dos dados como das metodologias utilizadas para a sua produção.

Empresas online, proprietários de mídias sociais, motores de busca ou aplicativos reconhecidos de Internet geralmente são bastante sigilosos sobre suas estatísticas. Os únicos que podem fornecer números em aplicativos (com forte foco em mídia social e tudo o que tem a ver com marketing no mundo digital) são as empresas de marketing na internet. Na maioria das vezes, essas empresas cobram um preço significativo para acessar seus dados estatísticos. Essas empresas costumam ter relacionamentos especiais com os principais aplicativos para os quais podem ser amplificadores de público ou ter os meios financeiros para desenvolver métodos poderosos, mas geralmente aproximados (como é o caso de Alexa[[14]](#footnote-14) agora uma subsidiária da Amazon, ou W3Techs).

Para encontrar fontes, existem duas opções:

1) coletar algumas informações gratuitas, mas muito parciais, que são filtradas por uma dessas empresas como forma de atrair clientes para pagar pela maioria das informações;

2) pagar para obter os dados.

A primeira abordagem foi aplicada de forma consistente com a implantação de intensa atividade de busca na Internet. No entanto, deve-se levar em conta que as metodologias geralmente não são transparentes e que esses números raramente se sustentam, o que é inconveniente para quem deseja manter uma observação perene. Recursos de estudo limitados levaram à seleção de Alexa e Statista[[15]](#footnote-15)como fontes pagas, por causa de sua melhor relação custo-benefício, e essas duas fontes alimentaram as estatísticas para a maioria dos aplicativos ou espaços desejados. A escolha de Alexa permitiu ampla cobertura[[16]](#footnote-16) mas é necessário saber que os números de tráfego por país para cada aplicativo fornecido por Alexa apresentam viés significativos (consulte o capítulo 7.4.6 Vieses potencial de W3Techs e Alexa.com). A lista completa das fontes selecionadas encontra-se detalhada no Anexo II. Fontes selecionadas.

### 2.4.1 O caso da língua materna (L1)

Para L1, este trabalho é baseado em uma contribuição significativa de Daniel Prado, que assumiu os dados da fonte do Projeto Joshua[[17]](#footnote-17)(que permite a distribuição de mais de 7.500 línguas em todo o mundo por país) e tem exigido, por um lado, ajustar os números, sempre que necessário, utilizando outras fontes diversas, e, por outro lado, adaptar os números à divisão por país escolhido. Ethnologue teria sido uma fonte de melhor qualidade para usar sem modificações, mas seu preço estava acima do orçamento.

A matriz LOC1 atende à definição de falantes, para as línguas selecionadas e todos os países selecionados: LOC1 (i, j) = Número de falantes L1 da língua i no país j.

A fonte fornece números para 7.500 línguas; faltava fazer uma escolha a partir da lista de línguas a serem processadas para estudo. Hoje, estima-se que o número de línguas com presença na Internet seja de cerca de 500. Uma possibilidade seria chegar o mais próximo possível dessas línguas. Outra possibilidade seria selecionar as línguas para os quais a Wikipédia fornece estatísticas (cerca de 300). Depois de muita reflexão e vários testes, a escolha final foi a lista de 140 línguas com mais de 5 milhões de falantes L1. Os valores para as demais línguas foram acumulados em uma categoria denominada "línguas restantes" que ainda é contabilizada na última linha da matriz LOC1.

A razão para esta escolha é a prudência em face de potenciais vieses, uma vez que a hipótese implícita do método de ponderação para transformar dados por país em dados por língua não fornece confiança suficiente nos resultados quando o número de falantes é baixo (para mais detalhes ver 5.1 Limitações específicas do método).

### 2.4.2 O caso do segundo língua (L2)

A primeira prioridade é levar em conta de forma consistente as implicações do multi-linguismo. Na verdade, quando uma segunda língua é mencionada, o seguinte fator é implicitamente introduzido: as pessoas contadas em L2 também têm uma primeira língua (L1) e, portanto, L1 + L2 inclui as mesmas pessoas mais de uma vez[[18]](#footnote-18). A evidência então diz que se alguém tentar medir a porcentagem global de pessoas que falam uma língua, seja como língua materna ou como segunda língua, esse número deve necessariamente ser baseado no número total de falantes por língua no mundo, um número muito maior do que a população mundial[[19]](#footnote-19). Essa evidência é frequentemente ignorada em muitas estatísticas e representa uma fonte potencial de erros graves. No cenário que foi adotado, com base nos números L2 do Ethnologue[[20]](#footnote-20), e com um suplemento médio global L2 de 18% dado como precaução (e ajustado para criar números arredondados) para os línguas ausentes na lista de 140 línguas selecionados, as cotas (não mais estritamente percentuais) são calculadas com base em uma soma que seria igual a 125% da população mundial (25% representando o grau mundial de multi-linguismo).

Também é importante entender que essa noção é igualmente aplicável a todos os conceitos processados: usuários, tráfego, usos, conteúdo, interfaces e índices. Na verdade, o multi-linguismo se aplicará a todos os elementos (os sites podem ser feitos em várias línguas, o mesmo para o fluxo de e-mails ou chats e trocas nas redes sociais) e depois se somarmos os números para as línguas do mundo, os totais calculados em relação à população mundial serão mais de 100%. Na realidade, é provável que o grau de multi-linguismo dos sites seja superior ao de seus visitantes, porém a aproximação de considerá-los equivalentes é aceitável.

O método ideal para lidar com o caso de L2 no modelo, obviamente, seria fazer o mesmo que para L1 e produzir uma matriz como:

LOC2 (i, j) = número de falantes L1 + L2 da língua i no país j.

Então bastaria aplicar o resto do método sem nenhuma alteração (levando em conta o multi-linguismo para os totais) para obter os resultados esperados com uma precisão muito apreciável. Infelizmente, os dados para construir tal matriz não existem e a existência de fortes diferenças entre as fontes demo-linguísticas para L2 torna a realização da referida matriz muito difícil no estado atual de conhecimento sobre o assunto. Por isso, propõe-se outra forma de tratar o problema, simplificando-o.

O princípio do método escolhido é simples e consiste, para cada língua, em obter um número que representa o incremento que será aplicado às quantidades de L1 para obter o valor L1 + L2. Este número será igual a 1 para a linguagem i se L2 (i) = 0. Para esta edição, o Ethnologue será tomado como fonte para L2, mantendo os valores existentes para L1. Os valores de L2 por língua diferente de zero são apresentados no Anexo III: Valores escolhidos para L2.

A taxa global de aumento resultante para os indicadores (1,25 com as hipóteses de trabalho[[21]](#footnote-21)) é o resultado da seguinte operação de ponderação:

j = L

Tg = **∑** L1 (j) x T12 (j)

j = 1

Onde:

L é o número total de línguas,

L1 (j) o número de falantes L1 para a língua j

T12 (j) a taxa de aumento de L1 para L1 + L2 para a língua j.

A seguinte equação mostra como o valor dos micro-indicadores para L1 + L2 é calculado a partir de seu valor para L1:

j = L

ML1 + L2 (i) = Tg x ML1 (i) / ∑ ML1 (j) x T12 (j)

j = 1

A divisão opera a ponderação quadrada em 100% e o produto por Tg está ali apenas para normalizar os resultados para um percentual igual a Tg.

Este método é obviamente menos preciso do que uma solução que poderia funcionar em nível de país com a matriz LOC2 ideal; suas limitações e vieses potenciais, e uma variante potencial simples do método para torná-lo mais preciso, são discutidos em 7.2.2 O caso de L2.

O método L1 + L2 é aplicado a todos os indicadores, exceto índices, conteúdos e interfaces e tradução que, por natureza, já têm significado direto para L1 + L2.

## 2.5 Processo

Os indicadores expressos diretamente em termos de linguagem (conteúdos e interfaces) não implicam qualquer dificuldade para o processo; Em seguida, será dada atenção aos indicadores expressos por país que merecem um tratamento especial. O princípio de cálculo para converter valores expressos como porcentagem mundial por país em valores expressos como porcentagem mundial por língua é o produto matricial. A primeira matriz é LOC, que possui as línguas selecionados no eixo x e os países selecionados no eixo y. A segunda matriz é o vetor MPn que contém os valores da fonte por país do micro-indicador n. O resultado do produto oferece o valor do micro-indicador expresso em porcentagem por língua (MLn).

LOC (i, j) = O número de falantes da língua i no país j.

MPn (j) = O valor medido para o micro-indicador n no país j.

j = P

MLn (i) = Σ LOC (i, j) x MPn (j) onde P é o número total de países.

j = 1

O produto da matriz MLn = LOC +. x MPn, em notação APL[[22]](#footnote-22), o = SUMPRODUCT (LOC; MPn) em notação Excel, é uma operação de ponderação que produz um novo vetor, desta vez o tamanho do número de línguas, e para o qual MLn (i) = o valor do micro-indicador n deduzido para língua i, com base nos dados demo-linguísticos dos línguas por país (LOC) e os valores medidos por país para o micro-indicador n (MPn). O coração do método é, na verdade, a ponderação dos valores medidos para este micro-indicador em cada país com a presença da língua em cada país. A validade dessa ponderação e seus possíveis vieses são discutidos no Capítulo 7.1 Limitações específicas do método.

Os totais do MLn são iguais aos do MPn, mas desta vez a distribuição é feita por língua e não por país.

A principal desvantagem deste método de cálculo é que ele não fornecerá resultados confiáveis ​​se a fonte de dados não relatar todos os países devido ao produto da matriz transformar essa ausência em um valor zero, o que leva a resultados inaceitáveis ​​para línguas com forte presença em países onde a fonte não informa.

### 2.5.1 Extrapolação

A forma de contornar esta situação é estimar, da melhor forma possível, os valores que faltam para os países não informados. É viável e razoável, na maioria dos casos, extrapolar os dados faltantes das informações existentes e outras informações disponíveis em outros lugares. Não é necessária precisão absoluta, mas um método simples onde os desvios da realidade são reais, mas com impacto limitado nos resultados do processamento estatístico.

Dois métodos diferentes foram adotados para resolver todos os casos:

**a) Extrapolação proporcional ao percentual de pessoas conectadas por país.**

Este método é aplicado quando é razoável considerar que os valores são naturalmente proporcionais ao percentual global de pessoas conectadas (é o caso, por exemplo, do tráfego do site). Se a fonte de dados for expressa em quantidades em vez de porcentagem, o total mundial deve primeiro ser extrapolado; se for expresso como uma porcentagem mundial, não será necessário. O percentual total ou restante é distribuído entre os países não informados na proporção de seus respectivos pesos em termos de conexão à Internet.

**b) Método de quartil**

Neste método, os países sem documentos têm os valores suplementados de um quartil dos números da fonte com base em sua porcentagem de pessoas conectadas. Após vários testes, a atribuição dos quartis foi determinada da seguinte forma:

* Se o país tiver menos de 15% de sua população conectada: a pontuação mais baixa
* Se o país tiver mais de 15%, mas menos de 35% de sua população conectada: primeiro quartil
* Se o país tiver mais de 35%, mas menos de 65% da população conectada: a mediana
* Se o país tiver mais de 65%, mas menos de 85% de sua população conectada: terceiro quartil
* Se o país tiver mais de 85% de sua população conectada: a pontuação mais alta.

Os micro-indicadores para os quais não existe um método de extrapolação que pareça óbvio são os mesmos para os quais o significado da transformação das figuras por país em figuras por língua não é claro e, portanto, são excluídos (ver a questão do significado da transformação de país em língua na seção 5.4.4 A questão do significado da transformação país > língua).

### 2.5.2 Processo para indicadores

Além do indicador UIT que fornece porcentagens de pessoas conectadas à Internet, por país, talvez outros indicadores de tecnologia e, no limite, algumas estatísticas da Wikimedia[[23]](#footnote-23), nenhum indicador único reflete uma verdade válida para todas as línguas. Na verdade, dificilmente se verificará o ideal de um grau absoluto de globalização, que nos permite considerar que o comportamento dos usos é o mesmo para um dado indicador, independente do país ou língua.

Só uma análise estatística de um importante conjunto de micro-indicadores, selecionados com os critérios mais cuidadosos possíveis, diversos e globalizados, permite, na ausência de uma medida confiável de conteúdo por língua.[[24]](#footnote-24), quantificar a presença de línguas na Internet, sob a condição de analisar e compreender os possíveis vieses nos resultados que podem ser derivados das hipóteses simplificadoras, das fontes e do próprio processamento estatístico.

O modelo também pode ajudar a compreender, por diferenciação temática, onde uma língua tem maior ou menor presença; na versão atual, apenas o indicador de tráfego composto por 316 micro-indicadores permite aventurar-se na análise temática.

Todos os indicadores de presença de línguas na Internet são apresentados nas duas tabelas a seguir. A primeira tabela mostra os indicadores e como são calculados. A segunda tabela mostra os macro-indicadores e como são calculados. Todos os indicadores são expressos em termos de participação mundial, com base na população total de falantes (população mundial multiplicada pelo fator de multi-linguismo, na versão atual, 125% da população mundial).

Tabela  : Descrição dos indicadores

|  |  |  |  |
| --- | --- | --- | --- |
| **INDICADOR** | **DEFINIÇÃO** | **TÉCNICA** | **CONFIABILIDADE** |
| **A: USUÁRIOS** | Indicador mono derivado dos números da UIT do % mundial de pessoas conectadas por país. | pesagem  país -> língua  sem extrapolação | Muito forte  Viés muito marginal |
| **B: USOS** | Inclui 11 micro indicadores  - Linhas telefônicas  - Mercado de comércio eletrônico  - Downloads do OpenOffice  - Uso de redes sociais + projeção 2021  - Uso da projeção Netflix + 2020  - Assinantes Facebook, Twitter, YouTube, LinkedIn | pesagem  país-> língua  extrapolado proporcionalmente  média truncada em 20% | Forte.  Baixo viés.  Mas teria que ser expandido para dar mais sentido ao número médio de micro-indicadores. |
| **C: TRÁFEGO** | Alexa mede o tráfego de uma seleção de 316 sites (ver detalhes no Anexo I). | pesagem  país-> língua  extrapolado proporcionalmente  média truncada para 20% | Relativamente bom  Mas o forte viés de Alexa pró-ocidental foi confirmado pelo tráfego vs. subscrito |
| **D: ÍNDICES**  **(sociedade da informação)** | Inclui 5 índices WebIndex dos seguintes critérios:  - E-governo  - Acesso universal  - E. participação  - Infraestrutura geral  - Índice Geral | pesagem  país-> língua  extrapolado pelo método quartil.  Transformação em dados percentuais mundiais ponderando com dados UIT  Média | Bom (são dados subjetivos quantificados por um órgão competente).  Esta categoria deve ser ampliada. |
| **E: CONTEÚDO**  **(Wikimedia e livros)** | Inclui 13 micro indicadores  - Número de livros na Amazon  - As medidas contidas na linguagem W3Techs  - 11 indicadores linguísticos da Wikimedia: artigos, usuários ou editores da Wikipedia, wikibook, Wikibooks, wikisource, Wikiversity, wiktionnnaire. | Uso direto de figuras por língua.  Média truncada para 20% dos micro-indicadores | Muito forte para Wikimedia e Amazon.  Mas presença muito baixa de algumas das principais línguas asiáticas.  o número de micro-indicadores teria que ser ampliado para dar mais força à média |
| **F: INTERFACES**  **(e línguas de tradução)** | Inclui 23 micro-indicadores binários  - 12 línguas de interface de Cortana, DuckDuckGo, Facebook, GoogleSearch, GoogleScholar, Skype, Telegram, Wikibook, Wikipedia, Wikisource, Wikiquote, Wikiversity  - 1 língua de conteúdo para Dmoz  - 10 línguas de tradução de Bing, Dictionnary, Duolingo, FreeTranslator, GoogleTranslate, IM, Online, Dicionário, SDL, Systran | % de presença em todos os 23 micro-indicadores.  % ponderação mundial com dados ITU. | Perfeito. |

Tabela  : Descrição dos macro-indicadores

|  |  |  |
| --- | --- | --- |
|  | **DEFINIÇÃO** | **TÉCNICA** |
| ***PODER*** | Mede a participação global da língua na Internet | Média dos 6 indicadores |
| ***CAPACIDADE*** | Ele mede a força da língua na Internet, independentemente do número de falantes. | Relação entre poder e % mundial de falantes |
| ***GRADIENTE*** | Ele mede a força dos falantes conectados, independentemente de seu número. | Relação entre poder e em todo o % mundial de falantes conectados |
| ***PRODUTIVIDADE*** | Ele mede a propensão dos falantes online de produzir conteúdo em sua língua.[[25]](#footnote-25) | Relação entre % conteúdos e % de falantes conectados |

Os primeiros 3 macro-indicadores estão ligados pelas seguintes equações:

**Poder (**língua) = habilidade (língua) x % população mundial de falantes (língua)

**Poder** (língua) = gradiente (língua) x % da população mundial conectada à Internet (língua)

Deve-se observar que:

* A- Os dados da UIT sobre os usuários (pessoas conectadas à Internet) são um indicador chave do modelo e "pesam" muito nos resultados, diretamente na média para calcular a potência que lhe dá um peso tão importante quanto o 300 microindicadores de tráfego Alexa, e indiretamente nas técnicas de extrapolação e ponderação, que afetam transversalmente todos os cálculos. Toda a estrutura de processamento é baseada nestes dados considerados muito confiáveis.
* C- Para os dados de tráfego obtidos no Alexa, a validade reside no grande número de sites medidos, embora os resultados tenham um forte desvio padrão causado pela seleção de sites altamente marcados em uma população local (ver seção 7.4.7 Limitações / enviesamento relacionadas com o grau de localidade das fontes). A técnica da média truncada permite eliminar valores extremos e dar mais confiança ao resultado. Além disso, esses dados mostram um forte viés em favor das línguas ocidentais (principalmente inglês e francês) e especialmente em detrimento das línguas asiáticas, o que é discutido na seção 7.4.6 Potencial viés de W3Techs e Alexa.com.
* D- Os índices abrem uma perspectiva muito interessante sobre o alcance dos elementos contextuais criados por organizações internacionais ou não governamentais. No momento, WebIndex[[26]](#footnote-26) foi parcialmente incluído; é um projeto da World Wide Web Foundation que mantém uma série de índices baseados no trabalho sério dos serviços das Nações Unidas, do Banco Mundial e de outras instituições. Os índices utilizados neste estudo incluem os serviços de governo eletrônico, a participação do cidadão pela Internet, o grau de universalização do acesso e a infraestrutura para a sociedade da informação, além de um índice geral que incorpora todos os fatores. É importante saber que WebIndex oferece outros indicadores além dos cinco que já foram utilizados e que outras organizações já oferecem ou oferecerão no futuro índices que seria útil incluir em uma versão futura[[27]](#footnote-27).
* E: Este indicador é chamado de conteúdo, mas no momento está muito focado na galáxia da Wikimedia, o que oferece a enorme vantagem de fornecer estatísticas detalhadas e confiáveis ​​para a presença de línguas em suas diferentes atividades. É complementado por dados da W3Techs (não muito credíveis) e livros da Amazon, o que dificilmente amplia a perspectiva. Este indicador, que fornece o segredo mais bem guardado da Internet, a distribuição dos conteúdos por língua, claramente merece uma ampliação do número de fontes ... desde que as fontes sejam credíveis, o que se tornou um terrível desafio. Deve-se levar em consideração, no entanto, que as principais línguas asiáticas estão muito mal representadas na Wikimedia[[28]](#footnote-28). Também é importante entender que por natureza (indicador direto de línguas) este indicador penalizará completamente (com notas nulas) línguas fora do universo da Wikimedia, que embora seja o grupo mais preocupado com a questão linguística, está limitado a cerca de 300 línguas.
* F: A penalidade para línguas fracamente conectados (mas também línguas que estão bem conectados, mas não levados em consideração nas traduções e interfaces) será ainda mais forte do que para E. Este indicador fará com que as línguas com pontuação mais baixa caiam ainda mais na classificação, embora reflita uma realidade que não pode ser ignorada.

### 2.5.3 Processo para alcançar resultados diferenciados por assunto

O único indicador com um número de elementos suficiente para permitir uma análise mais detalhada por assunto é o tráfego com seus 316 elementos. Cada site está associado a um tema que o representa e com a média truncada em 20%, calcula-se a diferença positiva ou negativa de cada tema em relação à média geral. No entanto, esses resultados devem ser vistos com cautela, especialmente se o número de sites para um determinado tópico for baixo. Até o momento, essa análise foi feita apenas pelo francês que mostra sua força no conteúdo da pesquisa científica, nos MOOCs[[29]](#footnote-29), em conteúdos profissionais, em redes sociais de música, em jogos e em motores de pesquisa.

### 2.5.4 Tipos de ponderação usados

O método estatístico utilizado baseia-se principalmente nos métodos de ponderação e extrapolação. É útil identificar e sintetizar os diferentes tipos de ponderação usados ​​no processo e tornar explícitas as hipóteses simplificadoras que afetam a validade dos resultados obtidos (a discussão detalhada das implicações e possíveis vieses resultantes das hipóteses são apresentados no capítulo 7.1 Limitações específicas para o método).

A principal ponderação é aquela que transforma as percentagens por país em percentagens por língua: os valores de um micro-indicador expressos em termos de países são ponderados em relação à distribuição das línguas nos países para fornecer a distribuição por língua. A suposição implícita que sustenta a validade desse processo é que o conceito medido pelo micro-indicador é expresso da mesma forma para todas as línguas do país.

A ponderação que permite converter os valores L1 em ​​valores L1 + L2 é feita com base na taxa de aumento de L1 para L1 + L2 para cada língua. A suposição simplificadora desse método é que as porcentagens de conexão à Internet são iguais para falantes L2 e L1, independentemente do país de residência.

A ponderação que permite converter os valores expressos de uma porcentagem de um determinado critério (interfaces, índices) em uma porcentagem mundial é feita em relação às porcentagens de pessoas conectadas por língua. O princípio por trás desse método é que as porcentagens mundiais obtidas para o critério são um desvio da porcentagem global de falantes conectados em função da distribuição por língua deste micro-indicador. Uma aproximação mais intuitiva do significado dessa ponderação é fornecida no capítulo 7.4.8 Sobre o princípio da ponderação.

Tabela : Os 3 tipos de ponderação usados

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Demo-linguística** | **Segundas línguas** | **Usuários da Internet por língua** |
| TIPO | País ---> Língua | L1 ---> L1 + L2 | % critério ->% mundo |
| ENTRADA | Dados por país | Resultados L1 | Dados em % critério |
| RESULTADO | Dados por língua | Resultados L1 + L2 | Dados em % mundial |
| DADOS DE  PESAGEM | Matriz LOC | Vetor L1 + L2 / L1 por língua | Dados UIT |
| ESCOPO | Todas as fontes por país | Indicadores de usuário, tráfego e uso | Índices e interfaces. |
| HIPÓTESE  IMPLÍCITA | Independência em relação às línguas do país | Independência das taxas de conexão à Internet dependendo dos países | Modulação das taxas de conexão à Internet de acordo com os critérios |

# Resultados

A tabela a seguir mostra todos os resultados sobre as 15 línguas mais «poderosas» da Internet. Os resultados, expressos em termos de participação mundial, são apresentados separados entre a primeira língua (L1) e o primeiro mais a segunda língua (L1 + L2), para cada indicador (usuários, tráfego, uso, conteúdos, índices e interfaces) e para o macro-indicador "poder".

Tabela : Indicadores para os 15 principais línguas em termos de poder

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **COMERCIAL** | | **TRÁFEGO** | | **FORMULÁRIOS** | | **CONTEÚDO** | **INTERFACE** | **ÍNDICE** | **PODER** |
|  | **L1** | **L1 + L2** | **L1** | **L1 + L2** | **L1** | **L1 + L2** | **L1 + L2** | **L1 + L2** | **L1 + L2** | **L1 + L2** |
| Inglês | 0,085 | 0,222 | 0,283 | 0,568 | 0,192 | 0,437 | 0,499 | 0,309 | 0,305 | **0,390** |
| Chinês | 0,218 | 0,205 | 0,081 | 0,062 | 0,125 | 0,108 | 0,030 | 0,248 | 0,194 | **0,141** |
| Espanhol | 0,080 | 0,091 | 0,109 | 0,094 | 0,095 | 0,095 | 0,053 | 0,121 | 0,090 | **0,091** |
| Francês | 0,020 | 0,056 | 0,036 | 0,080 | 0,029 | 0,074 | 0,093 | 0,074 | 0,073 | **0,075** |
| Russo | 0,028 | 0,050 | 0,015 | 0,022 | 0,025 | 0,041 | 0,054 | 0,066 | 0,051 | **0,047** |
| Alemão | 0,019 | 0,031 | 0,028 | 0,037 | 0,026 | 0,039 | 0,084 | 0,042 | 0,039 | **0,046** |
| Português | 0,040 | 0,040 | 0,026 | 0,020 | 0,048 | 0,043 | 0,033 | 0,053 | 0,039 | **0,038** |
| Japonês | 0,036 | 0,034 | 0,031 | 0,024 | 0,038 | 0,032 | 0,030 | 0,043 | 0,046 | **0,035** |
| Árabe | 0,044 | 0,042 | 0,042 | 0,031 | 0,032 | 0,028 | 0,013 | 0,048 | 0,032 | **0,032** |
| Hindi | 0,031 | 0,039 | 0,020 | 0,020 | 0,024 | 0,027 | 0,003 | 0,035 | 0,033 | **0,026** |
| Malaio | 0,006 | 0,026 | 0,005 | 0,015 | 0,006 | 0,022 | 0,013 | 0,027 | 0,021 | **0,021** |
| Italiano | 0,009 | 0,009 | 0,014 | 0,011 | 0,011 | 0,010 | 0,041 | 0,012 | 0,010 | **0,016** |
| Coreano | 0,015 | 0,014 | 0,012 | 0,009 | 0,011 | 0,009 | 0,009 | 0,016 | 0,016 | **0,012** |
| Polonês | 0,009 | 0,008 | 0,010 | 0,008 | 0,009 | 0,008 | 0,030 | 0,009 | 0,008 | **0,012** |
| Urdu | 0,009 | 0,017 | 0,009 | 0,013 | 0,008 | 0,013 | 0,001 | 0,011 | 0,014 | **0,011** |
| **RESTO** | **0,351** | **0,368** | **0,279** | **0,236** | **0,321** | **0,265** | **0,264** | **0,135** | **0,279** | **0,258** |
| **TOTAL** | **1.000** | **1.250** | **1.000** | **1.250** | **1.000** | **1.250** | **1.250** | **1.250** | **1.250** | **1.250** |

A linha RESTO representa os resultados para o conjunto completo de as línguas do mundo, exceto as 15 línguas que aparecem na tabela.

Deve ficar claro que a classificação em termos de poder privilegia as línguas que possuem o maior número de falantes. Para classificações que medem indicadores de língua que são independentes do número de falantes, as tabelas a seguir oferecem resultados interessantes.

Tabela  : Línguas ordenados por capacidade

|  |  |  |
| --- | --- | --- |
|  | **Capacidade** | **Classificação**  **Poder** |
| Hebraico | 5,40 | 35 |
| Finlandês | 5,40 | 38 |
| Holandês | 4,81 | 19 |
| Sueco | 4,46 | 28 |
| Inglês | 3,72 | 1 |
| Alemão | 3,40 | 6 |
| Dinamarquês | 3,30 | 49 |
| Italiano | 3,16 | 12 |
| Tcheco | 3,13 | 27 |
| Francês | 2,96 | 4 |
| Servo-croata | 2,83 | 22 |
| |  | | --- | | Húngaro | | 2,83 | 32 |
| Polonês | 2,64 | 14 |
| Japonês | 2,54 | 8 |
| Grego | 2,40 | 37 |

É claro que existe uma correlação forte e consistente entre a capacidade e a porcentagem de pessoas conectadas à Internet. No entanto, a correlação não é total, conforme mostrado na classificação a seguir ordenada pela porcentagem de pessoas conectadas.

Tabela  : Línguas ordenados por porcentagem de pessoas conectadas

|  |  |  |
| --- | --- | --- |
|  | **% Pessoas**  **conectadas** | **Classificação**  **Poder** |
| Dinamarquês | 95,67% | 49 |
| Finlandês | 92,30% | 38 |
| Holandês | 92,27% | 19 |
| Sueco | 90,54% | 28 |
| Japonês | 90,43% | 8 |
| Alemão | 86,43% | 6 |
| Alemão | 86,41% | 51 |
| Bávaro | 84,42% | 39 |
| Flamengo | 83,60% | 59 |
| Tcheco | 81,17% | 27 |
| Francês | 81,09% | 4 |
| Inglês | 78,05% | 1 |
| Eslovaco | 77,38% | 62 |
| Hebraico | 76,05% | 35 |
| Húngaro | 71,16% | 32 |

E por último, a última tabela, ordenada por gradiente, destaca o dinamismo das pessoas conectadas.

Tabela : Línguas ordenados por gradiente

|  |  |  |
| --- | --- | --- |
|  | **Gradiente** | **Classificação**  **Poder** |
| Hebraico | 2,62 | 35 |
| Finlandês | 2,16 | 38 |
| Holandês | 1,93 | 19 |
| Sueco | 1,81 | 28 |
| Inglês | 1,76 | 1 |
| Italiano | 1,73 | 12 |
| Servo-croata | 1,54 | 22 |
| Húngaro | 1,47 | 32 |
| Alemão | 1,45 | 6 |
| Tcheco | 1,42 | 27 |
| Polonês | 1,41 | 14 |
| Francês | 1,35 | 4 |
| Grego | 1,30 | 37 |
| Dinamarquês | 1,27 | 49 |
| Armênio | 1,20 | 60 |

É de se notar sem surpresa que as línguas nacionais dos países reconhecidos por suas políticas proativas para a sociedade da informação aparecem nas primeiras posições dos últimos 3 macro-indicadores. É de notar que várias línguas estão antes do inglês, apesar da sua vantagem estratégica na Internet, por ser a língua preferida para conteúdos multilíngues e a crença de muitos que é a língua franca da Internet.

# Análise dos resultados

É importante saber que dois dos seis indicadores (tráfego, medido pelo Alexa e conteúdo, baseado nas estatísticas da Wikimedia) podem ter, à sua maneira, um viés muito significativo a favor do inglês e contra, primeiro, de línguas não ocidentais (especialmente chinês, hindi, urdu, bengali e, em menor medida, árabe) e, em segundo lugar, embora em menor medida, de algumas outras línguas ocidentais, no caso de Alexa (português e, em menor medida, espanhol). Os vários vieses possíveis devem ser analisados.

## 4.1 Sobre a Wikipedia

As estatísticas da Wikimedia são impecáveis; no entanto, deve-se entender que, apesar de ser uma das aplicações mais globais da Internet, os números para algumas línguas asiáticas estão bem abaixo de sua presença relativa na Internet. A tabela a seguir compara as proporções entre os artigos da Wikipedia e os usuários da Internet. Grandes variações aparecem com valores anormalmente baixos para línguas asiáticos.

Tabela : Artigos da Wikipedia vs. presença na internet

|  |  |  |  |
| --- | --- | --- | --- |
| **LÍNGUA** | **% PESSOAS CONECTADAS** | **% ARTIGOS WIKIPEDIA** | **RELAÇÃO** |
| Sueco | 0,2% | 8,8% | 44 |
| Holandês | 0,4% | 4,5% | 11 |
| Francês | 5,6% | 4,3% | 0,8 |
| Inglês | 22,6% | 12,6% | 0,6 |
| Chinês | 20,9% | 2,8% | 0,13 |
| Urdu | 2,0% | 0,3% | 0,15 |
| Hindi | 3,9% | 0,3% | 0,08 |
| Bengali | 1,3% | 0,1% | 0,08 |

## Comparação dos resultados com os do InternetWorldStats (IWS)

É interessante comparar os resultados deste estudo com os dados fornecidos pela InternetWorldStats[[30]](#footnote-30), a fonte mais citada por vários anos para dados de usuários da Internet por língua (embora limitada a apenas 10 línguas com o maior número de usuários). Usando, em princípio, a mesma fonte para pessoas conectadas (UIT), os dois métodos devem, a partir dos mesmos dados demo-linguísticos, fornecer resultados muito semelhantes. Mas não é o caso. A análise mostra três causas diferentes para explicar diferenças significativas: 1) hipóteses muito diferentes para L2, 2) a gestão do multi-linguismo, 3) um erro que aparece na porcentagem de francófonos conectados à Internet (IWS apresenta um valor anormalmente baixo de 26 %, enquanto este estudo indica 81%).

Este estudo fornece uma porcentagem muito maior de números de falantes conectados do que o IWS para francês, russo e, em menor grau, para espanhol e alemão. As diferenças nos dados demo-linguísticos usados ​​explicam essas diferenças até certo ponto. No entanto, uma simulação do modelo usado neste estudo com os números demo-linguísticos exatos usados ​​pelo IWS continua a mostrar diferenças significativas em termos da porcentagem global de pessoas conectadas para francês (75%) e inglês (25%)[[31]](#footnote-31). Por outro lado, gerenciar o multi-linguismo com porcentagens da população mundial levaria rapidamente ao absurdo sobre o restante das línguas do IWS se sua lista de línguas fosse estendida.

Tabela : Diferenças de hipóteses com IWS para L1 + L2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LÍNGUA** | **L1 + L2** | **L1 + L2** | **diferenças** | **Diferenças** |
|  | **IWS** | **estude** | **absoluto** | **%** |
| Inglês | 1404089957 | 955652967 | 448436990 | **47%** |
| [Chinês](http://www.internetworldstats.com/stats17.htm) | 1417846280 | 1347846280 | 70000000 | **5%** |
| Espanhol | 454877800 | 525 877 800 | 71000000 | **-14%** |
| Árabe | 388915929 | 333 915 929 | 55000000 | **16%** |
| [Português](http://www.internetworldstats.com/stats20.htm) | 267372370 | 228 372 370 | 39000000 | **17%** |
| Japonês | 125388300 | 125388300 | 0 | **0%** |
| [Malaio](https://en.wikipedia.org/wiki/Malay_language) | 293 998 640 | 228 998 640 | 65000000 | **28%** |
| [Russo](https://en.wikipedia.org/wiki/Russian_language) | 146013300 | 243 287 120 | 97273820 | **-40%** |
| Francês | 394660700 | 230 146 470 | 164 514 230 | **71%** |
| Alemão | 95419000 | 122108000 | 26689000 | **-22%** |

Tabela : Dados da InternetWorldStats (junho de 2016)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LÍNGUA** | **USUARIOS**  **INTERNET** | **%**  **FALANTES** | **%** | **POPULAÇÃO** |
|  |  | **CONECTADO** | **MUNDIAL** |
| Inglês | 948 608 782 | 67,76% | 26,27% | 1 400 052 373 |
| [Chinês](http://www.internetworldstats.com/stats17.htm) | 751 985 224 | 53,12% | 20,82% | 1 415 572 934 |
| Espanhol | 277 125 947 | 61,55% | 7,67% | 450 235 963 |
| Árabe | 168 426 690 | 43,37% | 4,66% | 388 332 877 |
| [Português](http://www.internetworldstats.com/stats20.htm) | 154 525 606 | 57,93% | 4,28% | 266 757 744 |
| Japonês | 115 111 595 | 91,02% | 3,19% | 126 464 583 |
| [Malaio](https://en.wikipedia.org/wiki/Malay_language) | 109 400 982 | 37,76% | 3,03% | 289 702 633 |
| [Russo](https://en.wikipedia.org/wiki/Russian_language) | 103 147 691 | 70,48% | 2,86% | 146 358 055 |
| Francês | 102 171 481 | *25,94%* | 2,83% | 393 892 299 |
| Alemão | 83 825 134 | 88,26% | 2,32% | 94 973 855 |
| RESTO | 797 046 681 | *33,66%* | 22,07% | 2 367 750 664 |
| [TOTAL](http://www.internetworldstats.com/stats.htm) | 3 611 375 813 | 49,20% | 100,00% | 7 340 093 980 |

A tabela a seguir mostra os dados obtidos neste estudo e a comparação com o IWS.

Tabela : Comparação com IWS após redução para 100%

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **LÍNGUA** | **USUARIOS**  **INTERNET** | **% FALANTES** | **%**  **MUNDIAL** | **POPULAÇÃO** | **DIFFE** | **RENçAS** |
|  |  | **CONECTADO** |  |  | **No %** |
| Inglês | 746575851 | 78,18% | 23,94% | 954 971 600 | -2,32% | -9,71% |
| [Chinês](http://www.internetworldstats.com/stats17.htm) | 692 230 352 | 51,36% | 22,20% | 1347846000 | 1,38% | 6,21% |
| Espanhol | 306111437 | 58,21% | 9,82% | 525 877 800 | 2,14% | 21,83% |
| Árabe | 140 122 838 | 41,97% | 4,49% | 333 842 070 | -0,17% | -3,78% |
| [Português](http://www.internetworldstats.com/stats20.htm) | 134 142 584 | 58,74% | 4,30% | 228 372 370 | 0,02% | 0,54% |
| Japonês | 113372837 | 90,43% | 3,64% | 125376800 | 0,45% | 12,33% |
| [Malaio](https://en.wikipedia.org/wiki/Malay_language) | 87701961 | 38,30% | 2,81% | 228 998 640 | -0,22% | -7,70% |
| [Russo](https://en.wikipedia.org/wiki/Russian_language) | 166904263 | 68,60% | 5,35% | 243 287 120 | 2,50% | 46,64% |
| Francês | 186641894 | 81,10% | 5,99% | 230 146 420 | 3,16% | 52,73% |
| Alemão | 105 539 688 | 86,43% | 3,38% | 122108000 | 1,06% | 31,42% |
| **RESTO** | **438771072** | **16,05%** | **14,07%** | **2733112497** | **-8,00%** | -56,84% |
| [**TOTAL**](http://www.internetworldstats.com/stats.htm) | **3118114777** | **44,08%** | **100,00%** | **7073939317** | **0,00%** |  |

Algumas línguas apresentam uma diferença marcante, em ordem decrescente: o restante das línguas (-56%), francês (+ 53%), russo (- 47%), alemão (+ 31%), espanhol (- 22%). A diferença que aparece no resto é uma consequência clara de como o IWS lida com o multi-linguismo.

A tabela a seguir mostra os resultados quando os mesmos dados demo-linguísticos do IWS são aplicados ao modelo deste estudo. Com esses dados, o modelo estabelecido no presente estudo encontrou valores mais elevados para o inglês (+ 25%) e o francês (+ 72%). Também mostra um resto negativo, um sintoma de um tratamento inadequado do multi-linguismo. As diferenças são explicadas pelo tratamento do multi-linguismo (a aplicação da porcentagem de falantes L1 + L2 sobre a população mundial em vez do total mundial de falantes leva a uma contradição no IWS que está oculta pelo fato de ser um número limitado de línguas). Se isso fosse corrigido, ainda haveria um grande erro no tratamento do francês pela IWS, conforme mencionado acima.

Tabela : Simulação com dados IWS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LÍNGUA** | **USUÁRIOS DE INTERNET** | **% FALANTES** | **%** | **POPULAÇÃO** |  |
|  |  | **CONECTADO** | **GLOBAL** | **DIFERENÇAS** |
| Inglês | 1094530237 | 78,18% | 35,10% | 1400052373 | 25% |
| [Chinês](http://www.internetworldstats.com/stats17.htm) | 727013732 | 51,36% | 23,32% | 1415572934 | 11% |
| Espanhol | 262080616 | 58,21% | 8,41% | 450235963 | 9% |
| Árabe | 162994152 | 41,97% | 5,23% | 388332877 | 11% |
| [Português](http://www.internetworldstats.com/stats20.htm) | 156689591 | 58,74% | 5,03% | 266757744 | 15% |
| Japonês | 114 356 472 | 90,43% | 3,67% | 126 464 583 | 13% |
| [Malaio](https://en.wikipedia.org/wiki/Malay_language) | 110 950 392 | 38,30% | 3,56% | 289702633 | 15% |
| [Russo](https://en.wikipedia.org/wiki/Russian_language) | 100 407 220 | 68,60% | 3,22% | 146 358 055 | 11% |
| Francês | 319 434 926 | 81,10% | 10,24% | 393892299 | 72% |
| Alemão | 82087259 | 86,43% | 2,63% | 94973855 | 12% |
| **RESTO** | **12429819** | **-0,59%** | **-0,40%** | **2101596001** |  |
| [**TOTAL**](http://www.internetworldstats.com/stats.htm) | **3118114777** | **44,08%** | **100,00%** | **7073939317** |  |

## 4.3 Sensibilidade dos fatores

Todo o modelo descrito foi automatizado em uma planilha Excel e é possível obter instantaneamente os novos resultados de qualquer alteração de parâmetro. Essa flexibilidade, que foi utilizada para a comparação com o IWS, também permitiu testar a sensibilidade de cada fator nos resultados. Portanto, parecia, no caso do francês, que sua quarta posição em termos de potência seria mantida se os dados de L2 fossem alterados para um valor muito inferior. Pelo contrário, a sensibilidade do indicador de tráfego em relação à escolha dos sites foi considerada muito forte se for utilizada a média simples; isso levou à decisão de usar a média truncada de 20% para evitar viés de seleção.

Diferentes hipóteses sobre o número de segundas línguas, um método mais preciso para segundas línguas, a escolha de uma métrica diferente para determinar indicadores (média, média truncada, mediana) ou diferentes combinações de micro-indicadores, obviamente, leva a resultados de indicadores com algumas diferenças relativamente pequenas[[32]](#footnote-32) mas eles podem alterar a ordem dos línguas que estão muito próximos nos resultados[[33]](#footnote-33).

Os diversos testes realizados mostraram que o estudo fornece uma ferramenta de trabalho e simulação capaz de oferecer resultados imediatos quando os parâmetros são alterados, o que provavelmente o torna um instrumento único e inédito no campo da cibermetria de língua. É tentador estender o estudo a simulações baseadas não apenas em figuras L1 + L2, mas adicionando um conceito de Li ad hoc que poderia ser definido como a população com domínio suficiente de uma linguagem para fazer uso eficaz das várias características da Internet, embora continua sendo um grande desafio encontrar figuras demo-linguísticas confiáveis ​​e consistentes para Li em todas as línguas, incluindo as mais importantes...

# Limitações metodológicas, análise de viés e controles realizados

## 5.1 Cuidado com os viés!

A demo-linguística está longe de ser uma ciência exata e as divergências entre os especialistas em figuras de L1, L2 e Li, são importantes e crescem de L1 a Li, quando as definições se tornam menos precisas. Por outro lado, a Internet é um universo em expansão que atingiu cifras impressionantes quanto ao número de pessoas conectadas², servidores e páginas da web e para o qual não existem fontes disponíveis para o tamanho da Web (e os índices dos buscadores). Por outro lado, algumas fontes exigem muita latitude com os dados para responder ao domínio do marketing de cliques e da publicidade que destronou outras formas de modelo de negócios na Internet mais compatíveis com a teoria da informação.

A história da cibermetria linguística foi marcada pelo triunfo do viés... Em 2000, Inktomi, um mecanismo de busca extinto, declarou, com enorme poder de marketing, que mais de 85% da web estava ligada Inglês, com base em dados expostos sem qualquer informação sobre o método de cálculo. Apenas 10 línguas foram mostradas... e um total de 100%, como se as outras línguas não contassem para nada! Uma estimativa séria do resto das línguas teria levado a presença do inglês para 70% ...

Tabela : Distribuição de línguas na Web de acordo com Inktomi (2000) :

|  |  |
| --- | --- |
| LÍNGUA | PROPORÇÃO DO SITE (%) |
| Inglês | 86,54 |
| Alemão | 5,83 |
| Francês | 2,36 |
| Italiano | 1,55 |
| Espanhol | 1,23 |
| Português | 0,75 |
| Holandês | 0,54 |
| Finlandês | 0,50 |
| Sueco | 0,36 |
| Japonês | 0,34 |

Em 1997, uma empresa canadense, Alis Tecnologies, publicou os resultados de um estudo realizado com o apoio da Internet Society e apontou 80% do inglês na web. Um grupo de pesquisadores da OCLC[[34]](#footnote-34) retomou o método em 1999 e 2003 e alcançou o mesmo resultado de 80% em inglês de forma estável e atemporal... apesar do crescimento exponencial da Web e do declínio constante na porcentagem de falantes de inglês na Internet. Esses dados alimentaram a versão, retomada e ampliada pela mídia, de uma Web estável com 80% de páginas em inglês por 10 anos, entre 1987 e 2007. Porém, no mesmo período, o observatório da Funredes apresentou uma diminuição do espaço do inglês na web de 80% em 1997 para 45% em 2007. Foram necessárias duas publicações da UNESCO, em 2006 e 2009[[35]](#footnote-35), para acabar com essa desinformação. Qual foi o principal viés do método Alis adotado pela OCLC? O método consistiu na aplicação de um algoritmo de reconhecimento de linguagem em **uma amostra aleatória de 3.000 sites**, selecionados a partir da geração aleatória de números IP. Se o método consistisse em repetir a medição cem vezes (variando a amostra ao acaso) e tratasse os resultados como a distribuição de uma variável aleatória à qual as leis estatísticas foram aplicadas (média, covariância, lei de Fisher, intervalos de confiança ...) teria sofrido apenas vieses moderados (a escolha da página inicial favorece o inglês, os algoritmos de reconhecimento de língua favorecem o inglês e a ausência de gerenciamento de multi-linguismo faz o mesmo), mas o problema era que os resultados eram o fruto de **um único sorteio aleatório** e, portanto, apenas sujeito à vontade do acaso!

Um estudo que vai combinar dados específicos da Internet com dados demo-linguísticos para a construção de indicadores está na intersecção desses dois mundos incertos, o que, em termos estatísticos, implica a multiplicação de incertezas e aumenta o tamanho de intervalos de confiança. Portanto, você precisa ter muito cuidado para identificar vieses que podem afetar seus dados ou metodologia. Somente assim será possível validar a confiabilidade do método, inferir as possíveis implicações de vieses nos resultados e tentar encontrar as correções adequadas.

Este capítulo tenta sistematizar a análise de todos os vieses ou limitações que podem ser causados ​​direta ou indiretamente pelas opções e / ou fontes metodológicas nesse estudo e mostra os controles que foram feitos no tratamento dos dados para garantir resultados relativamente confiáveis, mesmo que seja impossível atingir a perfeição neste campo.

## 5.1 Limitações e vieses do método

O método se dedica a pesquisar, identificar, avaliar e utilizar micro-indicadores quantitativos para construir indicadores que caracterizem adequadamente a presença de línguas na Internet. Sendo as fontes por língua muito escassas, optou-se por utilizar as fontes por país para ultrapassar esta limitação e geri-la de forma a obter uma visão estatística suportada por suficientes micro-indicadores. Esta opção apresenta um desafio: o da transformação de figuras por país em figuras por língua. Este desafio foi resolvido através do método de ponderação entre a distribuição de falantes de diferentes línguas por país e os dados por país de algumas aplicações ou espaços na Internet. Agora é necessário entender completamente os critérios de validade e as limitações dessa abordagem de ponderação.

Foi criada uma matriz contendo os 7500 línguas e 192 países. Por meio da aritmética simples, é possível distribuir os dados por país para cada critério medido entre cada língua (na proporção da presença de línguas em cada país) e, assim, obter dados por língua para cada um dos critérios medidos. Qual é a hipótese que suporta este método e está correta?

A suposição implícita é que os dados sobre a Internet em um determinado país são válidos da mesma forma para todas as línguas faladas naquele país. Isso corresponde à realidade ou é pelo menos próximo?

Para tomar o exemplo mais óbvio, a porcentagem de pessoas conectadas à Internet em um país: essa porcentagem é a mesma para cada língua falado e, portanto, independente da língua materna dos habitantes desse país? É claro que esta é uma hipótese simplificadora que implica, neste exemplo, que a exclusão digital não se correlaciona com a língua dentro do país. Isso geralmente é falso: falantes de determinados línguas provavelmente terão uma taxa de conexão à Internet significativamente mais baixa ou mais alta do que a média nacional. Pode ser, porque representam uma população com níveis socioeconômicos, culturais ou de escolaridade muito diferentes da média. Por tanto, esta é uma abordagem aceitável que não introduz distorção violenta nos resultados sob a condição de não aplicá-la a espaços ou aplicações que possam agravar esse viés (por exemplo, está simplificação não poderia ser válida par uma aplicação que consiste em aprender um das línguas de estudo) e também **sob a condição de não considerá-la válida para diferenciar a respetiva situação das línguas dentro de um país[[36]](#footnote-36).**

Para cada língua, além do país (ou países) em que é o língua nacional, será utilizada a distribuição de falantes em diferentes países[[37]](#footnote-37)e permanece a questão se é reciprocamente correto considerar que falantes de uma língua que residem em um país onde esta não é uma língua nacional têm a mesma taxa de conectividade com a Internet que falantes da língua nacional desse país. Intuitivamente, pode-se entender que o risco de erro (ou o tamanho da possível diferença entre a taxa nacional e a taxa de uma língua não nacional) é maior quando se trata de línguas minoritárias com baixo número de falantes. Por esse motivo, a escolha foi finalmente feita em línguas com mais de 5 milhões de falantes (L1) como medida de prudência.

Este problema afeta apenas cálculos para línguas nativos (L1). O tratamento de L2 provavelmente causa outro tipo de risco de distorção entre os cálculos e a realidade e merece uma análise mais detalhada que será feita no próximo capítulo.

A questão das fontes de dados incompletas foi resolvida pela técnica de extrapolação, aplicada diretamente na proporção direta com a taxa de conectividade dos respectivos países, ou pelo método do quartil. Este método resolve a situação corretamente e pode-se considerar que o enviesamento introduzido (no caso em que os países realmente têm um valor para este micro-indicador muito superior ou muito inferior ao calculado por extrapolação, por motivos particulares relacionados com a natureza deste micro-indicador[[38]](#footnote-38)) serão compensados ​​em média e se tornarão marginais quando o número de casos for grande o suficiente.

Agora temos que nos questionar sobre as limitações ou vieses de cada um dos critérios relacionados aos micro-indicadores em cada elemento do complexo utilizado: línguas, países e fontes.

## 5.2 Línguas

### 5.2.1 Selecionando a fonte para o cálculo de L1

Nesta edição, o Projeto Joshua foi a principal fonte de dados demo-linguísticos na língua materna (L1). Na próxima edição, está contemplado o uso de dados do Ethnologue, o que seria um custo adicional para o estudo, a menos que um acordo institucional seja alcançado entre MAAYA e SIL.

### 5.2.2 O caso de L2

O método escolhido neste estudo para L2 não consegue diferenciar por país e, portanto, a correção fina obtida graças à ponderação não é aplicada. Aplicar a mesma taxa de crescimento de L1 a L1 + L2 aos resultados dos cálculos baseados em L1, na verdade, implica a mesma suposição simplificadora que a transformação dos dados do país em língua (dados da Internet em cada país têm o mesmo comportamento para todas as línguas faladas neste país), mas desta vez, a implicação não está dentro de um país, mas entre todos os países do mundo. A hipótese de simplificação, nesse caso, é que a taxa de conexão à Internet das pessoas em uma determinada língua materna também é idêntica para falantes dessa língua como L2, independentemente do país de residência.

No caso imaginário de uma língua L que teria 100 milhões de falantes L1 em ​​um país P1, com uma porcentagem de conectividade com a Internet de 80%, e os 100 milhões de falantes de segunda língua todos residem em um país P2 com uma taxa conectividade de 40%, então o erro do método é superestimar o número de alto-falantes L1 + L2 conectados em 50%.

Nc = número calculado de alto-falantes conectados

Nr = número real de alto-falantes conectados

Nc (L1) = Nr (L1) = 100 milhões x 80% = 80M

Nc (L1 + L2) = 80M x 2 = 160M

Nr (L1 + L2) = 100 M x 80% + 100 M x 40% = 120 M

Tamanho do erro = 40M ou 50% extra

É claro que esta é uma situação totalmente imaginária e rebuscada e só é exposta por necessidade pedagógica, uma vez que a realidade não tem diferenças tão extremas.

Seria possível, no entanto, superar esta desvantagem potencial, utilizando o método dos quartis desde que os dados correspondentes estejam disponíveis, por exemplo:

* P0: a proporção de L2 que reside em países onde a taxa de conexão é inferior a 15% da população: primeiro quartil (Q1 = pontuação mais baixa de pessoas conectadas: 0,08 com os valores atuais)
* P1: a proporção que reside em países onde a taxa é superior a 15%, mas inferior a 35%, segundo quartil (Q2 = 0,34)
* P2: a proporção que reside em países onde a taxa é superior a 35%, mas inferior a 65: terceiro quartil (Q3 = 0,56)
* P3: a proporção que reside em países onde a taxa é superior a 65%, mas inferior a 85: quarto quartil (Q4 = 0,76)
* P4: proporção de residentes em países onde a taxa é superior a 85%: último quartil (a nota mais alta: Q5 = 1).

e a fórmula para recalcular a taxa de crescimento a ser aplicada aos indicadores seria:

i = 4

Tc (j) = (T (j) x PL1 (j)) / Σ Pi x Qi

i = 0

Onde

T (j) é a relação L1 + L2 / L1 para a língua j

Tc (j) é a mesma relação, mas corrigida pelo método dos quartis a ser aplicado aos resultados para que L1 passe para os resultados L1 + L2

PL1 (j) é a % de pessoas conectadas para L1 (j).

No exemplo francês, com L2 = 153 485 770, T = 3 e PL1 = 0,81, se a distribuição fosse de 20% para cada quartil, o valor corrigido a ser aplicado seria:

Tr (francês) = 3 x 0,81 / (20% x 0,08 + 20% x 0,34 + 20% x 0,56 + 20% x 0,76 + 20% x 1 = 0,548)

Nesse caso, em vez de 3, seria usado um fator de 2, o que possivelmente está mais próximo da realidade.

Claro, também é possível simplificar esse método e dividi-lo apenas em duas categorias, altos e baixos, tudo realmente depende de quão claro podemos da proporção da população L2 entre essas duas categorias.

A qualidade dos resultados certamente seria melhor assegurada pela aplicação deste método simples. Embora este método mais refinado seja implementado (mas condicionado pelo fornecimento de dados mais detalhados sobre as populações L2), é razoável considerar que o método atual pode causar uma superestimação dos resultados de inglês e francês que têm grandes populações L2 em países com conectividade à Internet. bastante inferior aos Estados Unidos (ou Reino Unido) e França (ou Bélgica), respectivamente.

### 5.2.3 Reduzindo o número de línguas

O princípio fundamental da independência do processo em relação ao número de línguas tratadas foi verificado em provas, o que garante resultados idênticos seja qual for o número de línguas selecionadas (claro que sempre cuidando do cálculo da linha "línguas restantes"). Portanto, é teoricamente possível tratar todas as línguas com o método. Na prática, este é um volume computacional impressionante (7500 x 200 = 1 milhão e meio de células do Excel) e o tamanho do arquivo é muito pesado para um manuseio rápido (mais de 100 MB) e, portanto, representa um custo enorme. marginal para benefício marginal zero porque apenas 500 línguas em 7500 têm uma existência virtual. Por outro lado, a simplificação das hipóteses invalidaria os resultados para línguas com números de falantes não medidos em milhões. Depois de vários testes com valores diferentes, a seleção foi finalmente fixada com a lista de línguas com mais de 5 milhões de falantes (140 línguas) naquela primeira edição, restando a tarefa futura de tentar diminuir esse limiar na direção de línguas com mais de 1 milhão de falantes.

### 5.2.4 Verificando a invariância dos resultados com o número de línguas

Claro, era necessário garantir que, como dita a intuição, as operações aritméticas realizadas para transformar dados por país em dados por língua produzissem os mesmos resultados, independentemente do número de línguas processados. Isso foi verificado várias vezes para mudar de 7500 para 390, 100, 28, 85 línguas, a lista de línguas de mais de 10 milhões de falantes e, finalmente, para mais de 5 milhões de falantes. Esses testes às vezes permitiam a descoberta de erros nos dados e representavam um bom investimento na confiabilidade dos resultados.

## 5.3 Países

Após algumas iterações, a lista final de países a considerar foi determinada de forma simplificada: a eliminação de alguns países com muitos poucos habitantes e / ou porque a UIT não fornece dados e o agrupamento nos principais países dos estados satélites.

Os dados da UIT foram modificados para refletir mudanças na lista de países retidos, vinculando os dados da população e modificando os dados da porcentagem de pessoas conectadas em proporção às suas respectivas populações.

Um bom exemplo é a China, que possui um percentual de conexão à Internet muito menor do que Hong Kong ou Macau e, portanto, a integração desses dois estados nos dados da China será feita na proporção de suas respectivas taxas. Alguns países foram retirados da lista (como Coréia do Norte, San Marino ou as Malvinas), outros aderiram a outro país (por exemplo, as Ilhas Faroé e a Groenlândia na Dinamarca). O Reino Unido é o país que recebeu o maior número de territórios anexados.

Para alguns grandes países a UIT não fornece dados para usuários conectados à Internet e os falantes desses países não foram integrados ao modelo: é o caso da Coréia do Norte (que não está muito conectada à Internet), Saara Ocidental, Kosovo e Sudão do Sul.

Por fim, há uma lista de pequenos estados que não foram levados em consideração, seja porque a UIT não fornece dados (em itálico) ou porque sua população é muito limitada: *Bonaire, Santo Eustáquio e Saba, Curaçao, Ilha Christmas*, Ilhas Cocos, Ilha de Man, Ilha Norfolk, Pitcairn, Falklands, *Ilhas Marianas do Norte, Mayotte*, *Nauru, Palau*, *San Bartolome*, San Marino, San Martín, *São Pedro e Miquelão*, a ilha *Svalbard e Jan Mayen, Território Britânico do Oceano Índico, Cidade do Vaticano.*

Tabela  : Países : afiliação de territórios

|  |  |
| --- | --- |
| **País de afiliação** | **LISTA DE PAÍSES RELACIONADOS** |
| China | Hong Kong e Macau |
| Estados Unidos | Guam, Ilhas Virgens dos EUA, Porto Rico e Samoa Americana |
| Dinamarca | Groenlândia e Ilhas Faroe |
| França | Guadalupe, Guiana, Martinica, Nova Caledônia, Polinésia Francesa, Reunião e Wallis e Futuna |
| Nova Zelândia | Ilhas Cook, Niue e Tokelau |
| Países Baixos | Aruba, Antilhas Holandesas |
| Reino Unido | Anguila, Antígua e Barbuda, Ascensão, Bermudas, Gibraltar, Guernsey, Ilhas Cayman, Ilhas Virgens, Jersey, Montserrat, Santa Helena, São Cristóvão e Névis e Turks e Caicos |

## 5.4 Fontes

### 5.4.1 Conceitos básicos

Os três princípios a seguir foram geralmente respeitados:

1. sempre vá para a primeira fonte[[39]](#footnote-39);
2. sempre prefira as últimas datas das fontes;
3. não assume responsabilidade pela correção das aproximações detectadas nas fontes (exceto conforme descrito abaixo).

O primeiro princípio obviamente se aplica a fontes que fornecem porcentagens mundiais por país, não a fontes individuais para cada país (que ainda é o trabalho da primeira fonte, aquela que compõe os números mundiais). A celebração do primeiro princípio frequentemente implica o segundo princípio, mas deve ficar claro que os cálculos podem combinar datas diferentes por país dentro de uma cobertura global, mas também entre os elementos colocados em equação (ver 5.4.3 A questão das datas).

### 5.4.2 Exceções aos princípios básicos

No entanto, existem alguns casos em que o aprimoramento das fontes primárias se impõe por serem dados críticos para todo o processo: é o caso dos dados demográficos por países e das pessoas conectadas por país.

Outra exceção foi feita no caso de dados de assinatura oferecidos pelo Statista para Netflix. São duas séries: os dados atuais e uma projeção para 2021. A descoberta de uma redução, na projeção para o ano de 2021, do número de assinantes nos EUA, país de longe líder nesta aplicação, permitiu a verificação e a descoberta de que os dados para os Estados Unidos não foram documentados na projeção para 2021. A técnica de extrapolação aplicada a essa categoria de fonte foi capaz de limitar os danos, mas o resultado ainda era insatisfatório com uma diminuição nas assinaturas entre 2016 e 2021 para os Estados Unidos. Isso deve levar à remoção desta fonte, exceto ao obter dados confiáveis ​​de outra fonte para o crescimento de assinantes nos EUA neste mercado,

### 5.4.3 A questão das datas

A questão das datas de origem é complexa. É comum que algumas fontes de dados por país utilizem datas diferentes para países diferentes, situação obrigatória devido às diferenças no nível de atualização das fontes de cada país. A presença na mesma operação aritmética das fontes de dados com datas diferentes não representa um obstáculo, embora possa ter um impacto marginal nos resultados. Portanto, a fonte da UIT para a taxa de conectividade à Internet por país é 2015, mas os dados demográficos por país podem ser de datas diferentes; deve-se estar atento a essa situação ao processar os dados, mas isso não os invalida.

No caso de fontes que combinam várias fontes externas para construir indicadores compostos, infelizmente é comum que não usem os dados mais atualizados. O compromisso de atualizar essas fontes compostas está além do escopo deste estudo. É o caso do indicador "índice de serviços governamentais online" que WebIndex usa para criar indicadores compostos: dá à França a primeira classificação com uma pontuação de 100. A primeira fonte foi pesquisada depois e a pesquisa revelou que WebIndex usou uma fonte datado de 2013, enquanto a primeira fonte[[40]](#footnote-40) fornece os dados de 2016, onde a França passa para o quinto lugar com uma pontuação de 94,20. Os resultados não refletem essa evolução e isso tem efeitos totalmente marginais nos resultados.

### 5.4.4 A questão do significado da transformação país > língua

Se o significado dessa transformação é intuitivamente óbvio para os dados por país, pode ser mais difícil de entender, até mesmo incompreensível, sobre percentuais que não são expressos por país.

O exemplo que faz mais sentido é o número de usuários da Internet para cada país. Se esses dados forem expressos em termos de quantidade por país ou por porcentagem de pessoas conectadas por país (esta é a primeira fonte oferecida pela UIT) ou como porcentagem de usuários de Internet em um país em relação ao total mundial de usuários na Internet, a aritmética simples permite ir de um valor a outro. O produto matricial (língua x país) com os valores por país, portanto, permite com uma simples aritmética, obter o número de internautas para cada língua (ou a porcentagem de pessoas conectadas mundialmente para cada língua, dependendo do modo de cálculo selecionado).

No caso de dados expressos em percentagens nacionais, por vezes é muito difícil dar sentido à transformação pela língua ou mesmo impossível encontrar uma técnica de extrapolação para este micro-indicador (os dois pontos provavelmente estão relacionados). Nestes casos, a opção escolhida foi abandonar esta transformação e, portanto, a integração deste micro-indicador na produção de indicadores. É o caso, por exemplo, do número de pedidos por país de eliminação de dados da Google (o direito ao esquecimento). A extrapolação não é possível porque a informação está relacionada à cultura jurídica do país (e à existência ou não de uma lei) e não está diretamente relacionada ao percentual de pessoas online no país. A transformação para a língua deste indicador, por outro lado, causa confusão, uma vez que se pode deduzir que o apagamento dos dados é feito em relação à língua apresentada no cálculo, quando é a língua oficial do país. Se quiséssemos dar um significado estrito e consistente para a transformação para língua, neste caso específico, teríamos que expressar: "a porcentagem de falantes de uma determinada língua que solicitam exclusões ao Google na língua nacional de seu país de residência" ... isso não é de interesse óbvio nem fácil de entender!

### 5.4.5 Limitações devido a fontes

As limitações devidas às fontes incluem, em primeiro lugar, é claro, as limitações e possíveis vieses dos dados fornecidos pela fonte principal. No caso do indicador de pessoas conectadas à Internet por país, a análise do método utilizado pela fonte[[41]](#footnote-41), amostra:

* É o indicador HH7: proporção de indivíduos que usam a Internet com descrição precisa do método e suas limitações[[42]](#footnote-42);
* os valores foram estabelecidos a partir de inquéritos populacionais;
* a definição estabelece que as pessoas que se conectam à Internet são contadas independentemente do dispositivo que utilizam (que não é necessariamente um computador e também pode ser um celular, um tablet, um assistente digital pessoal, um console de jogos, TV digital, etc.) e que o acesso pode ser feito através da rede fixa ou da rede móvel;
* o padrão de contabilidade adotado é o de pessoas de 16 a 75 anos que se conectaram pelo menos uma vez nos últimos 3 meses.

Uma análise mais detalhada mostra que:

* aproximadamente 60% das fontes são estimativas da própria UIT e 15% derivam do Eurostat, utilizando exatamente as mesmas definições e critérios[[43]](#footnote-43);
* para o resto dos países, os dados são fornecidos por diferentes autoridades dependendo do país. Essas autoridades geralmente usam critérios diferentes. Portanto, a faixa etária pode começar aos 3, 5, 6, 7, 10, 12, 15 ou 20 anos e às vezes não há limite de idade;
* para alguns países os dados dos Estados associados podem ser contados com o país e em outros casos independentes[[44]](#footnote-44).

UIT é de longe a fonte mais confiável para este tipo de dados, e outras fontes costumam usá-lo sem citar e, às vezes, por meio de alterações não documentadas. Esses dados são, por outro lado, um elemento essencial em todo o processo do modelo exposto e grande cuidado foi tomado para considerar cuidadosamente todos os fatores, em particular a distribuição de números exatos no caso de países com territórios associados.

Quais são os possíveis vieses dos dados da UIT? Por um lado, os países europeus, para os quais a UIT utilizou dados do Eurostat, e os países nos quais a UIT divulgou as suas próprias estimativas, podem ser considerados dados particularmente fiáveis. Por outro lado, pode-se temer uma leve superestimação de alguns dos países onde as autoridades locais oferecem os dados, seja porque os critérios de definição são mais amplos, como já mencionado, ou também porque alguns desses países poderiam fizeram uma estimativa generosa para mostrar a boa validade de suas políticas contra a exclusão digital. É bom saber desse viés potencial, mas não é importante e geralmente pode ser considerado bastante marginal.

### 5.4.6 Potencial viés Alexa.com e W3Techs

A maioria dos micro-indicadores são dados de tráfego para sites onde Alexa.com foi usado como fonte. Quais são os possíveis vieses desta fonte?

Alexa oferece estatísticas de tráfego de aplicativos da web por meio de uma barra de ferramentas que supostamente uma grande amostra de usuários da Internet[[45]](#footnote-45) concordou em instalar no seu navegador. Alexa é informada das visitas aos sites por meio do software que é acionado a partir da barra de ferramentas e, com esses dados, faz um ranking dos 25 milhões dos sites mais populares. Alexa, como a maioria das empresas privadas na Internet, não é transparente sobre sua metodologia e os vieses que podem resultar; em particular, infelizmente não existem estatísticas sobre as barras de ferramentas instaladas que permitam conhecer o seu total, muito menos a sua distribuição por país ou língua.

Claro, Alexa só oferece estimativas de dados de tráfego para sites que estão no topo do ranking. A seleção de sites para o estudo foi feita com uma boa posição neste ranking (normalmente com um ranking Alexa inferior a 10.000, exceto para alguns tópicos menos populares para os quais foram selecionados rankings de até 50.000). As tendências que podem surgir do método Alexa são:

* concentração de atenção nos sites mais visitados (por definição);
* uma subestimação em países onde a barra de ferramentas Alexa não é muito baixada (especialmente uma forte subestimação do tráfego doméstico chinês, já que 80% dos sites não são acessíveis pelo sistema de nomes da ICANN[[46]](#footnote-46));
* uma superestimação relativa do tráfego dos países que instalaram a barra Alexa em maior proporção (com toda a probabilidade, especialmente os países ocidentais e mais particularmente os Estados Unidos[[47]](#footnote-47)).

Quais são as implicações desses vieses nos resultados do estudo?

Suspeita-se que haja uma super-representação dos países ocidentais e depois das línguas ocidentais das demais línguas e uma super-representação do inglês em comparação com todas as línguas, em uma porcentagem muito alta em comparação com as línguas não ocidentais.

Um estudo comparativo dos dados de assinatura do Facebook, Twitter e Linkedin confirmou essa suspeita.

O estudo comparou os dados em termos de assinaturas desses aplicativos com os dados em termos de tráfego fornecidos pela Alexa. As diferenças entre os dois resultados (após ter os valores alterados em porcentagem) são relativamente grandes e, previsivelmente, geralmente com dados de tráfego Alexa extremamente altos em comparação com o número de assinantes no caso das línguas ocidentais. As diferenças entre a porcentagem de assinantes e o tráfego estão se mostrando muito importantes na maioria dos países não ocidentais.

Nota-se que o português se comporta como línguas não ocidentais neste tipo de comparação, e o motivo ainda não foi elucidado (haveria um motivo específico no Brasil que tornaria o download da barra de ferramentas do Alexa um problema ou é que o tempo de conexão aos diferentes serviços é bem menor devido aos preços, ou talvez seja o que os brasileiros amam tanto nas redes sociais que costumam ter várias contas?). Este cheque evidenciou as qualificações favoráveis ​​da francês, não é o caso do espanhol e menos ainda do finlandês (o último caso provavelmente mostra que a técnica de extrapolação nem sempre é capaz de fornecer corretamente os níveis reais).

A tabela a seguir mostra o resultado da divisão dos valores propostos pelo Alexa (em termos de tráfego) pelos valores obtidos no Statista (em termos de assinaturas), classificados em ordem decrescente, a dicotomia entre as línguas ocidentais (exceto o português) e as línguas asiáticas aparecem muito claramente. Um fator de 1 indicaria a equivalência dos resultados, o resultado é superior a 1 quando os valores propostos pelo Alexa (tráfego) são superiores aos do Statista (assinaturas), denotando um viés favorável do Alexa.

A explicação para as diferenças entre os dados de comparação deve-se à combinação de dois fatores:

1) O acima mencionado: Alexa favorecerá significativamente os países ocidentais como um reflexo natural do layout da barra de ferramentas.

2) Mas também um fator adicional que pode desempenhar um papel está relacionado ao tempo médio de permanência dos usuários em sessão, que, por razões de preço, geralmente deve ser maior nos países industrializados.

No entanto, o segundo fator não pode explicar sozinho o tamanho das diferenças observadas e este é certamente o primeiro fator, o viés de Alexa.[[48]](#footnote-48), que é o principal responsável.

Tabela : Relações tráfego / assinante

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **FACEBOOK** | **TWITTER** | **LINKEDIN** | **MÉDIA** |
| Francês | 2,81 | 4,40 | 2,49 | **3,23** |
| Inglês | 3,11 | 2,65 | 1,63 | **2,46** |
| Tailandês | 2,35 | 1,23 | 2,38 | **1,99** |
| Catalão | 1,69 | 1,82 | 2,25 | **1,92** |
| Japonês | 1,41 | 1,65 | 2,33 | **1,79** |
| Norueguês | 3,57 | 1,46 | 0,22 | **1,75** |
| Sueco | 2,36 | 2.07 | 0,74 | **1,72** |
| Dinamarquês | 3,18 | 0,34 | 1,22 | **1,58** |
| Grego | 2,42 | 1,48 | 0,45 | **1,45** |
| Alemão | 1,19 | 1,57 | 1,33 | **1,36** |
| Polonês | 2,14 | 0,75 | 0,85 | **1,25** |
| Húngaro | 1,59 | 0,35 | 0,97 | **0,97** |
| Russo | 0,68 | 0,48 | 1,33 | **0,83** |
| Tcheco | 1,76 | 0,32 | 0,39 | **0,82** |
| Holandês | 1,54 | 0,34 | 0,59 | **0,82** |
| Italiano | 0,85 | 1,17 | 0,41 | **0,81** |
| Coreano | 0,63 | 0,81 | 0,95 | **0,80** |
| Urdu | 0,41 | 0,44 | 1,46 | **0,77** |
| Romena | 1,21 | 0,43 | 0,51 | **0,72** |
| Espanhol | 0,80 | 0,72 | 0,62 | **0,71** |
| Turco | 0,28 | 0,88 | 0,52 | **0,56** |
| Árabe | 0,51 | 0,48 | 0,57 | **0,52** |
| Lahnda | 0,52 | 0,28 | 0,67 | **0,49** |
| Tâmil | 0,20 | 0,18 | 0,99 | **0,46** |
| Marati | 0,16 | 0,14 | 1,00 | **0,43** |
| Canarim | 0,15 | 0,13 | 0,96 | **0,41** |
| Chinês | 0,55 | 0,23 | 0,44 | **0,41** |
| Bengali | 0,27 | 0,25 | 0,68 | **0,40** |
| Hindi | 0,21 | 0,19 | 0,74 | **0,38** |
| Português | 0,33 | 0,19 | 0,27 | **0,26** |
| Cebuano | 0,10 | 0,12 | 0,57 | **0,26** |
| Javanês | 0,07 | 0,05 | 0,30 | **0,14** |
| Sudanês | 0,07 | 0,05 | 0,29 | **0,14** |

A conclusão geral dessas comparações é que os resultados do macro-indicador de tráfego exageram os valores para as línguas ocidentais e, nessas línguas, os do inglês[[49]](#footnote-49). Essa realidade deve ser colocada em perspectiva ao usar os resultados. Qual é o tamanho do viés contra as línguas não ocidentais?

Para se ter uma ideia, temos que verificar se a W3Techs está usando sistematicamente um algoritmo de identificação de linguagem aplicado aos 10 milhões de sites principais de acordo com Alexa e irá refletir naturalmente esses vieses em seu resultado[[50]](#footnote-50). Na realidade, os W3Techs não só refletem essas tendências, mas também às vezes amplificam ou as distorcem conforme outros fatores se somam:

* algoritmos de reconhecimento de língua também são tendenciosos em favor do inglês, embora isso possa ser um fator marginal de cerca de 10%[[51]](#footnote-51);
* a escolha da detecção da língua na página inicial dos sites, muitas vezes multilíngue, com o inglês como língua adicional, aumentará substancialmente a pontuação em inglês (poderia levar a uma superestimação da ordem de 1/3);
* a consideração ineficaz do multi-linguismo.

Determinar qual dos fatores é mais influente é um exercício muito difícil. No entanto, existe um "truque" que permitirá que você entenda a situação até certo ponto. Em estudos realizados pela FUNREDES e União Latina, no período 1998-2007, o macro-indicador de produtividade (relação entre o percentual de conteúdo online para uma língua e o percentual de pessoas conectadas por essa língua) foi estudado e monitorado durante um longo período. Este indicador da propensão para criar conteúdo por língua mostrou alguma estabilidade e consistência ao longo do tempo com valores muito claramente centrados em 1. As diferenças eram da ordem de mais ou menos 50% dependendo se a língua era o objeto de um grande esforço de produção de conteúdo (1.5) ou, pelo contrário, se havia fraqueza nessa direção (0,5).

Estes são os valores deste indicador em 2005 e 2008 (fonte FUNREDES[[52]](#footnote-52)):

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Francês** | **Inglês** | **Espanhol** | **Português** | **Alemão** | **Italiano** | **Romena** | **Catalão** |
| **2005** | 1,14 | 1,57 | 0,66 | 0,55 | 1.06 | 0,81 | 0,70 |  |
| **2008** | 0,87 | 1,42 | 0,43 | 0,34 | 1,16 | 0,98 | 0,66 | 0,74 |

Vê-se que o indicador permanece na janela de 0,3 a 1,5. Aqui fazemos a mesma comparação, usando os dados UIT (convertidos em termos de língua) e os W3Techs para conteúdo.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Francês** | **Inglês** | **Espanhol** | **Português** | **Alemão** | **Italiano** | **Romena** | **Catalão** |
| **2017** | 0,71 | 2,34 | 0,53 | 0,62 | 1,72 | 2,41 | 0,92 | 0,44 |

Os resultados para inglês, alemão e italiano não são confiáveis. Eles mostram uma grande superestimação dos valores atribuídos pela W3Techs aos conteúdos nessas línguas. As próximas duas tabelas mostram os mesmos indicadores calculados da W3Techs para uma amostra de outras línguas.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Chinês** | **Hindi** | **Árabe** | **Malaio** | **Urdu** | **Bengali** | **Coreano** | **Vietnamita** | **Tailandês** |
| **2017** | 0,09 | 0,02 | 0,19 | 0,23 | 0,05 | 0,07 | 0,57 | 0,45 | 0,36 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Russo** | **Japonês** | **Polonês** | **Hebraico** | **Húngaro** | **Sueco** | **Turco** | **Ucraniano** | **Tcheco** |
| **2017** | 1,27 | 1,60 | 2,00 | 1,65 | 1,59 | 2,2 | 1,58 | 0,17 | 3,22 |

Vale ressaltar que a subestimação de certas línguas atinge proporções gigantescas (chinês, hindi, urdu, bengali), o que invalida completamente os números apresentados pela W3Techs para essas línguas, ou muito grandes para serem críveis (árabe, malaio, ucraniano, tailandês), enquanto outros estão claramente superestimados (italiano, tcheco, sueco, polonês). Valores abaixo de 0,1 mostram que o déficit é enorme e absurdo; Valores entre 0,1 e 0,3 mostram que é enorme e não é confiável, enfim, valores acima de 2 mostram um superávit difícil de acreditar.

Essas comparações mostram que a W3Techs não é confiável como fonte de medição de conteúdo. Isso não é para duvidar da capacidade da empresa de realizar essas ações, mas para entender que dois vieses cumulativos estão envolvidos nos resultados:

* O viés específico da amostra Alexa que é usado pela W3Techs para apoiar as medidas.
* O viés do método W3Techs que claramente favorece o inglês ao medir as “home pages” e não levar em conta o multi-linguismo.

A análise dos dados mostra que se existe um viés notável inerente à W3Techs e deve inflar os resultados do inglês em uma certa proporção (não levando em conta sites multilíngues e trabalhando nas primeiras páginas, pode causar uma superestimação do inglês da ordem de 30%), o principal viés, de longe, é induzido por Alexa. Até certo ponto, W3Techs é uma revelação do viés Alexa! Na verdade, parece que há países cuja presença no banco de dados Alexa é subestimada por um fator de cerca de 500%: China e Índia juntas respondem por mais de 1 bilhão de pessoas conectadas à Internet (ou seja, quase um terço de Internautas), como poderiam ter juntos apenas 6,6% dos conteúdos da Internet, conforme indicado pela W3Techs?

### 5.4.7 Correção de vieses W3Techs

Existe uma técnica que foi usada com frequência nos primeiros estudos da FUNREDES / Union Latina, no período 1998-2007, para determinar o peso do inglês[[53]](#footnote-53). Era tentar obter resultados coerentes dos indicadores para o resto das línguas e a partir daí todos os resultados foram fixados. Se levarmos em consideração as primeiras 15 línguas classificadas neste estudo, a porcentagem de pessoas conectadas das demais línguas é de aproximadamente 35%. Para este conjunto de línguas, a W3Techs atribui um resto igual a pouco mais de 10% do conteúdo mundial, o que não é consistente. Partindo do pressuposto lógico de uma forte superestimação do conteúdo em inglês por W3Techs (medido em 53%), a forte subestimação de algumas línguas asiáticas e do indicador para o resto das línguas, podemos tentar fazer uma aproximação razoável ... embora puramente especulativo, sobre qual é o valor mais provável para a distribuição de conteúdo por língua na Internet;

A tabela a seguir fornece os seguintes dados por coluna:

1: Indicador de conteúdo

2: Porcentagem de pessoas conectadas

3: Figura proposta para o conteúdo por W3Techs

4: Relacionamento entre os dados da W3Techs e as pessoas conectadas (3/2 :calculado como item 3 dividido por item 2)

5: Capacidade medida (1/7)

6: Capacidade especulativa (8/7)

7: Porcentagem da população mundial

8: Especulação sobre a porcentagem de conteúdo

9: Produtividade de conteúdo resultante de especulação (8/2)

10: Especulação sobre o percentual de conteúdo apresentado na base de 100%

Tabela : Classificação especulativa

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|  |  |  | % |  | **W3TECHS /CONNECT** |  |  |  | **CONTEUDOS** | **PROD..** | **CONTEUDOS** |
|  |  | **PODER** | **CONNECT.** |  |  | **CAPACID..** | **CAPACID..** | **POP.** | **ESPEC.** | **CONT** | **ESPEC.** |
|  |  | **L1 + L2** | **L1 + L2** | **W3T** | **L1 + L2** | **L1 + L2** | **Espec.** | **MUNDO** | **L1 + L2** | **L1 + L2** | **/ 100%** |
| 1 | **INGLÊS** | 0,404 | 0,226 | 0,519 | 2.299 | 3,74 | 3,33 | 0,108 | **0,360** | 1,59 | 29% |
| 2 | **CHINÊS** | 0,140 | 0,209 | 0,020 | 0,096 | 0,92 | 1,18 | 0152 | **0,180** | 0,86 | 14% |
| 3 | **ESPANHOL** | 0,093 | 0,093 | 0,051 | 0,551 | 1,56 | 1,35 | 0,059 | **0,080** | 0,86 | 6% |
| 4 | **FRANCÊS** | 0076 | 0,056 | 0,041 | 0,726 | 2,93 | 2,31 | 0,026 | **0,060** | 1.06 | 5% |
| 5 | **RUSSO** | 0,047 | 0,050 | 0,065 | 1.288 | 1,71 | 1,45 | 0028 | **0,040** | 0,79 | 3,2% |
| 6 | **PORTUGUÊS** | 0,038 | 0,041 | 0,026 | 0,641 | 1,49 | 1,36 | 0,026 | **0,035** | 0,86 | 2,8% |
| 6 | **ALEMÃO** | 0,046 | 0,032 | 0,055 | 1.723 | 3,31 | 2,53 | 0,014 | **0,035** | 1,10 | 2,8% |
| 6 | **JAPONÊS** | 0,033 | 0,034 | 0,056 | 1.633 | 2,34 | 2,47 | 0,014 | **0,035** | 1.02 | 2,8% |
| 9 | **HINDI** | 0,027 | 0,039 | 0,001 | 0,023 | 0,48 | 0,54 | 0,056 | **0,030** | 0,76 | 2,4% |
| 9 | **ÁRABE** | 0,032 | 0,042 | 0,007 | 0,165 | 0,86 | 0,79 | 0,038 | **0,030** | 0,71 | 2,4% |
| 11 | **MALAIO** | 0,021 | 0,027 | 0,060 | 2.262 | 0,80 | 0,77 | 0,026 | **0,020** | 0,75 | 1,6% |
| 12 | **COREANO** | 0,013 | 0,014 | 0,009 | 0,637 | 1,41 | 1,69 | 0,009 | **0,015** | 1.06 | 1,2% |
| 12 | **URDU** | 0,014 | 0,020 | 0,001 | 0,044 | 0,51 | 0,55 | 0,027 | **0,015** | 0,74 | 1,2% |
| 14 | **ITALIANO** | 0,016 | 0,009 | 0,023 | 2.515 | 3,22 | 1,97 | 0,005 | **0,010** | 1.09 | 0,8% |
| 15 | **BENGALI** | 0,008 | 0,013 | 0,001 | 0,069 | 0,29 | 0,37 | 0,027 | **0,010** | 0,76 | 0,8% |
| 16 | **POLONÊS** | 0,012 | 0,009 | 0,017 | 1.996 | 2,55 | 1,94 | 0,005 | **0,009** | 1.06 | 0,7% |
|  | RESTO | 0,230 | 0,335 | 0,048 | 0,144 | 0,48 | 0,45 | 0,630 | **0,286** | **0,85** | 2,3% |
|  | TOTAL | 1.250 | 1.250 | 1.000 |  |  |  | 1,2500 | 1,2500 |  | 100% |

Vemos nesta tabela que nas cifras especulativas a produtividade do indicador de conteúdo foi reduzida a valores consistentes para o resto das línguas. A especulação se baseia justamente em um equilíbrio razoável entre a produtividade do conteúdo para as demais línguas e a manutenção das línguas abrangidas em uma janela razoável entre 0,7 e 1,3 (com exceção do inglês, que pode chegar a um valor mais alto). Há uma sensibilidade bastante forte desses dois fatores; portanto, se passarmos o inglês em 30% (sem alterar o resto), a produtividade do resto das línguas passa a ser superior a 1, o que não corresponde ao que se pode esperar; se elevarmos o inglês para 40%, a produtividade do inglês chegará a 1,8. Um valor especulativo de um terço da proporção do conteúdo em inglês daria uma produtividade para o restante acima de 0,9, o que é muito importante. Portanto, a especulação foi ajustada a este valor para o inglês: 36%.

A escassez de fontes de informação sobre línguas na Internet criou uma situação alarmante em que as únicas fontes existentes não são confiáveis, mas são referenciadas por um grande número de interlocutores sérios que constroem inferências distorcidas por dados enviesados ​​com superestimação do papel de inglês na internet**[[54]](#footnote-54)**!

### 5.4.8 Limitações / vieses relacionados ao grau de localidade das fontes

Muitos usuários da Internet em países ocidentais podem pensar que aplicativos famosos como Wikipedia, Facebook, Google, Twitter ou YouTube são padrões internacionais cuja popularidade e uso são iguais em todos os países do mundo. A realidade é muito diferente: alguns países (como China ou Rússia) desenvolveram aplicativos alternativos que drenam grande parte do tráfego nacional potencial para esse conjunto de grandes aplicativos da Internet. É o caso, por exemplo, do Vkontakte que substitui o Facebook na Rússia ou do Badiu que substitui o Google na China. Se pudéssemos atribuir uma classificação qualitativa de "globalidade" (ou, ao contrário, grau de "localidade") para aplicações na Internet estaremos cientes que este critério pode variar significativamente e quase nunca atingirá o seu horizonte de 100%. Entre as aplicações mais "globais"[[55]](#footnote-55) provavelmente existe o Twitter e sem dúvida a Wikipedia (e não é por acaso que esta última é de longe a que apresenta maior diversidade linguística) e talvez o Google, com algumas ressalvas, mas este é menos o caso do Facebook, YouTube, Linkedin, mesmo que estes os aplicativos ainda são muito mais "globais" do que seus concorrentes asiáticos Vkontakte ou Badiu.

Nesse contexto, qual seria a validade da análise usando diferentes sites Alexa? Embora seja claro que cada indicador de país para uma dada aplicação ou área deve ser tomado com cuidado para evitar que a qualificação de um língua seja determinada por um alto grau de localidade dos sítios Web selecionados, não é menos verdade que a multiplicação de sítios (mais de 300) fornecem, graças à média truncada, um meio de neutralizar esses efeitos. A suavização da curva também é alcançada pela técnica de extrapolação que fornecerá valores para os países ausentes na fonte e que trabalha com critérios globais.

Nesse sentido, é interessante identificar os sites que são fortemente locais a uma língua na seleção e verificar 1) se a distribuição dessas línguas não está muito enviesada e 2) se a média truncada em 20% corrige o viés de seleção.

A tabela a seguir quantifica, para cada língua, o viés de localidade na amostra de sites usada para medir o tráfego.

Tabela : Distribuição de sites com forte localidade

|  |  |
| --- | --- |
| LÍNGUA | Número de sites |
| Alemão | 2 |
| Árabe | 2 |
| Chinês | 10 |
| Coreano | 3 |
| Espanhol | 9 |
| Francês | 12 |
| Hindi | 6 |
| Inglês | 25 |
| Português | 3 |
| Russo | 4 |

O Anexo IV lista os sites fortemente locais das principais línguas. Ressalta-se que a média truncada em 20% eliminará, da amostra de 316 sites, cerca de 30 das pontuações mais baixas e igual número das mais altas para cada língua, o que abrange amplamente os valores apresentados na tabela anterior.

A sensibilidade desse fator nos resultados ainda precisa ser verificada. Parece que a sensibilidade é muito alta se a média simples ou a mediana for usada. Por exemplo, o valor médio do indicador de tráfego para o francês é 10,31 e com esse valor o poder do macro-indicador do francês seria 8,07. Se retirarmos o site Fnac.com da lista abaixo da média (tráfego) cai para 10,00; se também eliminarmos Gameblog.com, a média cai para 9,69; se também eliminarmos Fun-mooc.fr, Openclassrooms.com, Qwant.com, Skyrock.com, Theses.fr e Viadeo.com, a média do indicador de tráfego cai para 8,09 e o poder do macro-indicador cai para 7,70. Esta análise mostra que nem a média nem a mediana são eficazes e é necessário descartar valores extremos. Esta análise conduziu à solução finalmente adotada para o indicador de tráfego: a média truncada em 20%.

Essa situação afeta ainda mais o esforço de diferenciação temática. Por este motivo, esta abordagem só se aplicará aos dados de tráfego mais numerosos, não sendo considerados os resultados no caso de temáticas com muito poucos micro-indicadores, uma vez que refletiria apenas o viés de seleção dos sítios.

### 5.4.9 Sobre o princípio de ponderação

Embora o princípio de ponderação seja fácil de entender para conceitos como conteúdo, usos, tráfego e principalmente pessoas conectadas, ele pode se tornar bastante pouco claro ou misterioso, quando aplicado a avaliações de 0 a 100, como no caso de índices ou percentagens para um total de aplicações, como no caso das interfaces. Isso merece uma explicação.

Por qual mecanismo uma pontuação de 0 a 100 atribuída a uma língua como resultado de uma avaliação de um país em um determinado critério pode ser convertida em uma porcentagem mundial e qual é o significado desse novo valor? Por qual mecanismo uma porcentagem que representa o número de vezes que uma língua está presente em 24 interfaces de aplicativo pode ser convertida em uma porcentagem mundial e qual é o significado dessa porcentagem?

O mecanismo e o significado são os mesmos em ambos os casos. O mecanismo é a ponderação dos valores obtidos para os países pelo percentual de pessoas conectadas por país. Isso é calculado assim:

Pinter (j) = Tg x INTER (j) x UIT (j) / Po

i = L

com Po = ∑ UIT (i) x INTER (i) (fator de ponderação)

i = 1

Onde:

Pinter (j) é o peso do indicador de interface calculado para a língua j

Tg é a taxa global de multi-linguismo (para normalizar o total de L1 + L2)

INTER (j) é a porcentagem encontrada da língua j em relação às interfaces

ITU (i) é a porcentagem de pessoas com acesso à Internet para a língua i.

L é o número total de línguas (incluindo o resto).

O sentido desse percentual mundial é aproximar o valor do percentual de interfaces de uma determinada língua do percentual de pessoas conectadas à Internet para essa língua, com variações dependendo da distribuição dos percentuais de interface para as línguas. Se esse valor for exatamente a média ponderada das pontuações do índice (ou porcentagens para as interfaces), o resultado seria idêntico ao percentual de pessoas conectadas. Se esse valor for muito maior do que a média, o resultado aumentará; caso contrário, será reduzido. Deve-se entender que o efeito não depende apenas da nota (ou porcentagem), mas de toda a distribuição dos valores da porcentagem de pessoas conectadas de todas as línguas.

Para desenvolver uma abordagem mais intuitiva para esse mecanismo, é melhor simular a situação no caso mais simples de interfaces. O francês está presente em 23 dos 24 elementos medidos, dando-lhe uma porcentagem de 23/24 = 95,83%, o que elevará sua porcentagem mundial ponderada. Na verdade, este micro-indicador eleva o valor das pessoas conectadas de língua francesa (5,64%) a um valor de 7,37%. A simulação jogando com diferentes valores de porcentagem de presença nas interfaces permite sentir intuitivamente o que está acontecendo com o indicador com este método e como ele impacta os macro-indicadores.

Tabela : Simulação para interfaces

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Número de interfaces** | 0 | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 23 | 24 |
| **Porcentagem de**  **as interfaces** | 0% | 12,5% | 25% | 37,5% | cinquenta% | 62,5% | 75% | 87,5% | 95,8% | 100% |
| **Valor**  **calculado por ponderação** | 0% | 1,01% | 2,01% | 2,99% | 3,96% | 4,9% | 5,8% | 6,8% | 7,4% | 7,7% |
| **Cota de poder** | 6,4% | 6,6% | 6,7% | 6,9% | 7,1% | 7,2% | 7,4% | 7,5% | 7,6% | 7,7% |

Portanto, se não houvesse interface em francês entre as 24 da amostra, o indicador "interface" assumiria o valor de 0% e afetaria o valor do poder, que seria reduzido em 1,2%. Pelo contrário, se as 24 interfaces existissem em francês (100%), o indicador da interface ganha 0,3% e a potência ganha 0,1%. Vê-se que o valor de 5,6% dos francófonos conectados é muito próximo ao valor de 17 para o número da interface francesa e, portanto, existe o limiar de equilíbrio (que depende da distribuição desses valores) que irá mover o indicador para cima ou para baixo do valor de referência da porcentagem de falantes de francês conectados.

O mesmo exercício com o indicador governamental, por exemplo, dá os seguintes resultados.

Tabela : Simulação para índice

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Francês classificação no índice** | 0 | 25 | 50 | 75 | 96,1 | 100 |
| **Classificação francês** | 19 | 38 | 57 | 76 | 91 | 94 |
| **Porcentagem calculada para francês no índice e.Gov** | 0,18% | 3,5% | 5,1% | 6,7% | 8,0% | 8,3% |

O efeito sobre o valor do poder é bastante marginal, pois é um dos 5 micro-indicadores cuja média dá a marca para o indicador do índice; com pontuação de 50, a potência é reduzida para 7,5% e com pontuação de 0 a 7,4%. Nesse caso, o limite de equilíbrio é próximo a 57. Esses dois exemplos mostram intuitivamente como funciona o mecanismo de ponderação e como ele é sensível às condições iniciais.

# Conclusões e perspectivas

O monitoramento e o aprimoramento desse método pragmático de mensuração de linguagens na Internet requerem, no melhor dos casos:

* uma parceria entre MAAYA, Ethnologue e UIT para sincronizar uma observação anual após receber atualizações da UIT e Ethnologue;
* ampliação da lista de micro-indicadores e monitoramento contínuo da disponibilidade de novas opções;
* a tarefa de separar os dados L2 por país, ou pelo menos por grupo de países, a fim de melhorar a mudança para os indicadores L1 + L2;
* a investigação de micro-indicadores adicionais que poderiam servir à delicada questão de conteúdo;
* o aprofundamento dos controles e comparações do lado do modelo;
* a sistematização da captura de dados de fontes por país e língua do modelo, a fim de reduzir os riscos de erros;
* a migração da implementação do modelo Excel para APL, que seria a linguagem de programação ideal para essas operações matriciais.

Este método "artesanal" não anula de forma alguma as ambições do projeto DILINET de dar respostas a uma ampla gama de situações e áreas relacionadas com as linguagens no ciberespaço. No entanto, esse modelo fornece uma aproximação razoável a uma importante questão da Internet que foi totalmente deixada para o setor privado, com a enorme desvantagem de perder o controle de um grande problema societal. Nos últimos anos, a obtenção de dados sobre línguas na Internet foi o resultado da convergência de interesses privados, poucos transparentes sobre os métodos utilizados, obrigando os interessados ​​nesses dados a aceitá-los de olhos fechados.

# Bibliografia

[1] - D. Pimienta, D. Prado, "Étude sur la place des Langues de France dans l'Internet", Ministère de la culture de France, 2014.- [http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et- la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet](http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et-%20la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet)

[2] - "Le français dans l'Internet", Rapport 2014 "La langue française dans le monde", pp. 501-541, OIF, Nathan, 2014 -<http://www.francophonie.org/Rapports-Publications.html>

[3] - D. Pimienta, D. Prado, "Exploring the status of languages ​​of France in the Internet: methods and reflection as possible of approaches for --other groups of languages", Proc. of third international conference on linguistic and cultural diversity in cyberspace, IFAP/UNESCO,Yakutsk, 7/2014

[http://mcbs.ru/files/2015/hmao\_2015/yak\_mling\_2015.pdf#page=140](http://mcbs.ru/files/2015/hmao_2015/yak_mling_2015.pdf%23page=140)

[4] - D. Pimienta, Chapter “Language and content” in “Accelerating Development Using the Web: Empowering Poor and Marginalized Populations”, W3C Foundation, 2012

http://g3ict.org/download/p/fileId\_928/productId\_241

[5] - MAAYA, “NET.LANG: vers un cyberespace multilingue”, C&F Éditions, 2012 - <http://net-lang.net>

[6] - D. Pimienta, D. Prado, Á. Blanco, Douze années de mesure de la diversité linguistique sur l’Internet : bilan et perspectives, UNESCO, 2009 -

<http://unesdoc.unesco.org/images/0018/001870/187016f.pdf>.

[7] - D. Pimienta, Chapter “Target 9 contents and languages" in Measuring the Information Society: the ICT Development Index, ITU, ISBN 92-61-12831-9, 2009.

<http://www.itu.int/ITU-D/ict/publications/idi/2009/index.html>

[8] - J. Paolillo, D. Pimienta, D. Prado, et al. (2005), Mesurer la diversité linguistique sur Internet, UNESCO, 12/2005- <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>.

[9] - I. Suzuki, Y. Mikami, et al. (2002), “A Language and Character Set Determination Method Based on N-gram Statistics”, in ACM Transactions on Asian Language Information Processing, Vol.1, No.3, 9/2002, pp.270-279.

[10] - D. Pimienta, D. Prado, "Medición de la presencia de la lengua española en la Internet: métodos y resultados", en Revista Española de Documentación Científica 39 (3), julio-septiembre 2016, e141

ISSN-L:0210-0614. doi: <http://dx.doi.org/10.3989/redc.2016.3.1328>

# Anexo I. Lista de micro-indicadores

|  |
| --- |
| **Micro-indicadores USOS** |
| Penetração 3G |
| Assinaturas de celular |
| Videochamadas via dispositivo móvel |
| Videochamadas via dispositivo móvel - outra fonte |
| Chamadas de vídeo via dispositivo móvel - De acordo com Statista |
| Contas fixas de banda larga |
| Contas de banda larga móvel |
| Custo de banda larga |
| Famílias conectadas à Internet |
| ATMs Bitcoin |
| Infecções de PC |
| Telefones fixos |
| Mercado de comércio eletrônico |
| Mensagens instantâneas via celular |
| Caixa eletrônico |
| E-Wallets |
| Solicitações formais para o Google (dados) |
| Solicitações formais para o Google (excluído) |
| Servidores de Internet Seguros |
| Telefones inteligentes |
| Downloads do OpenOffice |
| Telefonia IP via dispositivo móvel |
| Tráfego móvel |
| Twitter via celular |
| Uso de servidores por sites |
| Usuários do Facebook |
| Usuários do Facebook de acordo com Owloo |
| Usuários do Facebook de acordo com Statista |
| Usuários do Linkedin |
| Usuários da Netflix - de acordo com a Netflix |
| Usuários da Netflix - de acordo com o Statista |
| Usuários de mídia social |
| Usuários de mídia social (projeção 2021) |
| Usuários do Twitter |
| Usuários do YouTube |
| Usuários do YouTube de acordo com Alexa |
| Usuários do YouTube de acordo com Statista |
| Usuários do Google |

|  |
| --- |
| **Micro-indicadores INDICES** |
| Indicador de acesso universal |
| indicador de e.gov |
| e. indicador de participação |
| WebIndex macro-indicador |
| Infraestrutura de macro-indicadores |

|  |
| --- |
| **Micro-indicadores INTERFACE** |
| Android |
| Apple |
| Línguas de tradução - Bing |
| Interface Cortana- |
| Dictionnary.com - línguas de tradução |
| DMOZ - linguagens de conteúdo |
| DuckDuckGo - Interface |
| Duolingo - línguas de tradução |
| Facebook - Interface |
| FreeTransductor - línguas de tradução |
| Google - Interface |
| Google Scholar - Interface |
| GoogleTranslate - línguas de tradução |
| Tradutor de mensagens instantâneas - línguas de tradução |
| Ios |
| Linux |
| On-Line - Língua de tradução |
| Preferências do Bing |
| Preferências do Google |
| Preferências do Yahoo |
| Linguagens de tradução reversa |
| SDL - línguas de tradução |
| Skype - Interface |
| Systran - línguas de tradução |
| Interface de telegrama |
| janelas |

|  |
| --- |
| **CONTEÚDO do micro-indicador** |
| Amazon USA - número de livros |
| W3Techs - conteúdo |
| Wikibooks - Interface |
| Wikibooks - Número de usuários |
| Wikibooks - número de itens |
| Wikibooks - número de editores |
| Wikipedia - Editores casuais |
| Wikipedia - Línguas dos artigos |
| Wikipedia - Editores confirmados |
| Wikipedia - Interface |
| Wikipedia - Profundidade |
| Wikinews - artigos |
| Wikinews - Interface |
| WikiQuote - Interface |
| WikiQuote - Artigos |
| Wikiversity - Interface |
| Wikiversity - Artigos |
| Wikcionário - Interface |
| Wikcionário - Artigos |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Micro-indicador TRAFEGO** |  |  |  |  |
| 1and1.com | Dmoz.org | Jurn.org | Reverbnation.com | Yelp.com |
| 4shared.com | Doaj.org | Justanswer.com | Rumble.com | Youku.com |
| 500px.com | Douban.com | Kaixin001.com | Rutube.ru | Youtube.com |
| A2hosting.com | Draugiem.lv | Kakao.com | Toad.pt | Zhihu.com |
| Abilogic.com | Dreamhost.com | Kongregate.com | Scielo.org | Zoosk.com |
| Sobre mim | Dreamwidth.org | Last FM | Scienceopen.com |  |
| Academia.edu | Dropbox.com | Librarything.com | Search.com |  |
| Acfun.tv | Drupal.org | Linkedin.com | Secondlife.com |  |
| Adam4Adam.com | Duckduckgo.com | Liquidweb.com | Semanticscholar.org |  |
| Addictinggames.com | DXY.cn | Liveleak.com | Sharecare.com |  |
| Adobe.com | Eclipse.org | Logoslibrary.eu | Similarweb.com |  |
| Adultfriendfinder.com | Edx.org | Mamba.ru | Sitebuilder.com |  |
| Alexa.com | Egnyte.com | Match.com | Skyrock.com |  |
| Alivedirectory.com | Eharmony.com | Mediafire.com | Slideshare.net |  |
| Anastasiadate.com | Etoro.com | Meetic.fr | Smugmug.com |  |
| Angel.co | Europa.eu/ | Meetup.com | Snapchat.com |  |
| Anobii.com | Exalead.com | Mega.nz | Socolar.com |  |
| Answers.com | Excite.com | Mendeley.com | Sogou.com |  |
| Apple.com | Experienceproject.com | Metacafe.com | Somuch.com |  |
| Archive.org | Fetlife.com | Metafilter.com | Sony.com |  |
| Archives-ouvertes.fr | Filefactory.com | Microsoft.com | Soso.com |  |
| Armorgames.com | Fileserve.com | Miniclip.com | Soundcloud.com |  |
| Arvixe.com | Filmaffinity.com | Mixi.jp | Spaces.ru |  |
| Arxiv.org | Filmow.com | Mocospace.com | Spip.net |  |
| Ashleymadison.com | Flickr.com | Moodle.org | Spotify.com |  |
| Ask.com | Flipboard.com | Mouthshut.com | Squarespace.com |  |
| Ask.fm | Flixster.com | Mozilla.org | Stackexchange.com |  |
| Atom.io | Fotki.com | Mubi.com | Startpage.com |  |
| Avvo.com | Fotolog.com | Myheritage.com | Steampowered.com |  |
| Badoo.com | Foursquare.com | Mylife.com | Straightdope.com |  |
| Baidu.com | Fun-mooc.fr | Myspace.com | Stumbleupon.com |  |
| Bandcamp.com | Funnyordie.com | Napster.com | Sublimetext.com |  |
| Base-search.net | Futurelearn.com | Naver.com | Tagged.com |  |
| Beyond.com | G2a.com | Netbeans.org | Taringa.net |  |
| Bilibili.com | Gaiaonline.com | Netcraft.com | Theses.fr |  |
| Bing.com | Gameblog.com | Netflix.com | Tinyurl.com |  |
| Bit.ly | Gamefaqs.com | Netlog.com | Trombi.com |  |
| Bitbucket.org | Geni.com | Newgrounds.com | Tudou.com |  |
| Bitshare.com | Gfycat.com | Nicovideo.jp | Twitch.tv |  |
| Blackle.com | Gigablast.com | Ning.com | Twoo.com |  |
| Bluehost.com | Gigasize.com | Notepad-plus-plus.org | Udacity.com |  |
| Blurtit.com | Girlsaskguys.com | Novoed.com | Udemy.com |  |
| Box.com | Github.com | Oatd.org | Uploaded.net |  |
| Brackets.io | Godaddy.com | Odnoklassniki.ru | Uploading.com |  |
| Business.com | GOG.com | Office.com | Veoh.com |  |
| Busuu.com | Goodreads.com | Okcupid.com | Viadeo.com |  |
| C9.io | Google com | Openclassrooms.com | Vimeo.com |  |
| Cafemom.com | Gotinder.com | Opengrey.eu | Vine.co |  |
| Cairn.info | Grindr.com | Openlibrary.org | Visualstudio.com |  |
| Care2.com | Hi5.com | Openoffice.org | Vk.com |  |
| Caringbridge.org | Hightail.com | Openthesis.org | Wattpad.com |  |
| Chacha.com | Hostgator.com | Opera.com | Wayn.com |  |
| Chrome.com | Hulu.com | Origin.com | Wdl.org |  |
| Classmates.com | Icloud.com | Paypal.com | Webcrawler.com |  |
| Codeanywhere.com | Infinit.io | Periscope.com | Webometrics.info |  |
| Codepen.io | Influenster.com | Periscope.tv | Weebly.com |  |
| Commonsensemedia.org | Inmotionhosting.com | Photobucket.com | Weheartit.com |  |
| Contentful.com | Instagram.com | Pinterest.com | Wetransfer.com |  |
| Couchsurfing.com | Iqiyi.com | Playstation.com | Wikimedia.org |  |
| Coursera.org | Italki.com | Plurk.com | Wistia.com |  |
| Crunchyroll.com | Itch.io | Qq.com | Wix.com |  |
| Cyworld.com | Ixquick.com | Quora.com | Wolframalpha.com |  |
| Dailymotion.com | Ixquick.eu | Qwant.com | Wordpress.com |  |
| Dart-europe.eu | Jasminedirectory.com | Raptr.com | Worldcat.org |  |
| Daum.net | Jetbrains.com | Ravelry.com | Worldwidescience.org |  |
| Deezer.com | Joinhouse.party | Reddit.com | Xbox.com |  |
| Delicioso | Joomla.com | Rediff.com | Xing.com |  |
| Depositfiles.com | Journalseek.net | Renren.com | Yacy.net |  |
| Deviantart.com | Jstor.org | Researchgate.net | Yahoo.com |  |

# Anexo II. Fontes selecionadas

|  |  |  |
| --- | --- | --- |
| **NOME DA FONTE** | **ENDEREÇO ​** | **NÚMERO DE INDIC.** |
| Alexa. dados sobre o tráfego para sites[[56]](#footnote-56) | http://www.alexa.com/ | 310 |
| Broadband Commission Report 2016. Dados sobre conectividade e acesso à informação | [http://broadbandcommission.org/](http://www.broadbandcommission.org/Documents/reports/bb-annualreport2016.pdf) | 5 |
| INC. Dados demográficos e acesso à informação | [https://www.cia.gov/](https://www.cia.gov/library/publications/the-world-factbook/fields/2150.html#xx) | 1 |
| Owloo. Porcentagem de uso do Facebook | https://www.Owloo.com/ | 1 |
| Imperativo de progresso social. | http://socialprogressimperative.org/ | 1 |
| Statista. Dados na consulta do site[[57]](#footnote-57) | https://www.statista.com/statistics/ | 31 |
| Apache OpenOffice. Número de downloads do OpenOffice | http://openoffice.org/stats/countries | 1 |
| Estatísticas do mundo da Internet. Dados sobre o número de usuários por país e língua | [http://www.Internetworldstats.com/](http://www.internetworldstats.com/) | 2 |
| Traduzido. Índice de interesse econômico por língua | <http://www.translated.net/> | 1 |
| W3Techs. conteúdo por língua | https://W3Techs.com/ | 1 |
| WebIndex Data 2014. Índices sobre a Sociedade da Informação | <http://thewebindex.org/> | 9 |
| Wikipedia. Demografia e demo-linguística | <https://wikipedia.org/> | 2 |
| Amazon. Livros | https://www.amazon.com/ | 2 |
| Dados da Wikimedia sobre os vários wikis no universo da Wikimedia | <https://stats.wikimedia.org/> | 11 |
| Projeto Joshua. dados demográficos e linguísticos | http://legacy.joshuaproject.net/ | 1 |
| Ethnologue. dados demo-linguísticos. | https://www.ethnologue.com/ | 1 |

# Anexo III: Valores escolhidos para L2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Código ISO** | **Língua** | **L2** | **L1** | **L1 + L2 / L1** |
| awa | Awadhi | **45 400** | 11 821 647 | 1,0038 |
| ben | Bengali | **19 202 880** | 248 094 781 | 1.0774 |
| bho | Bhojpuri | **160.000** | 48 656 324 | 1,0033 |
| ces | Tcheco | **2540000** | 10 121.000 | 1.251 |
| deu | Alemão | **52 689 000** | 69 419 000 | 1.759 |
| ele | Grego | **57.000** | 11 455 800 | 1.005 |
| eng | Inglês | **611 563 010** | 344 179 157 | 2,7769 |
| fas | Persa | **18 434 114** | 43 495 306 | 1,4238 |
| fra | Francês | **153 485 770** | 76 660 700 | 3,0021 |
| hau | Hausa | **19.500.000** | 45 494 090 | 1,4286 |
| hin | Hindi | **121 230 000** | 374 692 040 | 1,3235 |
| ita | Italiano | **3 085 000** | 41769700 | 1.0739 |
| jpn | Japonês | **11.500** | 125 376 800 | 1,0001 |
| kan | Canarim | **9.000.000** | 43 729 877 | 1.2058 |
| khm | Khmer | **1 000 000** | 16 030 000 | 1.0624 |
| mar | Marati | **3.000.000** | 82 230 980 | 1.0365 |
| msa | Malaio | **175 200.000** | 53 798 640 | 4,2566 |
| mya | Birmanês | **10.000.000** | 31 704 180 | 1,3154 |
| orm | Oromo | **170.000** | 33 129 700 | 1,0051 |
| pol | Polonês | **454.000** | 40 479 300 | 1.0112 |
| por | Português | **11 180.000** | 217 192 370 | 1.0515 |
| rus | Russo | **113 273 820** | 130 013 300 | 1,8712 |
| sem | Cingalês | **2.000.000** | 11 838 680 | 1,1689 |
| som | Somali | **95 600** | 19 801 000 | 1,0048 |
| spa | Espanhol | **91 308 400** | 434 569 400 | 1,2101 |
| tam | Tâmil | **8.000.000** | 74 551 992 | 1,1073 |
| tel | Telugu | **5.000.000** | 87 584 764 | 1.0571 |
| isso | Tailandês | **40.000.000** | 28 549 500 | 2,4011 |
| tur | Turco | **380 300** | 64 175 300 | 1,0059 |
| urd | Urdu | **94 045 800** | 99 137 906 | 1.9486 |
| yor | Ioruba | **2.000.000** | 38 104 300 | 1.0525 |
| zha | Zhuang | **20.000** | 16.305.000 | 1,0012 |
| zul | Zulu | **15.700.000** | 12 159 200 | 2,2912 |
| **zzz** | **RESTO** | **236 200.000** | **891 034 646** | **1,2651** |
|  | **TOTAL** | **1824031594** | **7296012569** | **1,2500** |

notas:

- A figura malaia inclui indonésio

- Os valores L1 para chinês sendo maiores do que os valores L1 + L2 indicados pelo Ethnologue, L2 foi deixado em 0 para o chinês.

- Por precaução, foi registrado um valor de 1.265 para os demais línguas: o efeito marginal nos cálculos é evitar que os resultados dos 140 línguas selecionados sejam muito altos. O valor foi finalmente escolhido para permitir um arredondamento para aproximadamente 25% do efeito do multi-linguismo .

# Anexo IV: Sites totalmente locais

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **LOCAL** | **linguagem dominante** | **% língua**  **dominante** |  | **LOCAL** | **linguagem dominante** | **% língua**  **dominante** |
| Gmx.net | alemão | 57 |  | Filmaffinity.com | espanhol | 72 |
| Xing.com | alemão | 53 |  | Fotolog.com | espanhol | 52 |
| Adam4adam.com | | Inglês | 50 |  | Gigasize.com | espanhol  hindi | 21  17 |
| Adictinggames.com | | Inglês | 53 |  | Megaupload.com | espanhol | 21 |
| Aim.com | | Inglês | 62 |  | Rapidshare.com | espanhol | 17 |
| Amazon.com | | Inglês | 50 |  | Taringa.net | espanhol | 78 |
| Avvo.com | | Inglês | 67 |  | Tuenti.com | espanhol | 65 |
| Beyond.com | | Inglês | 68 |  | Archives-ouvertes.fr | francês | 21 |
| Cafemom.com | | Inglês | 51 |  | Bnf.fr | francês | 45 |
| Caringbridge.org | | Inglês | 68 |  | Cairn.info | francês | 41 |
| Chacha.com | | Inglês | 48 |  | Deezer.com | francês | 36 |
| Classmates.com | | Inglês | 67 |  | Exalead.com | francês | 36 |
| Commonsensemedia.org | | Inglês | 59 |  | Fnac.com | francês | 70 |
| Crunchyroll.com | | Inglês | 50 |  | Fun-MOOC .com | francês | 48 |
| Eharmony.com | | Inglês | 62 |  | Gameblog.com | francês | 69 |
| Excite.com | | Inglês | 53 |  | Openclassrooms.com | francês | 52 |
| Fetlife.com | | Inglês | 52 |  | Qwant.com | francês | 68 |
| Flixter.com | | Inglês | 50 |  | Skyrock.com | francês | 47 |
| Gaiaonline.com | | Inglês | 56 |  | Theses.fr | francês | 51 |
| Gfycat.com | | Inglês | 55 |  | Viadeo.com | francês | 51 |
| Influenster.com | | Inglês | 64 |  | Yacy.com | francês | 34 |
| Joinhouse.party | | Inglês | 57 |  | Mouthshut.com | hindi | 43 |
| Justanswers.com | | Inglês | 53 |  | Novoed.com | hindi | 14 |
| Match.com | | Inglês | 53 |  | Rediff.com | hindi | 40 |
| Metafilter.com | | Inglês | 54 |  | Rediffmail.com | hindi | 44 |
| Mocospace.com | | Inglês | 55 |  | Somuch.com | hindi | 36 |
| Mylife.com | | Inglês | 66 |  | Anobii.com | italiano | 38 |
| Oovoo.com | | Inglês | 51 |  | Nicovideo.jp | japonês | 90 |
| Ravelry.com | | Inglês | 57 |  | 4shared.com | Português  árabe  espanhol | 22  16  11 |
| Rumble.com | | Inglês | 64 |  | Filmow.com | Português | 83 |
| Sharecare.com | | Inglês | 51 |  | Toad.pt | Português | 71 |
| Smugmug.com | | Inglês | 54 |  | Scielo.org | Português  espanhol | 52  21 |
| Straightdope.com | | Inglês | 55 |  | Icq.com | russo | 31 |
| Telegram.com | | Inglês | 54 |  | Mail.ru | russo | 41 |
| Yelp.com | | Inglês | 65 |  | Mamba.ru | russo | 43 |
| Jurn.org | | árabe | 13 |  | Spaces.ru | russo | 47 |
| Baidu.com | | chinês | 88 |  | Playstore.com | turco | 41 |
| Douban.com | | chinês | 85 |  |  |  |  |
| Ikiyi.com | | chinês | 88 |  |  |  |  |
| Kaixin001.com | | chinês | 85 |  |  |  |  |
| Odnosklassniki.ru | | chinês | 75 |  |  |  |  |
| Plurk.com | | chinês | 64 |  |  |  |  |
| Qq.com | | chinês | 79 |  |  |  |  |
| Renren.com | | chinês | 84 |  |  |  |  |
| Sogou.com | | chinês | 92 |  |  |  |  |
| Weibo.com | | chinês | 89 |  |  |  |  |
| Youku.com | | chinês | 89 |  |  |  |  |
| Zihu.com | | chinês | 86 |  |  |  |  |
| Cyword.com | | coreano | 78 |  |  |  |  |
| Kakao.com | | coreano | 84 |  |  |  |  |
| Naver.com | | coreano | 86 |  |  |  |  |

1. Em particular, o ambicioso Language Observatory Project (LOP) liderado por Yoshiki Mikami da Universidade de Nagaoka. Veja [9]. [↑](#footnote-ref-1)
2. Parece que o tsunami sofrido pelo Japão em março de 2011 desencadeou a paralisia do projeto LOP. [↑](#footnote-ref-2)
3. Vejo [https://web.archive.org/web/20180809042117/http://dilinet.org/](https://web.archive.org/web/20180809042117/http:/dilinet.org/) Y <http://funredes.org/lc/dilinet/es/> [↑](#footnote-ref-3)
4. Em Janeiro de 2012, convite à apresentação de propostas ICT-2011.4.4 Gestão inteligente da informação; em janeiro de 2013, denominado ICT-2013.4.1 Análise de conteúdo e tecnologias de linguagem - Análise transversal de conteúdo. [↑](#footnote-ref-4)
5. Esta análise foi realizada para o francês, no entanto, este processo deve ser tratado com cautela devido à redução dos dados estatísticos que envolve a segmentação da amostra de micro-indicadores. [↑](#footnote-ref-5)
6. Por exemplo, a distribuição percentual global de pessoas conectadas por país é transformada na distribuição percentual global por língua. [↑](#footnote-ref-6)
7. O termo "cota" é preferido em vez de "porcentagem" porque uma compreensão consistente do multilinguismo implicará que os cálculos não estão mais relacionados à população mundial, mas ao número total de falantes por língua, um valor que é igual à população mundial aumentado pelo número pessoas bilingues. [↑](#footnote-ref-7)
8. Amazonas (<http://www.amazon.com>) é uma empresa líder mundial na venda de livros online e transformada em uma plataforma de e-commerce de diversos tipos de produtos. [↑](#footnote-ref-8)
9. A média truncada é a média dos valores do vetor após a remoção dos valores extremos mais altos e mais baixos. É usado como um recurso para reduzir consideravelmente os vieses resultantes da seleção da fonte. Por exemplo, este método permite reduzir, para o indicador de tráfego, o enviesamento devido à escolha de aplicações que privilegiam determinados línguas (ver 7.4.7 Limitações / enviesamentos relacionados com o grau de localidade das fontes). [↑](#footnote-ref-9)
10. Um portal de tecnologia para pesquisas e estatísticas. <https://w3techs.com/> [↑](#footnote-ref-10)
11. http://thewebindex.org/ [↑](#footnote-ref-11)
12. As únicas fontes disponíveis com dados para todos os países do mundo são pessoas conectadas à Internet, contas de banda larga fixa ou móvel e o número de downloads do Open Office. [↑](#footnote-ref-12)
13. Em inglês: código aberto, dados abertos, conhecimento aberto. [↑](#footnote-ref-13)
14. Alexa (<http://alexa.com>) é uma empresa que fornece estatísticas e classificações em sites com base no tráfego estimado. [↑](#footnote-ref-14)
15. Statista (<http://statista.com>) é um portal de estatísticas gerais que foi organizado por eles. [↑](#footnote-ref-15)
16. Mais de 80% dos micro-indicadores são medidos com a ajuda deste serviço. [↑](#footnote-ref-16)
17. Um portal para dados demo-linguísticos. <https://joshuaproject.net/> [↑](#footnote-ref-17)
18. Quando uma quantidade (ou uma porcentagem mundial) mede os falantes de uma língua (L1 ou L2) que têm acesso à Internet, deve-se entender que se o total for feito para todas as línguas do mundo será logicamente maior do que a população total de usuários da Internet (ou se for expresso como uma porcentagem mundial, será superior a 100%). Claro que a mesma pessoa, se for bilingue, aparecerá duas vezes nesta contabilidade (3 vezes se falar 3 línguas, 4 vezes se falar 4 línguas, etc.). [↑](#footnote-ref-18)
19. Há também a simplificação de contar bebês como falantes de sua língua materna desde o nascimento, embora isso realmente aconteça depois de alguns anos de aprendizagem. [↑](#footnote-ref-19)
20. Um portal para dados demo-linguísticos sobre todas as línguas do mundo (<http://ethnologue.com>), um projeto da SIL Internacional (<http://sil.org>) [↑](#footnote-ref-20)
21. Isso significa que 25% da população mundial possui uma segunda língua. [↑](#footnote-ref-21)
22. APL, "A Programming Language" em inglês, é um formalismo matemático e sua implementação na forma de uma linguagem de programação, projetada por Kenneth Iverson. Para mais detalhes, veja https://pt.wikipedia.org/wiki/APL\_(linguagem\_de\_programação). [↑](#footnote-ref-22)
23. Mas não esqueçamos que embora a Wikipedia seja uma das aplicações mais globalizadas da Internet, ela não está muito presente em alguns países importantes da Internet, como a China ou a Índia, e oferece informações em no máximo 300 línguas. [↑](#footnote-ref-23)
24. A análise realizada mostra que a W3Techs não pode ser considerada uma medida confiável, longe disso. [↑](#footnote-ref-24)
25. Obviamente, também poderia medir a propensão de falantes de outras línguas para produzir conteúdo em inglês. [↑](#footnote-ref-25)
26. http://thewebindex.org/ [↑](#footnote-ref-26)
27. The Economist acaba de publicar um relatório com macro-indicadores para medir o "grau de inclusão" da Internet (<https://theinclusiveInternet.eiu.com>) A ONG Open Knowledge publica um índice para medir dados abertos (<http://index.okfn.org/>) A UNESCO pretende desenvolver o grau de universalidade da Internet no país ([http://unesco.org/new/en/media-services/single-view/news/unesco\_call\_for\_proposals\_defining\_the\_Internet\_universality\_in/](http://www.unesco.org/new/en/media-services/single-view/news/unesco_call_for_proposals_defining_the_Internet_universality_in/)) [↑](#footnote-ref-27)
28. Por exemplo, a Wikpedia em chinês teria cerca de 700.000 artigos, em contraste com as duas opções de enciclopédia chinesa (Baidu Baike e Hudong) que juntas representam quase 12 milhões de artigos, ou seja, mais de 17 vezes (fonte: <https://fr.wikipedia.org/wiki/Wikipédia_en_chinois>) [↑](#footnote-ref-28)
29. Cursos online abertos e massivos. [↑](#footnote-ref-29)
30. Este é um serviço online gratuito da Miniwatts Marketing Group, uma empresa que oferece em seu site (http://InternetWorldStats.com) estatísticas de população e uso da Internet interessantes e bem organizadas e alguns números para línguas (http://www.Internetworldstats.com/stats7.htm) e ganhou notoriedade bem merecida. [↑](#footnote-ref-30)
31. De fato, dado que as percentagens de pessoas conectadas nos principais países francófonos (França, Bélgica, Suíça, Canadá) são, respectivamente, 83%, 85%, 87% e 88% e que esses países representam 115 milhões de falantes de francês, para obter uma média ponderada de 26% como indica o IWS, os outros 279 milhões de falantes de francês (o IWS indica 394 milhões como a população total de falantes de francês) devem ter menos de 2% de pessoas conectadas. Isso é indicado pela equação: 26% = (115M \* 85% + 279M \* x%) / 394M ==> x = 1,7%. Embora a África francófona sofra de exclusão digital, suas taxas de conectividade por país variam entre uma janela de 4-7% (Burundi, África Central, Chade), passando por uma janela de 10-14% (Benin, Mali, Madagascar, Níger) e muitos com mais de 45% (Costa do Marfim, Gabão, Marrocos, [↑](#footnote-ref-31)
32. Exceto para a decisão de usar a média truncada, outras alterações provavelmente afetariam o decimal em uma amplitude máxima de + - 0,3% se considerarmos apenas os métodos de micro-indicador ou agrupamento de cálculos e mais se os desvios em L2 forem importantes. [↑](#footnote-ref-32)
33. Isso é especialmente verdadeiro para a ordem alemã e russa em termos de poder e habilidade. [↑](#footnote-ref-33)
34. Online Computer Library Center, uma rede acadêmica líder que atende bibliotecários. [↑](#footnote-ref-34)
35. Veja [6] e [8]. [↑](#footnote-ref-35)
36. Portanto, essa limitação evita tirar conclusões para línguas não nacionais que são faladas dentro de um país. No caso da França, por exemplo, onde o Ministério da Cultura está muito interessado neste assunto, este estudo não fornece nenhuma luz particular e é preferível referir-se ao estudo apoiado pela Direção da Língua Francesa deste Ministério (ver [1] ). [↑](#footnote-ref-36)
37. Os línguas com maior número de países nos quais são distribuídos, para aqueles com mais de 50 países, na ordem: inglês (151 países), árabe (117), francês (105), chinês (94) alemão (81) Grego (74), Rússia (70), Espanhol (62), Italiano (58), Hindi (56), Armênio (55) e Português (51). [↑](#footnote-ref-37)
38. Deve ficar claro que às vezes a extrapolação será aplicada em alguns países onde o aplicativo não é usado e, em outros casos, subestimará os valores reais em alguns países onde o aplicativo é amplamente usado. [↑](#footnote-ref-38)
39. Em outras palavras, sempre dê preferência às fontes primárias. [↑](#footnote-ref-39)
40. <https://publicadministration.un.org/egovkb/en-us/reports/un-e-government-survey-2016> [↑](#footnote-ref-40)
41. [http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals\_la Internet\_2000-2015.xls](http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals_Internet_2000-2015.xls) [↑](#footnote-ref-41)
42. Manual sobre a medição do acesso de famílias e indivíduos às tecnologias de informação e comunicação (TIC) e uso dessas tecnologias, 2014 (ver página 74) - <https://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ITCMEAS-2014-PDF-F.pdf> [↑](#footnote-ref-42)
43. <http://ec.europa.eu/eurostat/statistics-explained/index.php/Information_society_statistics_-_households_and_individuals/fr> [↑](#footnote-ref-43)
44. Por exemplo, Martinica, Guadalupe, Guiana e Reunião estão incluídas na França, mas a Polinésia Francesa, Nova Caledônia, Wallis e Futuna e São Pedro e Miquelon não estão. [↑](#footnote-ref-44)
45. Alexa não relata claramente os "milhões de barras de ferramentas instaladas" (http://www.alexa.com/about) e um artigo da Wikipedia menciona a cifra de 10 milhões em 2005. [↑](#footnote-ref-45)
46. De acordo com especialistas do OpenRoot, isso diz respeito a 80% dos sites chineses, o que leva a uma subestimação do conteúdo (ou tráfego) em chinês por um fator de 5! [↑](#footnote-ref-46)
47. Esta afirmação vem de um raciocínio simples e não pode ser apoiada pelos dados na ausência de dados estatísticos na barra de ferramentas; no entanto, a análise dos resultados do W3Techs confirmará isso mais tarde. [↑](#footnote-ref-47)
48. Com altas pontuações em inglês no ranking Alexa, um mero fator de superestimação de 1,7 como no caso do LinkedIn, eles representam mais de 20% da superestimação generalizada em relação a outros línguas, o que é enorme. [↑](#footnote-ref-48)
49. E também pode ser para o francês, a julgar pelos testes de comparação realizados. [↑](#footnote-ref-49)
50. <https://W3Techs.com/technologies/overview/content_language/all> [↑](#footnote-ref-50)
51. Marginal ... mas quase sempre a favor do inglês. [↑](#footnote-ref-51)
52. <http://funredes.org/lc> [↑](#footnote-ref-52)
53. É claro que o método então utilizado produzia resultados em percentuais em relação ao inglês, e determinar a porcentagem absoluta do conteúdo em inglês (o que permitia, na sequência, deduzir os demais percentuais absolutos) era uma etapa delicada que só poderia ser resolvida. em aproximação cruzando vários dados. [↑](#footnote-ref-53)
54. Veja, como um exemplo típico, o recente relatório Mozilla sobre a saúde da Internet que dá atenção especial aos aspectos linguísticos, infelizmente, a partir de suposições tendenciosas (<https://Internethealthreport.org/v01/>). [↑](#footnote-ref-54)
55. Embora uma análise mais aprofundada tenha mostrado que esta avaliação deve ser qualificada porque a representação das principais línguas asiáticas na Wikipedia está muito abaixo de sua presença demográfica. [↑](#footnote-ref-55)
56. Os dados são pagos. [↑](#footnote-ref-56)
57. O Statista foi escolhido como a principal fonte paga após o estudo de possíveis fontes. [↑](#footnote-ref-57)