RELATÓRIO DO ESTADO DOS IDIOMAS DA INTERNET

Resumo do Relatório · 2022

Por que fizemos este relatório? Quem somos nós? 2

Como ler este relatório? 5

Quão multilíngue é a internet? 6

O que aprendemos sobre a internet multilíngue? 24

Como podemos melhorar?: Contextos e ações para uma internet multilíngue 25

Por fim, o que você pode fazer? 37

Gratidão 40

Definições 41

Resumo do relatório

Por que fizemos este relatório? Quem somos nós?

Os dicionários e gramáticas nos dizem que a linguagem é uma forma estruturada de expressar informação , principalmente entre seres humanos. Mas a linguagem é muito mais do que isso: é a herança fundamental que oferecemos uns aos outros, na maioria das vezes passada para nós por nossos ancestrais e, se tivermos sorte, uma herança a ser transmitida àqueles que virão depois de nós. Quando pensamos, falamos, ouvimos, imaginamos... usamos a linguagem para nós mesmos e para os outros. É a essência de quem somos e como somos no mundo. É o que nos ajuda a contar histórias e compartilhar o que sabemos sobre nós e os outros. Que língua você fala? Você sonha nessa língua? Você pensa em um idioma diferente daquele que fala no trabalho? A música que você ama está numa língua que nem sempre você entende?

Cada idioma é um sistema de ser, fazer e se comunicar no mundo. E o mais importante, de saber e imaginar. Cada língua é um sistema de conhecimento em si: nossa língua é uma forma fundamental de entendermos nosso mundo e o explicarmos aos outros. **Nossas línguas podem ser orais (faladas ou sinalizadas), escritas, ou transmitidas também por sons**, como um assovio ou um tambor! Em todas essas formas, **a língua é um proxy para o conhecimento**. Em outras palavras, a língua é a maneira mais óbvia pela qual expressamos o que pensamos, acreditamos, e sabemos.

Agora, pense nas línguas que você fala, nas quais pensa, sonha ou escreve. De todas elas, com quantas você consegue se comunicar plenamente em espaços digitais? Qual é a sua experiência no uso de idiomas ou idiomas online? O hardware que você usa tem os caracteres do seu idioma? Você precisa modificar os teclados para usá-los com o seu idioma? Quando você procura informações usando um mecanismo de pesquisa, os resultados são mostrados no idioma que você deseja? Você teve que aprender uma língua diferente da sua para acessar a internet e contribuir com ela? Se a sua resposta a alguma ou muitas dessas perguntas for “não”, então você está entre as poucas pessoas privilegiadas no mundo que conseguem usar a internet no seu próprio idioma facilmente. E seu idioma provavelmente é... o inglês.

A internet e seus diferentes espaços digitais oferecem uma das infraestruturas mais críticas de conhecimento, comunicação e ação da atualidade. Mesmo assim, com mais de 7.000 idiomas no mundo (incluindo línguas faladas e sinalizadas), **quantos desses idiomas podemos experimentar totalmente online? Como seriam a aparência, a sensação e o som de uma internet verdadeiramente multilíngue?**

Este relatório é uma forma de tentarmos responder a essa pergunta. Somos uma colaboração de três organizações: Whose Knowledge?, Oxford Internet Institute e The Center for Internet and Society (da Índia). Nós nos reunimos para fornecer diferentes percepções, experiências e análises sobre o estado dos idiomas na Internet e, em parceria com outros grupos que se preocupam com essas questões, esperamos criar uma internet com tecnologias e práticas digitais mais multilíngues.

Este relatório pretende fazer três coisas:

- **Mapear a atual situação dos idiomas na internet**: estamos tentando compreender quais e como os idiomas estão representados atualmente na internet. Para isso, trabalhamos tanto com dados quantitativos (analisando números em diferentes plataformas, ferramentas e espaços digitais) como com dados qualitativos (aprendendo com a nossas próprias histórias e experiências com idiomas online).
- **Aumentar a conscientização sobre os desafios e oportunidades para tornar a internet mais multilíngue**: a criação e o gerenciamento de tecnologias, conteúdo e comunidades para os idiomas do mundo apresentam desafios significativos, além de possibilidades e oportunidades fascinantes. Este relatório apresentará alguns desses desafios e possibilidades.
- **Promover um plano de ação:** Com essas percepções e consciência, propomos algumas maneiras pelas quais nós — e muitos outros que trabalham com essas questões em todo o mundo — gostaríamos de fazer planos e agir para assegurar uma internet mais multilíngue.

O que este relatório é e o que não é?

Este relatório é um trabalho em curso ou um trabalho em processo!

Muitos indivíduos, comunidades e instituições diferentes têm trabalhado em diferentes aspectos dos idiomas por muito tempo e, mais recentemente, também em diferentes aspectos das línguas online. Somos inspirados por eles, mas este relatório não pretende ser um levantamento completo e exaustivo de cada um e de seu trabalho. Também não conhecemos todo mundo que trabalha com idiomas e internet, embora tenhamos tentado incluir a maioria dos que conhecemos e nos inspiraram, de uma forma ou de outra, incluindo-os nas nossas seções de Recursos e Gratidão.

Estamos limitados pelos dados que podemos reunir e discutimos algumas dessas restrições em nossa seção Números. Agradecemos por comentários e sugestões para melhorar e atualizar as informações que oferecemos aqui e adoraríamos ouvir aqueles que já estão trabalhando nessas questões e gostariam de ser incluídos em futuras atualizações deste relatório.

Demos o melhor de nós para escrever este relatório no estilo mais acessível possível. Desejamos que várias gerações e comunidades de pessoas se juntem a nós em nosso trabalho e não queremos que jargões ou uma linguagem "acadêmica" sejam barreiras para a leitura e reflexão. Também queremos que seja traduzido para o maior número de idiomas possível (tradutores: entrem em contato conosco!), e embora tenhamos escrito parte deste documento primeiramente em inglês, não queremos que a língua seja uma barreira para reflexão ou ação.

Esperamos que este relatório sirva como uma "linha de base" para futuras pesquisas, discussões e ações sobre as questões de que ele trata, ao mesmo tempo em que nos fundamentamos em muitos esforços feitos antes de nós.

Quem somos e por que nos reunimos para fazer este relatório?

Três organizações se reuniram para fazer a pesquisa para este relatório: The Center for Internet and Society, Oxford Internet Institute e Whose Knowledge?. Todos nós estamos interessados nas implicações da internet e das tecnologias digitais a partir de diferentes perspectivas de pesquisa, política e advocacy.

Nos últimos anos, temos trabalhado o nosso próprio jeito de entender as desigualdades e injustiças dos conhecimentos na internet: quem contribui com o conteúdo online e como? Logo percebemos que havia poucos dados sobre o conhecimento em diferentes idiomas na rede. Em seguida, quisemos saber mais: qual é o alcance das línguas do mundo na internet atualmente? Até que ponto a internet é multilíngue? Nossa exploração foi limitada a apenas algumas áreas nas quais pudemos encontrar informações públicas úteis e abertas, mas esperamos que seja uma contribuição a mais para todos nós que lutamos por uma internet multilíngue.

Uma rápida nota sobre a Covid-19 e este relatório: Começamos a trabalhar neste relatório em 2019, antes da Covid-19, mas a maior parte do trabalho de análise, entrevistas e redação aconteceu durante a pandemia global que mudou a nossa vida individual e coletivamente. Todos os que contribuíram para este relatório foram afetados de alguma forma e demoramos muito mais do que esperávamos para compartilhá-lo com o mundo. Mas a Covid-19 também nos ajudou a lembrar como estamos interconectados, como é essencial para nós conseguir transmitir ideias complexas em diferentes idiomas, e como é essencial ter uma infraestrutura (digital) resiliente e acessível que seja verdadeiramente multilíngue.

Como ler este relatório?

Este relatório é o que chamamos de digital first, ou seja: a melhor forma de ler, ouvir e aprender com ele é por meio deste site, pois ele possui algumas camadas e níveis diferentes. Ele reúne Números e Histórias. Saber mais sobre o estado dos idiomas online de uma perspectiva estatística nos dá uma visão geral dos problemas e nos ajuda a entender os diferentes contextos que as pessoas estão experimentando. Mas as experiências linguísticas das pessoas na internet do mundo inteiro, em diferentes contextos, ajudam-nos a aprender mais profundamente sobre como é fácil ou difícil para elas utilizarem a internet nas suas respectivas línguas. Com histórias e números, podemos começar a abordar alguns dos contextos, desafios e oportunidades subjacentes.

É por isso que existem três camadas principais neste relatório:

- O resumo do que está no Relatório do Estado dos Idiomas e como o criamos (o que você está lendo agora!)
- Números que analisam alguns problemas críticos de linguagem em certas plataformas digitais, aplicativos e dispositivos que usamos todos os dias. Nossos amigos do Oxford Internet Institute conduziram este trabalho, e você encontrará visualizações e análises de dados fascinantes aqui. Observe que esta análise é limitada aos dados que pudemos acessar, a partir de conjuntos de dados e materiais abertos e publicamente disponíveis. Outras restrições metodológicas são discutidas em mais detalhes nesses ensaios, mas o mais importante é: é difícil encontrar uma maneira única e consistente de identificar idiomas e é igualmente difícil estimar quantas pessoas usam idiomas específicos, especialmente porque as línguas e seu uso são dinâmicos, mudando com o tempo.
- Histórias que nos dão uma compreensão mais profunda de como diferentes pessoas e comunidades pelo mundo experimentam a internet em seus próprios idiomas e, em muitos casos, como é difícil atualmente encontrar as informações de que precisam em suas próprias línguas. Fizemos uma convocatória em que solicitamos essas histórias nas formas escrita e falada, então você encontrará ensaios tanto textuais como entrevistas em áudio e vídeo. Nossos amigos do The Center for Internet and Society lideraram esse trabalho, reunindo essa rica tapeçaria de experiências linguísticas de todo o mundo. Temos contribuições sobre línguas indígenas como xindali, cree, ojibway, mapuzugun, zapotec e arrernte da África, Américas e Austrália; línguas minoritárias como bretão, basco, sardo e careliano na Europa; bem como idiomas dominantes regional e globalmente, como bengali, indonésio (Bahasa Indonesia) e cingalês na Ásia, e diferentes formas de árabe no norte da África.

Mais importante ainda, nossos colaboradores escreveram ou falaram em seus próprios idiomas e também em inglês, e nosso resumo também foi escrito e falado em diferentes idiomas, então esperamos que você goste de ler ou ouvir em mais idiomas do que apenas um!

Também fizemos o nosso melhor para dar vida a essas contribuições na forma visual, com ilustrações e animações criativas que reúnem cuidadosamente os diferentes aspectos sociais e técnicos da linguagem. Como tudo o mais em nosso relatório, eles também foram desenvolvidos de forma colaborativa por nossa ilustradora em conversas com nossos colaboradores.

Quão multilíngue é a internet?

A internet ainda não é (infelizmente, nem de longe) tão multilíngue quanto somos na vida real. Tentamos entender por que observando alguns números e experiências de pessoas em todo o mundo. Aqui, oferecemos apenas um breve resumo e análise da riqueza e profundidade do trabalho realizado por nossos colaboradores no mundo inteiro; portanto, veja os ensaios deles para ter mais detalhes e inspiração.

Em primeiro lugar, examinamos os contextos em que as pessoas estão usando a internet pelo mundo, em diferentes idiomas. Examinamos as maneiras pelas quais as informações e o conhecimento são distribuídos (ou não) em diferentes idiomas e geografias. Em seguida, examinamos mais de perto as principais plataformas e aplicativos que usamos para criar conteúdo, comunicar e compartilhar informações online e quantos idiomas cada um suporta. Examinamos em detalhes o Google Maps e a Wikipédia como dois espaços multilíngues de conteúdo que todos nós usamos no nosso dia a dia e como eles funcionam em diferentes idiomas.

Ao longo da estrada, compartilhamos histórias e experiências de pessoas que acessam e contribuem com o conhecimento na internet em seus próprios idiomas. Como acabamos concluindo, a maioria de nossos colaboradores precisa mudar de seu primeiro idioma de escolha para outro a fim de acessar as questões que valorizam e dar suas contribuições.

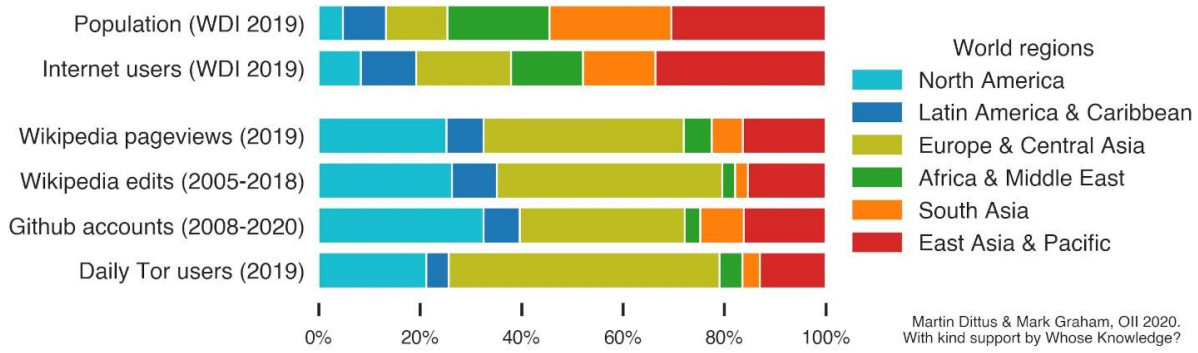
Contexto linguístico: desigualdades geográficas e de conhecimento digital

As línguas com tradição oral não cabem na web que temos hoje.
Ana Alonso

Parece-nos que estas plataformas, em geral, perpetuam a ideia colonialista de que existem línguas com maior valor e capacidade de comunicação, o que é uma visão preconceituosa de línguas minoritárias como o mapuzugun.
Kimeltuwe project

Sabemos que há uma estimativa de que mais de 60% do mundo esteja conectado digitalmente, a maioria de nós por meio de nossos telefones e dispositivos móveis. De todos os que estão online, três quartos de nós são do Sul Global: Ásia, África, América Latina, Caribe e Ilhas do Pacífico. Mas até que ponto o nosso acesso é significativo e equitativo? Podemos criar e produzir conhecimento público online na mesma medida em que o consumimos?

Na pesquisa de Martin e Mark sobre a população global em comparação com o número de usuários da internet, descobrimos que certas populações podem acessar a internet de maneira muito mais significativa do que outras, inclusive em espaços digitais muito conhecidos. Por exemplo, embora muitos dos que estamos online sejamos do Sul Global, não podemos acessar a internet como criadores e produtores de conhecimento; apenas como consumidores. A maioria das edições da Wikipédia, das contas no Github (um repositório de código) e o maior número de usuários do Tor (um navegador seguro) são da Europa e da América do Norte.

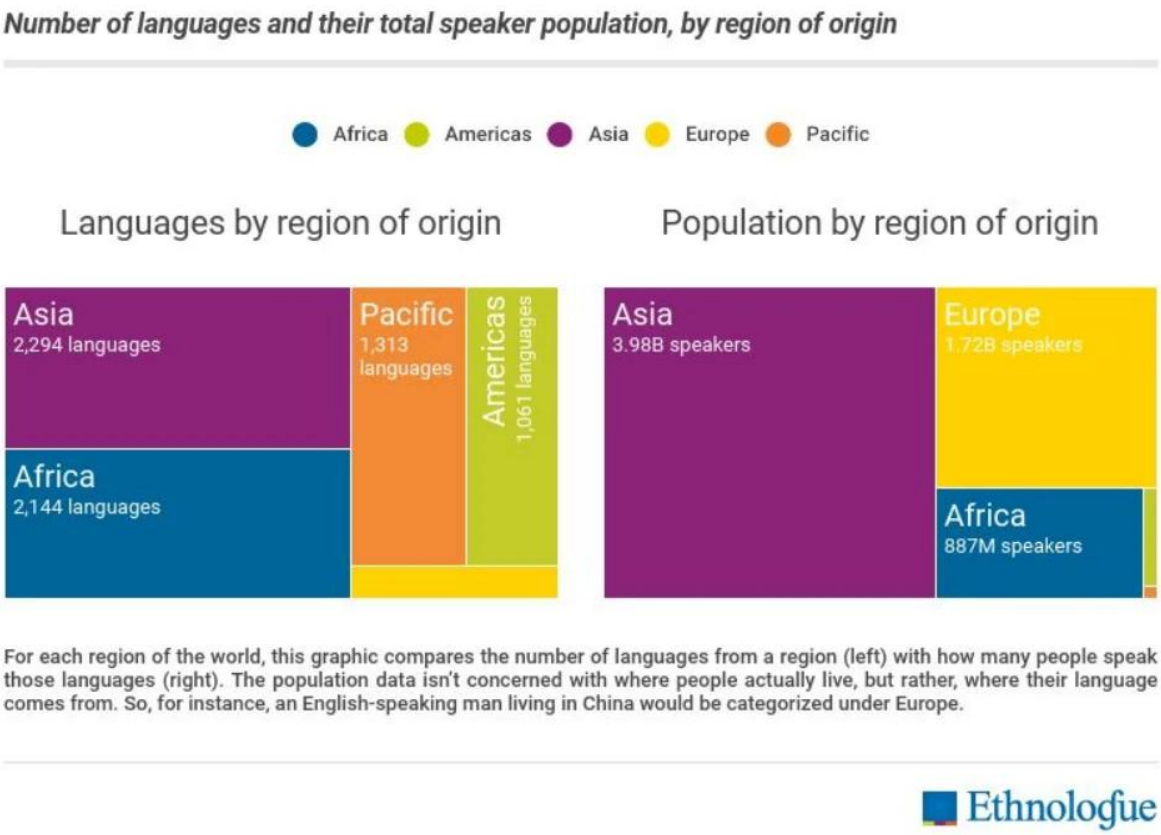


Medidas de participação digital por regiões do mundo. (Dados: Banco Mundial 2019, Wikimedia Foundation 2019, Wikipedia 2018, Github 2020, Tor 2019)

E o que esse acesso desigual significa para os idiomas? Todos nós podemos acessar a internet em nossas próprias línguas? Podemos criar conteúdo e informações em nossos idiomas?

Como outras projeções nos mostram, mais de 75% dos que acessam a internet o fazem em apenas 10 línguas — a maioria das quais com uma história colonial europeia (inglês, francês, alemão, português, espanhol...) ou idiomas dominantes em regiões específicas onde outras línguas lutam para permanecer relevantes (chinês, árabe, russo...). Em 2020, estimou-se que 25,9% de todos os usuários da internet estavam online em inglês, enquanto 19,4% a acessaram em chinês. A China é o país com mais usuários de internet em todo o mundo, e vale lembrar que o que chamamos de "chinês" não é um único idioma, mas uma família de muitas línguas diferentes.

Curiosamente, as línguas da internet do presente podem ser principalmente de origem europeia, mas a Europa tem o menor número de idiomas no mundo em comparação com todos os outros continentes. Dos mais de 7.000 idiomas do planeta, mais de 4.000 vêm da Ásia e da África (mais de 2.000 idiomas cada) e as Ilhas do Pacífico e as Américas têm mais de 1.000 idiomas cada. Papua Nova Guiné e Indonésia são os países com o maior número de línguas: mais de 800 em Papua Nova Guiné e mais de 700 na Indonésia.



Número de idiomas e população total de falantes por região no mundo. Fonte: Ethnologue

Muitos idiomas do sul da Ásia (hindi, bengali, urdu) estão entre os 10 principais idiomas do mundo pelo número de pessoas que os falam como primeira língua (ou nativa), mas estas não são as línguas em que as pessoas do Sul da Ásia podem acessar a internet. E, claro, conforme aprendemos com Ishan, cujo primeiro idioma é o bengali, mesmo que você consiga acessar o conhecimento digital no idioma escolhido, o tipo de informação que você procura pode não existir nele (no caso de Ishan, conteúdo sobre deficiência e direitos sexuais). A situação no Sudeste Asiático — que tem um dos países com maior uso de internet no mundo e uma grande diversidade de idiomas — é semelhante. Paska encontra exatamente os mesmos problemas com conteúdo sobre direitos sexuais em indonésio (Bahasa Indonesia) que Ishan encontra em bengali.

Também sabemos que dos mais de 7.000 idiomas no mundo, somente cerca de 4.000 deles possuem sistemas escritos ou scripts. Entretanto, a maioria desses scripts não foi desenvolvida pelos falantes dos idiomas, mas como parte de processos de colonização ao redor do mundo. A mera existência de scripts não significa que sejam compreendidos ou utilizados amplamente. A maioria das línguas do mundo é transmitida de forma oral ou sinalizada, e não pela escrita. Mesmo em línguas com formas escritas, sua publicação tende a ser feita em línguas coloniais europeias e, em muito menor medida, nas línguas regionais dominantes. Em 2010, o Google estimou que havia cerca de 130 milhões de livros já publicados, e uma proporção razoável deles está em cerca de 480 idiomas. Os periódicos acadêmicos mais bem estabelecidos em ciências ou ciências sociais estão em inglês. O livro mais traduzido do mundo é a Bíblia (em mais de 3.000 idiomas), e o documento mais traduzido no mundo é a Declaração Universal dos Direitos Humanos das Nações Unidas (em mais de 500 idiomas).

Por que isso importa? Porque as tecnologias de linguagem digital dependem do processamento automatizado de materiais publicados em diferentes idiomas para melhorar seu suporte e conteúdo de idioma. Portanto, quando a publicação de textos em todo o mundo favorece certos idiomas — e não consegue incluir nenhum idioma não escrito — isso aprofunda as desigualdades de linguagem que vivenciamos. E, é claro, as línguas não baseadas em texto — as baseadas em sinais, sons, gestos, movimentos — estão completamente ausentes da indústria editorial e, consequentemente, muitas vezes ausentes das tecnologias de linguagem digital.

Por exemplo, como Ana nos disse, “a web não foi projetada para responder aos usuários de línguas que contam somente com a tradição oral”. No âmbito das linguagens escritas na internet, é difícil encontrar conteúdos oriundos das tradições das línguas orais e visuais. Não podemos facilmente fazer uma busca de gestos, sinais e assobios, por exemplo. Em nossa entrevista com Joel e Caddie, eles nos falaram sobre o primeiro conjunto de emojis indígenas da Austrália feito em terras arrernte em Mparntwe/Alice Springs, e como o gesto físico é frequentemente combinado com a palavra falada para se expressar em arrernte. Emna disse o mesmo sobre a Tunísia e as diferentes línguas faladas por seu povo: “Quando se trata de preservar uma língua, não devemos nos concentrar apenas na escrita; precisamos fazê-lo em formas orais, gestos, sinais, assobios, etc., que não podem ser captados totalmente por escrito.”

As tecnologias digitais nos oferecem essas possibilidades de representar a pluralidade de formas de linguagem que se baseiam em texto, som, gesto e muito mais. Elas também podem nos ajudar a preservar e trazer de volta à vida as línguas que estão sob risco de extinção: mais de 40% de todas as línguas. Todos os meses, duas línguas indígenas e os saberes que elas expressam morrem e se perdem para nós.

Por que esses diferentes contextos de linguagem não são mais bem representados online?

Em seu ensaio, Claudia nos propõe três dimensões para entender a relação entre línguas e tecnologia: disponibilidade, usabilidade e como a tecnologia é desenvolvida. Como descobrimos ao longo deste relatório, o que Claudia chama de "línguas majoritárias" (e descobrimos que são principalmente línguas coloniais europeias ou línguas dominantes regionalmente) têm uma amplo leque de mídia, serviços, interfaces e aplicativos disponíveis, enquanto outras línguas têm muito menos disponibilidade de oferta, inclusive de infraestrutura como teclados, tradução automática ou reconhecimento por voz. As empresas de tecnologia também gastam muito mais tempo e recursos na usabilidade dessas línguas majoritárias, pois é aí que elas obtêm mais lucro. Finalmente, ela descobriu que a maior parte das tecnologias linguísticas é desenvolvida por meio de processos que vêm de cima para baixo, com pouca colaboração com as comunidades linguísticas, ou os poucos esforços para trabalhar com essas comunidades são mal planejados e coordenados.

Essas preocupações e desafios de contexto também nos fornecem caminhos para a criação de uma internet mais multilíngue, e voltaremos a essas possibilidades mais adiante.

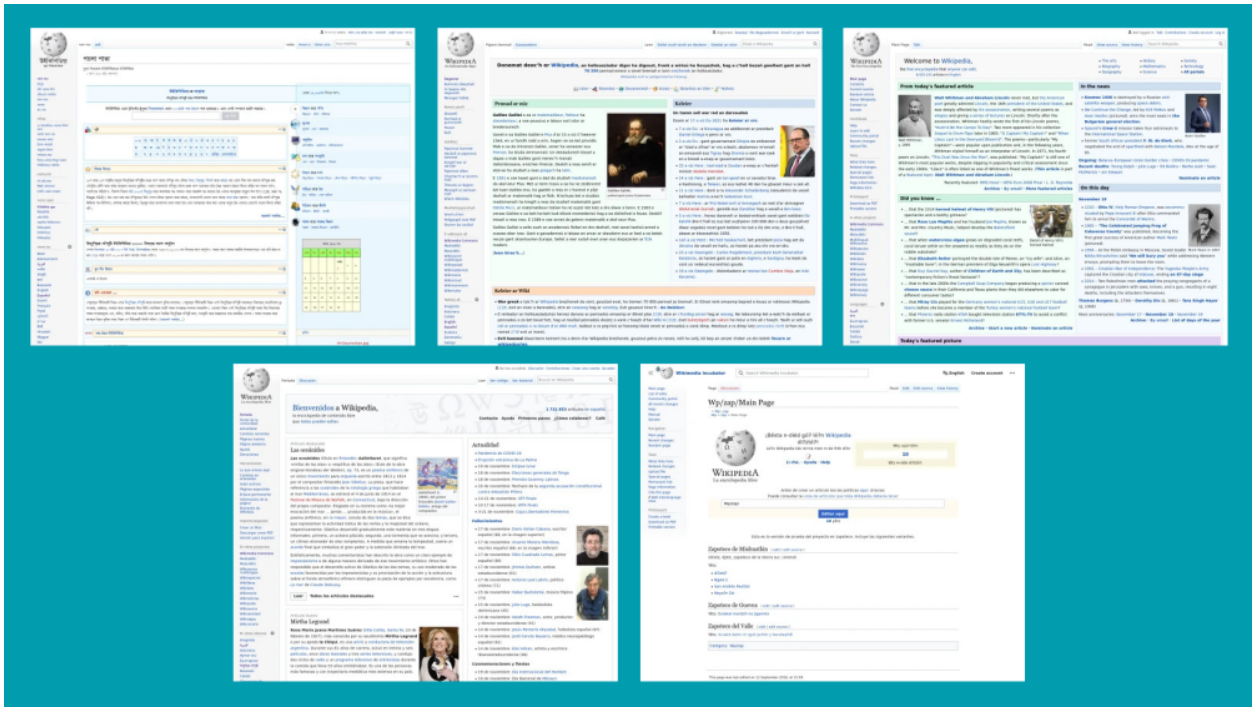
Suporte a idiomas: plataformas e aplicativos de mensagens

Quando você escrever a expressão "bom dia", antes de terminar de digitar o telefone ou computador já sugere a palavra. Mas quando escrevo a mesma expressão em chindali ("mwalamusha"), tenho que digitar a palavra inteira e isso leva muito tempo; e ainda por cima será sublinhado porque o computador ou telefone não reconhece essa palavra.
Donald Flywell Malanga

Os teclados com letras cingalês e tamil são raros. Nossos pais imprimiram etiquetas com letras minúsculas em cingalês, recortaram e colaram nas teclas em cima dos caracteres originais em inglês. Embora várias fontes cingalês tenham sido desenvolvidas, nenhuma funciona tão bem quanto as fontes Unicode.
Uda Deshapriya

Se os aplicativos populares e as principais interfaces de software não forem fornecidos em bretão em pouco tempo, o idioma certamente será menos atraente para as gerações mais jovens e não vai conseguir competir com os aplicativos franceses.
Claudia Soria

Mergulhamos mais fundo para entender como a internet ainda não é tão multilíngue quanto o mundo em que vivemos. Examinamos quais tipos de suporte linguístico — ou seja, interfaces de usuário em diferentes idiomas — as principais plataformas e aplicativos digitais nos oferecem para comunicar, criar e compartilhar conteúdo em nossos idiomas.



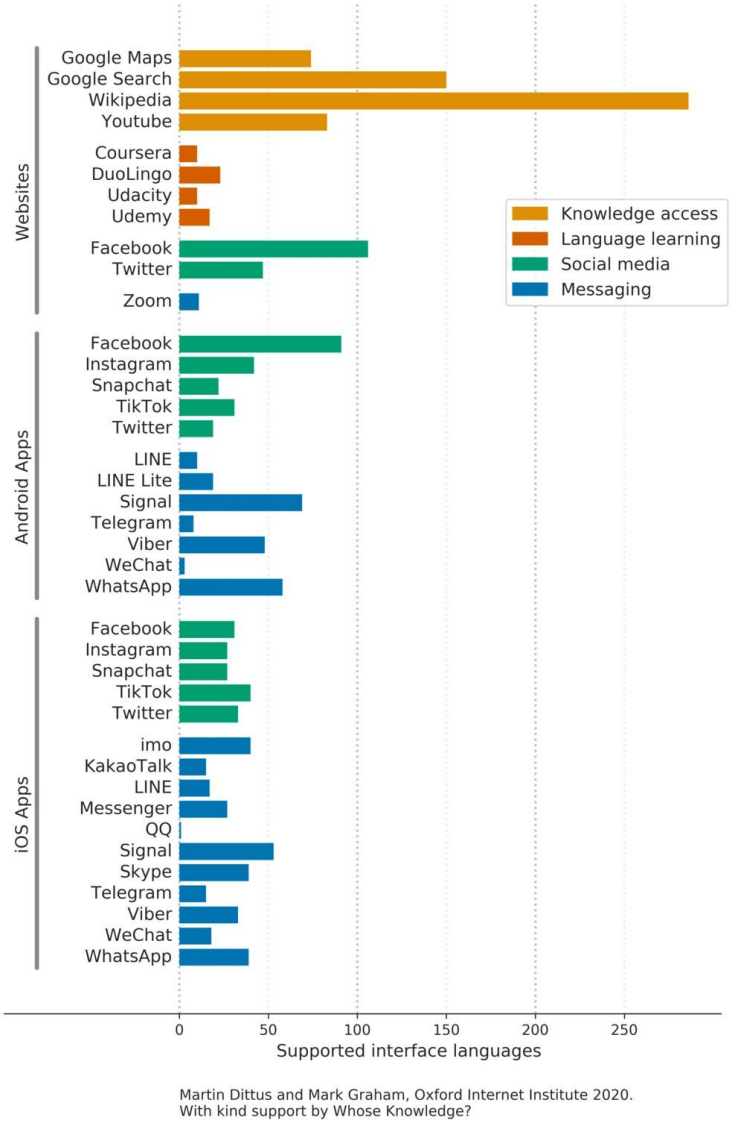
Interface da Wikipedia em vários idiomas.

Martin e Mark analisaram o suporte de idiomas para 11 sites, 12 aplicativos Android e 16 aplicativos iOS. Eles escolheram plataformas amplamente utilizadas que se especializaram na coleta e compartilhamento de conhecimento, particularmente aquelas que buscam ter uma presença global e um público em todo o mundo. Eles também tiveram que contar com dados disponíveis publicamente para essas plataformas e aplicativos, e focaram conteúdos e interfaces escritos, em vez de falados.

As plataformas foram agrupadas em quatro grandes categorias (reconhecendo que elas se sobrepõem):

- **Acesso ao conhecimento** (plataformas de conhecimento e informação, incluindo mecanismos de busca): Google Maps, Google Search, Wikipédia e YouTube.
- **Aprendizagem de línguas** (plataformas de aprendizagem autônoma de línguas): Duolingo e as plataformas educacionais Coursera, Udacity e Udemy.

- **Redes sociais** (plataformas de mídia social voltadas ao público): Facebook, Instagram, Snapchat, TikTok e Twitter.
- **Mensagens** (mensagens privadas e em grupo): imo, KakaoTalk, LINE, LINE Lite, Messenger, QQ, Signal, Skype, Telegram, Viber, WeChat, WhatsApp e Zoom.



Número de idiomas de interface suportados por cada plataforma, por categoria de plataforma.

Descobrimos que o suporte a idiomas baseados em texto é distribuído de forma altamente desigual em diferentes plataformas digitais. As principais plataformas da web, como Wikipédia, Google Search e Facebook são as que oferecem o maior suporte a idiomas atualmente. Curiosamente, a Wikipédia (que é uma organização sem fins lucrativos e editada por voluntários do mundo inteiro) é, de longe, a plataforma traduzida de forma mais abrangente. A Wikipédia suporta mais de 400 idiomas uma interface de usuário básica, e há Wikipédias em cerca de 300 línguas com ao menos 100 artigos . O Google Search suporta 150 idiomas, enquanto o Facebook, conta com 70-100 línguas. Signal lidera os aplicativos de mensagens, com

quase 70 idiomas suportados no Android e 50 no iOS. Por outro lado, a maioria das plataformas concentra seu suporte a idiomas apenas em um pequeno número de línguas amplamente faladas, deixando a maioria dos idiomas sem suporte. O aplicativo de mensagens QQ, por exemplo, oferece suporte apenas para chinês.

O pequeno número de idiomas que tende a ser suportado pela maioria das plataformas pesquisadas inclui idiomas europeus — como inglês, espanhol, português e francês — e alguns idiomas asiáticos — como mandarim, indonésio (Bahasa Indonesia), japonês e coreano. Idiomas importantes, como árabe e malaio, têm menos suporte, e outros idiomas falados por dezenas a centenas de milhões não são bem representados no que diz respeito a suporte de interface.

O que essa falta de suporte ao idioma significa para a maioria das pessoas no mundo? Em 2021, a estimativa é que somos cerca de 7,9 bilhões de humanos no planeta, a maioria dos quais vive na Ásia (quase 4,7 bilhões) e na África (quase 1,4 bilhão). No entanto, a maior parte do mundo não é atendida pelas línguas da internet:

- *Falantes de línguas africanas*: a grande maioria das línguas africanas não é suportada como linguagem de interface por nenhuma das plataformas pesquisadas e, consequentemente, mais de 90% dos africanos precisam mudar para uma segunda língua para usar a plataforma. Para muitos, isso pode significar uma língua colonial europeia ou uma língua mais dominante em sua região.
- *Falantes de idiomas do sul da Ásia*: no sul da Ásia, quase metade das plataformas pesquisadas não oferece suporte de interface para nenhum idioma regional, e até mesmo os principais idiomas da região, como hindi e bengali (falados por centenas de milhões de pessoas) são menos suportados que outros idiomas.
- *Falantes de idiomas do sudeste asiático*: O suporte para idiomas do sudeste asiático funciona de maneira semelhante. Embora os idiomas indonésio, vietnamita e tailandês tendam a ter um suporte muito bom nas plataformas que pesquisamos, a maioria dos outros idiomas da região não é compatível com a maioria das plataformas pesquisadas.

As descobertas de Martin e Mark são intensificadas pela realidade cotidiana das pessoas que vivem nessas regiões do mundo. Donald, do Malaui, descobriu, por exemplo, que quando perguntava a falantes de chindali (uma língua bantu ameaçada de extinção) como eles se comunicavam com seus smartphones, todos eles relataram como custava tempo e trabalho se comunicar em chindali, já que a maioria de seus telefones eram integrados com suporte a idiomas em inglês, francês ou árabe e não reconheciam o chindali. Esses desafios tecnológicos, é claro, se somam às restrições econômicas e sociais que limitam a capacidade dos falantes de chindali de comprar um smartphone ou planos de dados. Mesmo para quem usa a língua nacional do Malaui, chichewa, a falta de suporte linguístico é um fator de dificuldade: “Por que devo comprar um telefone caro ou perder o meu tempo acessando a internet quando a língua é o inglês, que não consigo entender?”

Na verdade, a falta de suporte a idiomas para a maioria das línguas africanas foi notada de forma acentuada em 2018, quando o Twitter reconheceu o suaíli pela primeira vez, uma língua falada por mais de 50-150 milhões de pessoas na África Oriental e além desse território (como primeira ou segunda língua). Até então, o suaíli e a maioria das outras línguas africanas eram confundidos com o indonésio na plataforma. O reconhecimento de palavras em suaíli e o suporte à tradução também não foram iniciados pela empresa de tecnologia: foi, na verdade, o resultado de uma campanha de usuários do Twitter que falam suaíli.

A situação não é muito melhor na América Latina para as línguas indígenas. Na entrevista com o projeto Kimeltuwe trabalhando em mapuzugun, língua falada pelos povos mapuche do Chile e da Argentina atualmente, destacou-se que seria ótimo poder postar em mapuzugun em plataformas como YouTube ou Facebook. Isso nem tanto no caso de interfaces traduzidas, mas apenas poder marcar, nos menus que estão disponíveis, o idioma como mapuzugun. Por exemplo, ao fazer upload de um vídeo no YouTube ou no Facebook, você não pode adicionar uma transcrição em mapuzugun, já que a língua não aparece na lista de idiomas disponíveis. Então, se você quer fazer o upload de uma transcrição em mapuzugun, deve selecionar espanhol ou inglês."

Martin e Mark não analisaram o suporte a idiomas em dispositivos específicos, como telefones celulares, mas sabemos que os teclados digitais são um dos poucos espaços críticos em que linguistas e tecnólogos mais progrediram. O Gboard, teclado do smartphone do Google para o sistema operacional Android, por exemplo, suporta mais de 900 variedades de idiomas e se baseia em um trabalho significativo com comunidades e estudiosos de diferentes línguas. Ainda assim, só podemos acessar um teclado de smartphone com esses recursos através de um smartphone relativamente de última geração.

Ao mesmo tempo, a experiência de Uda com o cingalês — uma língua falada por mais de 20 milhões de pessoas no Sri Lanka como primeira ou segunda língua a experiência de Uda com o cingalês — mostra que ainda é difícil criar conteúdo em um idioma cuja escrita não seja facilmente compreendida por alguns tecnólogos que trabalham com suporte linguístico, especialmente se a forma for muito diferente da escrita latina das línguas da Europa Ocidental. Ela diz: "O principal problema do Unicode Sinhala está relacionado à ordem em que os diferentes caracteres devem ser inseridos para criar uma letra. Essa ordem requer que você insira diacríticos após as consoantes. Esse pensamento segue as regras das línguas europeias baseadas na escrita latina. No entanto, em cingalês, às vezes os diacríticos precedem a consoante."

O Unicode é o padrão de tecnologia para codificar o texto que é expresso no sistema de escrita ou script de um idioma. A versão 13 tem 143.859 caracteres para mais de 30 sistemas de escrita em uso atualmente, uma vez que o mesmo sistema de escrita pode ser usado para mais de um idioma (por exemplo, a escrita latina para a maioria dos idiomas da Europa Ocidental, a escrita

Han para os idiomas japonês, chinês e coreano, e a escrita Devanagiri para diferentes idiomas do sul da Ásia). Ele também contém caracteres para scripts históricos de idiomas que não estão mais em uso. O Unicode Consortium (uma organização sem fins lucrativos com sede na Califórnia) também toma decisões sobre os emojis — os símbolos que usamos todos os dias em diferentes interfaces.

A pesquisa de Martin e Mark sobre suporte a idiomas e as experiências de outros colaboradores de todo o mundo acrescentam mais detalhes a esta breve descrição do suporte técnico desigual para a maioria dos idiomas em plataformas e aplicativos agora. Vale a leitura!

Conteúdo linguístico: acessibilidade e produção

O conteúdo feminista é particularmente inacessível em línguas locais. A Women's Development Foundation é um grupo de mulheres rurais que trabalha com direitos das mulheres desde 1983. Mas foi somente em 2019 que começamos a compartilhar conteúdo feminista em língua cingalesa sobre questões sociopolíticas e econômicas online.
Uda Deshapriya

Infelizmente era — e ainda é — muito difícil encontrar conteúdo queer educativo e positivo em indonésio na internet. Se pesquisarmos "LGBT" ou "homoseksualitas" (homossexualidade) no Google — o maior e mais popular mecanismo de pesquisa —, encontraremos muitos resultados contendo a palavra "penyimpangan" (desvio), "dosa" (pecado) e "penyakit" (doença).
Paska Darmawan

As informações sobre a interseção entre queer e deficiência (ou mesmo a ausência dessas informações) disponíveis em bengali na internet é em grande medida moldada por homofobia e capacitismo e, por sua vez, um instrumento para a criação deles.
Ishan Chakraborty

Queríamos entender o conteúdo da internet analisando qual versão do mundo e o conhecimento de quem nós experimentamos quando estamos online. Afinal, mais de 63% de todos os sites usam o inglês como principal idioma de conteúdo.

Em seus ensaios e entrevistas, nossos colaboradores discutem as diferentes constelações de desafios históricos, sociopolíticos, econômicos e tecnológicos para um acesso significativo à internet em seus idiomas. Mais importante, todos eles focam nos desafios de encontrar conteúdo relevante na internet em suas línguas e criar conteúdo significativo para eles nessas línguas. Em outras palavras, para nós, não é suficiente poder acessar informações e conhecimentos criados para nós em outras línguas por pessoas que podem não compreender nossos contextos e experiências e, pior ainda, ser hostis a eles. Precisamos poder produzir conhecimentos significativos para nós e as nossas comunidades ou, pelo menos, apoiar a produção e expansão desse conteúdo em todas as nossas diversas línguas.

Esse é particularmente o caso para aqueles que têm problemas de acessibilidade e experimentam diferentes formas interligadas de marginalização e exclusão.

Como Joel nos disse na entrevista sobre o projeto Indigemoji, tudo começou com um tweet frustrado feito em seu carro, em que ele simplesmente parou no acostamento da estrada um dia e começou a colocar palavras em arrernte ao lado dos emojis para descrever seu significado. Por décadas, depois que os emojis foram introduzidos pela primeira vez na internet, os Povos Originários ou povos indígenas fizeram petições sem sucesso para que eles expressassem suas línguas orais e visuais, como arrernte. Como dissemos antes, é o Consórcio Unicode que delibera a respeito de solicitações públicas de novos emojis, e petições como a por um emoji para a Bandeira Aborígine Australiana foram rejeitadas. Para Joel, Caddy e outros, o projeto Indigemoji foi um esforço intergeracional para superar essas várias formas de marginalização física e virtual e criar seu próprio conteúdo de maneira significativa para suas identidades e línguas indígenas.

Tweet com uma lista de emojis com palavras arrernte ao lado. Fonte: Indigemoji

É importante lembrar que as línguas indígenas são línguas "minoritárias" em nosso mundo hoje por causa das histórias de genocídio em massa pela colonização, pela qual as nações indígenas foram destruídas ou se tornaram populações minoritárias mesmo sendo os primeiros habitantes de uma determinada região ou território. E esses processos de colonização também afetam as línguas dominantes, faladas por milhões de pessoas em todo o mundo.

Ishan é um acadêmico queer com deficiência visual. Para ele, estar online já é, em si, um esforço contra todos os prognósticos. Além disso, ele se esforça para encontrar informações relevantes em bengali sobre deficiência, sobre queerness e, mais ainda, sobre as interseções entre essas questões. Isso leva ao que ele chama de 'marginalidade dentro da marginalidade':

“Por um lado, as atitudes homofóbicas e capacitistas da sociedade e, por outro, a homofobia internalizada e/ou o capacitismo de indivíduos (queer e/ou deficientes). Juntas, essas condições complementares perpetuam o mecanismo de marginalização. A localização social de um indivíduo queer com deficiência pode ser descrita como ‘marginalidade dentro da marginalidade’.”

Em outras palavras, processos críticos de acesso e informação, mesmo em uma língua dominante como o bengali — falado por cerca de 300 milhões de pessoas em todo o mundo — estão ausentes da internet.

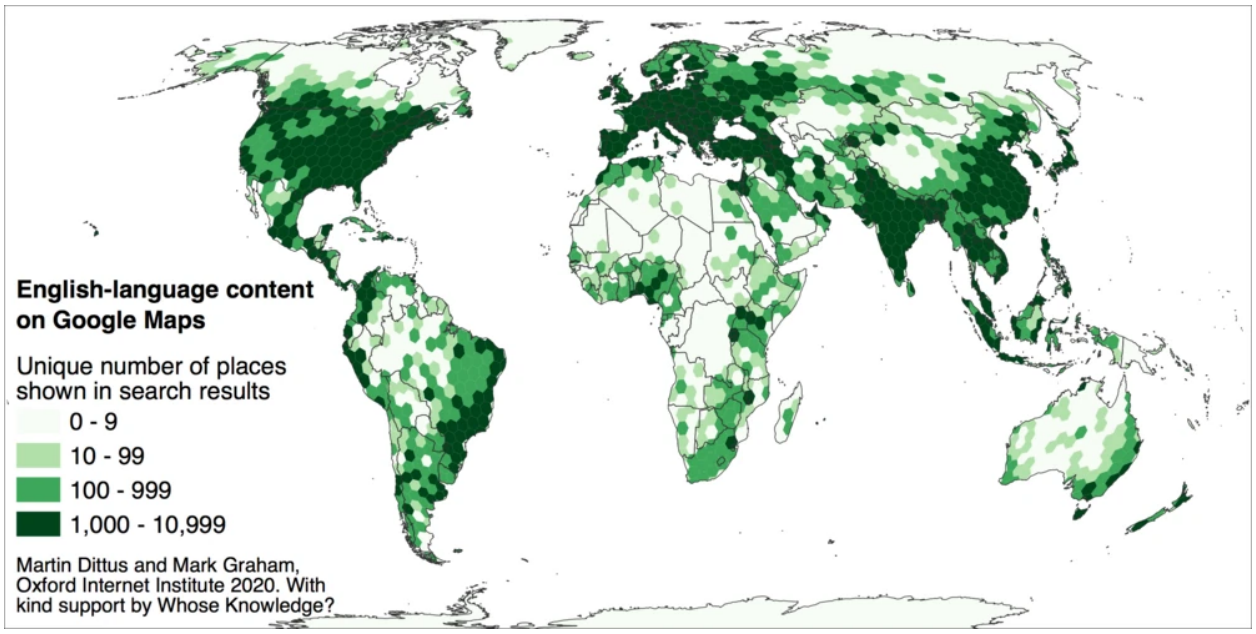
Martin e Mark decidiram se aprofundar na análise da extensão e tipo de conteúdo em diferentes idiomas, em dois tipos diferentes de plataformas de informação e conhecimento: Google Maps e Wikipédia.

Google Maps

Podemos acessar o Google Maps em todas as nossas diversas línguas? O idioma que usamos muda a versão do mundo que vemos através do Google Maps?

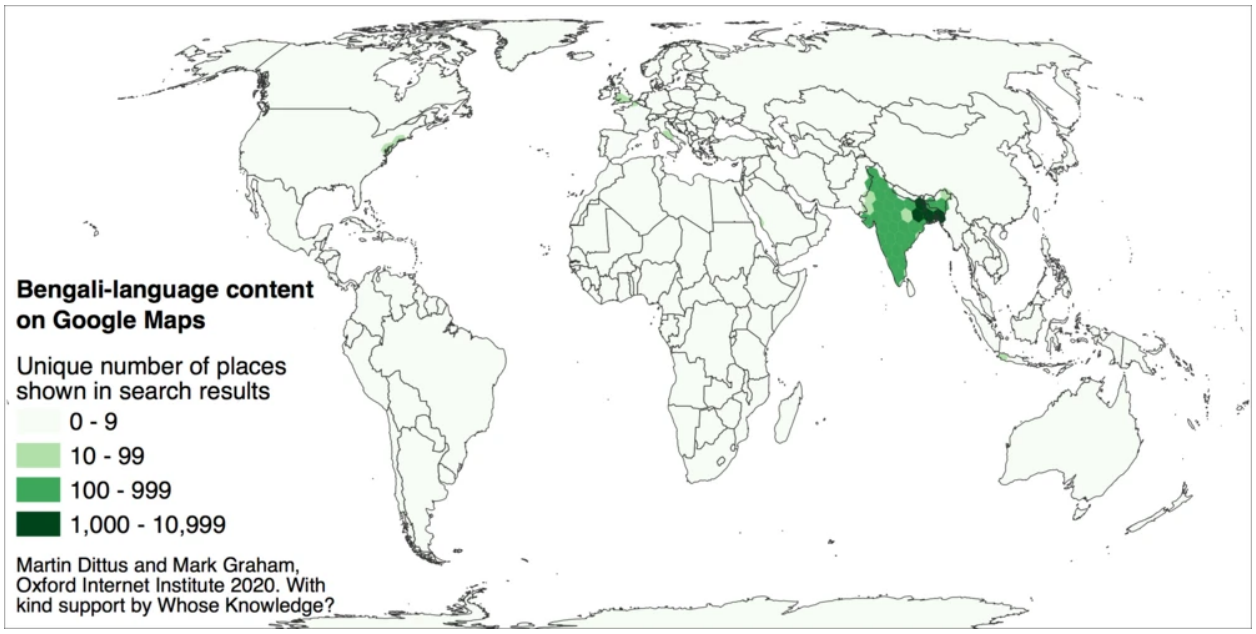
Para responder a essas perguntas, Martin e Mark coletaram dados sobre a cobertura de conteúdo global do Google Maps nos 10 idiomas mais falados: inglês, chinês mandarim, hindi, espanhol, francês, árabe, bengali, russo, português e indonésio (Bahasa Indonesia). Eles coletaram dezenas de milhões de resultados de busca individuais nesses idiomas e, entre eles, identificaram e mapearam cerca de três milhões de lugares individuais (lugares determinados e outras localizações).

Não é de surpreender que os mapas englobam mais conteúdo quando acessamos o Google Maps em inglês. O mapa em inglês do Google cobre o mundo inteiro, embora seja muito mais denso (ou seja, contenha mais informações) no Norte Global, com foco especial na Europa e na América do Norte. Ele também cobre o sul da Ásia e regiões do sudeste asiático, bem como grandes porções da América Latina, relativamente bem. Por outro lado, porém, muitas partes da África têm conteúdo comparativamente mais esparso.



Densidade de informações do Google Maps para falantes de inglês. As cores mais escuras indicam onde os resultados de busca incluem maior número de locais.

Em comparação com o mapa em inglês com cobertura relativamente boa, descobrimos que o bengali (que é a primeira língua de Ishan) está no outro extremo: sua cobertura é restrita principalmente ao sul da Ásia, especialmente Índia e Bangladesh, e o Google Maps tem pouco ou nenhum conteúdo para falantes de bengali no resto do mundo. Para descobrir conteúdos adicionais e navegar em lugares fora da Índia e Bangladesh, os falantes de bengali precisam mudar para um segundo idioma, como o inglês. Esse também é o caso do Google Maps em hindi (o terceiro idioma mais falado no mundo, depois do inglês e do mandarim).



Densidade de informações do Google Maps para falantes de bengali. As cores mais escuras indicam onde os resultados de busca incluem maior número de locais.

Aqui você pode ler muito mais sobre o Google Maps em diferentes idiomas no ensaio detalhado de Martin e Mark.

Wikipedia

Como mostrado na pesquisa sobre plataformas de Martin e Mark, a Wikipédia está na vanguarda do suporte a idiomas na internet, com sua interface de usuário traduzida para mais idiomas do que qualquer outra plataforma comercial que examinamos, incluindo Google e Facebook.

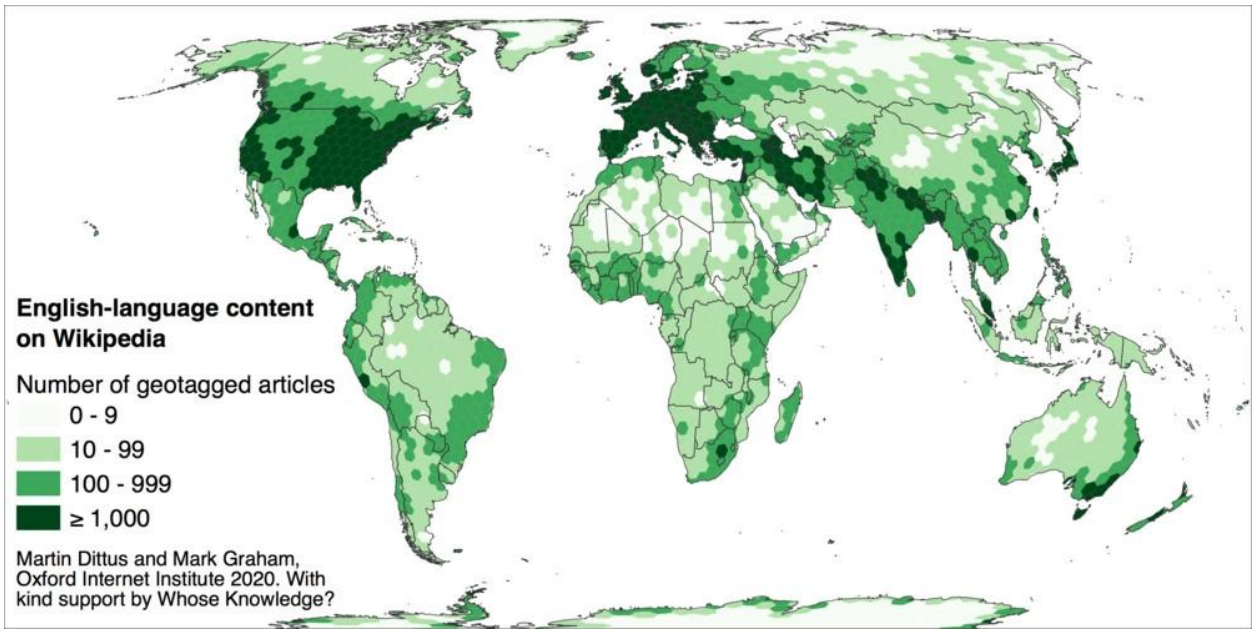
Em termos de conteúdo real — informações e conhecimento em artigos da Wikipédia —, a Wikipédia tem mais de 300 idiomas de edição. Mesmo assim, os falantes dessas línguas não têm acesso ao mesmo conteúdo ou à mesma quantidade de informações. Queríamos ir mais fundo em perguntas e respostas: Qual é a qualidade de cobertura de conteúdo da Wikipédia em seus idiomas de edição? Alguns idiomas estão mais bem representados do que outros? Alguma comunidade linguística tem acesso a mais conteúdo do que outras? Respondemos a algumas dessas perguntas em detalhes na análise da Wikipédia de Martin e Mark.

Usamos dados de 2018 com geotags (uma forma de incorporar referências geográficas, como coordenadas, nos artigos da Wikipédia) e analisamos o número de artigos e o crescimento do conteúdo em diferentes idiomas. Também baseamos nossa análise em idiomas "locais", definidos como os idiomas classificados como idioma oficial em Unicode CLDR (o código que oferece suporte a idiomas na internet), ou que estão em uso por pelo menos 30% da população em qualquer país.

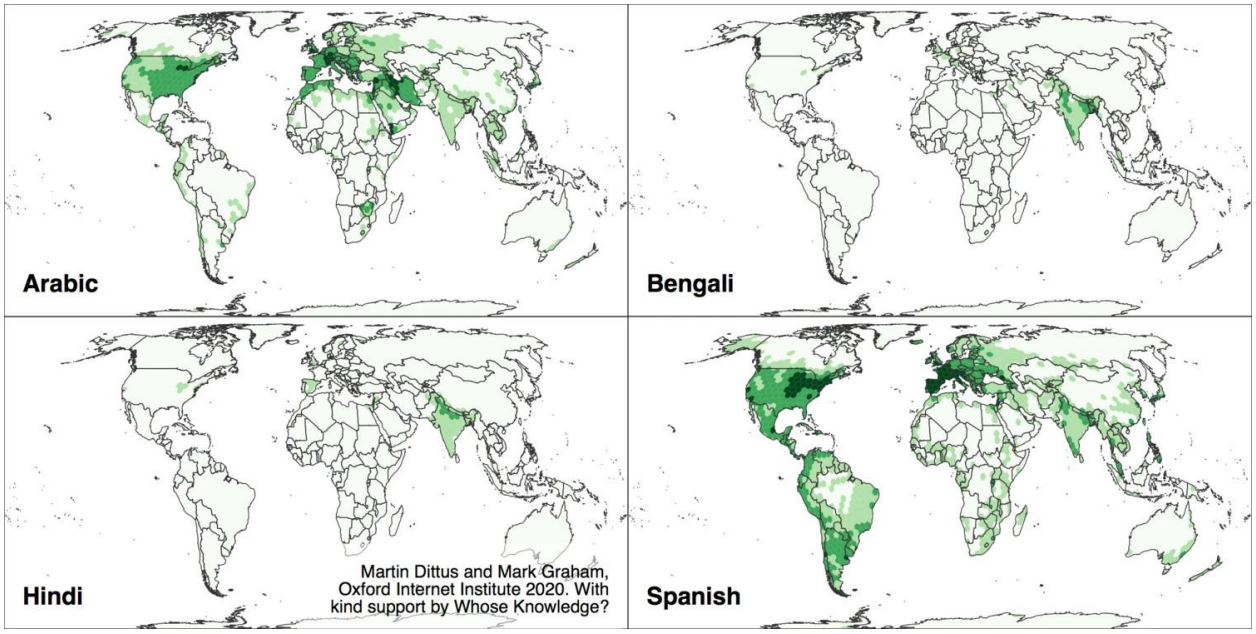
Em seguida, identificamos o idioma local mais prevalente, ou seja, falado pelo maior número de pessoas em cada país. Encontramos 73 destes, os mais prevalentes em pelo menos um país. O inglês é o mais falado e mais prevalente em 34 países. É seguido por árabe e espanhol (18 países), francês (13 países), português (sete países), alemão (quatro países) e holandês (três países). Chinês, italiano, malaio, romeno, grego e russo são os idiomas mais prevalentes em dois países e os 60 idiomas restantes são os mais prevalentes em um único país.

Para comparar essa distribuição em um idioma local com o conteúdo da Wikipédia em cada país, identificamos a edição no idioma da Wikipédia com o maior número de artigos sobre aquele país. Encontramos uma propensão para o conteúdo em inglês, o idioma dominante da Wikipédia em 98 países. Ele é seguido pelo francês (nove países), alemão (oito países), espanhol (sete países), catalão e russo (quatro países), italiano e sérvio (três países) e holandês, grego, árabe, servo-croata, sueco e romeno (dois países). Os 21 idiomas restantes da Wikipédia são mais prevalentes em um único país.

Embora o número de artigos em cada idioma da Wikipédia seja dinâmico e continue crescendo, ficou claro que as línguas de edição da Wikipédia variam muito em tamanho e escala — tanto em relação ao número de artigos quanto em relação ao tamanho de suas comunidades de editores. A Wikipédia em inglês é de longe a maior, com mais de seis milhões de artigos e quase 40 milhões de colaboradores inscritos. As maiores comunidades de colaboradores depois do inglês estão em espanhol, alemão e francês, cada uma tendo entre quatro e seis milhões de colaboradores e cerca de dois milhões de artigos. As edições em outros idiomas são pequenas em comparação com estas: apenas cerca de 20 idiomas de edição têm mais de um milhão de artigos e apenas 70 têm mais de 100.000 artigos. A maioria das línguas de edição da Wikipédia tem apenas uma pequena fração do conteúdo da Wikipédia em inglês.



Densidade de informações da Wikipédia em inglês no início de 2018. As cores mais escuras indicam maior número de artigos marcados geograficamente.

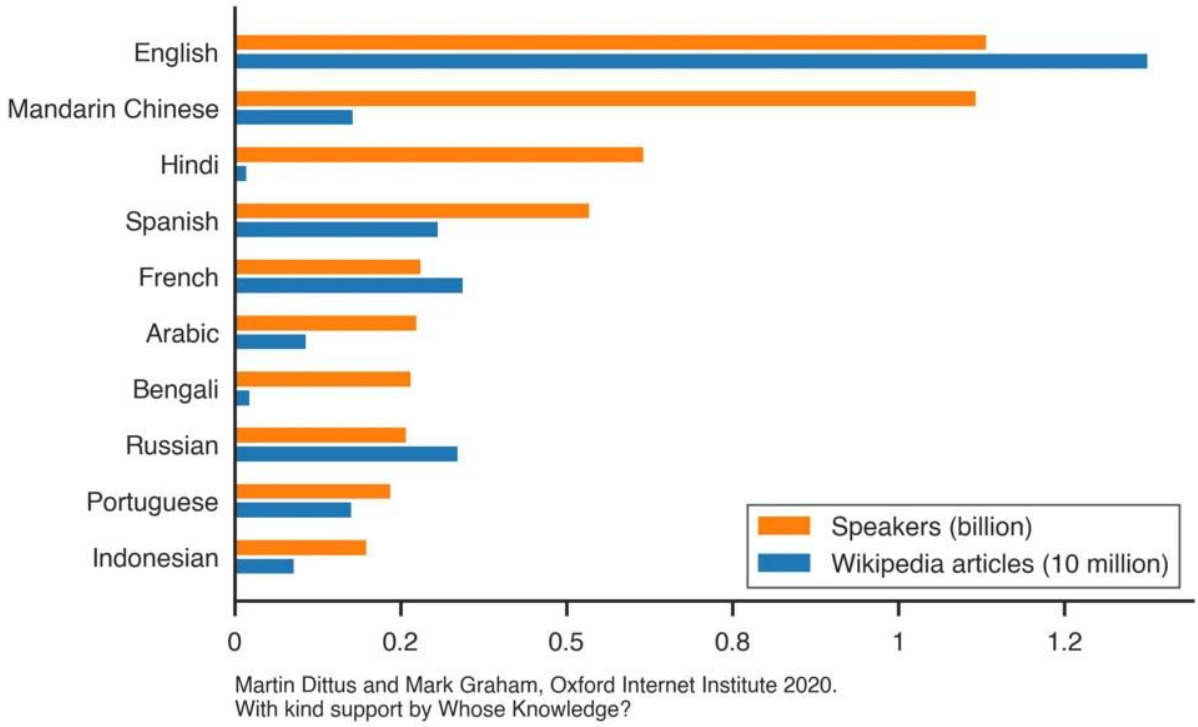


Densidade de informações das Wikipédias em árabe, bengali, hindi e espanhol no início de 2018. As cores mais escuras indicam maior número de artigos marcados geograficamente.

O que impressiona é como o conteúdo nas Wikipédias em vários idiomas ecoa a mesma distribuição do Google Maps que vimos anteriormente.

Isso também se aplica quando examinamos o número de artigos em Wikipédias de diferentes idiomas em comparação com o número de falantes dessas línguas (como primeira e segunda

línguas). Verificamos que em idiomas europeus como inglês, francês, espanhol, russo e português, o número de artigos nessas Wikipédias é proporcional ao número de falantes. Mas esse não é o caso quando se trata de outras línguas amplamente faladas: chinês mandarim, hindi, árabe, bengali e indonésio (Bahasa Indonesia) são falados por centenas de milhões de pessoas, mas suas edições da Wikipédia são muito menores, com um número menor de artigos em comparação às edições em línguas europeias. Existem mais artigos na Wikipédia em francês, espanhol ou português do que nas versões em mandarim, hindi ou árabe, embora estes estejam entre os cinco principais idiomas falados do mundo, com mais falantes que o francês e o português.



Conteúdo da Wikipédia e número de falantes das 10 línguas mais faladas no mundo. (Estimativa populacional: Ethnologue 2019, que incluem falantes de uma segunda língua.)

O ensaio de Martin e Mark tem muito mais análises e visualizações de dados na Wikipédia em diferentes idiomas, mas o que fica claro a partir desses números é o que é ecoado pelas experiências vividas por nossos colaboradores em todo o mundo.

As marginalizações e exclusões de línguas que não são as coloniais europeias são profundas nos mundos físico e virtual, incluindo línguas dominantes e globais como o árabe. Para escrever a Wikipédia em seu próprio idioma, usando referências baseadas no contexto daquele idioma, precisamos de fontes publicadas extensas e confiáveis, o que, como descobrimos anteriormente, é raro na maioria dos idiomas do mundo. Como wikipedista, Emna fala sobre como é difícil encontrar recursos e referências em diferentes línguas pela África: "...a luta para

mim, por exemplo, como wikipedista, para encontrar referências em nossas próprias línguas — e quando digo 'nossas próprias línguas', não é apenas a língua tunisiana, nosso dialeto, ou a língua árabe — mas também quando percorremos a África, é encontramos uma enorme lacuna de recursos e referências."

Mesmo na Europa, os falantes de línguas minoritárias têm dificuldade em usar ou editar Wikipédias em seus idiomas. Enquanto Claudia descobriu que muitos de seus entrevistados em bretão estavam cientes da existência de uma Wikipédia bretã, sendo que "19% deles até contribuem para ela, editando artigos existentes ou escrevendo novos (8%)", ela tem a impressão de que a maioria dos falantes de línguas minoritárias muda para uma língua dominante porque é mais fácil: "A disponibilidade não implica que os serviços, interfaces, aplicativos e Wikipédias sejam realmente usados. Alguns estudos revelam que falantes de línguas minoritárias mudam facilmente para sua língua dominante ao usar tecnologias digitais baseadas em linguagem, seja porque a tecnologia é inerentemente melhor, ou porque a gama de serviços disponíveis é muito mais ampla."

A Wikipédia e sua miríade de projetos de conhecimento gratuitos e de fonte aberta (em que o código está disponível abertamente e é construído coletivamente) constituem um dos espaços mais promissores e úteis para conhecimento multilíngue online. Por exemplo, suas comunidades de voluntários sabem e entendem que não existe uma única variante de inglês, árabe ou chinês, mas expressar essa pluralidade de contexto e conteúdo de linguagem nem sempre é fácil. Como podemos ver em nossas análises e experiências, a Wikipédia também está sob estruturas históricas e contínuasde poder e privilégio que distorcem os caminhos e formas pelas quais criamos e compartilhamos conhecimento em diferentes idiomas, bem como no âmbito das famílias de idiomas.

Quais são alguns caminhos a seguir para nós como indivíduos, organizações e comunidades que desejamos uma internet mais multilíngue? Nas próximas e últimas seções, nós nos baseamos nas descobertas feitas a partir de todos os números e histórias que compartilhamos com você até agora para traçar um panorama geral do que aprendemos, assim como os contextos, entendimentos e ações que podem nos levar a uma internet verdadeiramente multilíngue.

O que aprendemos sobre a internet multilíngue?

Aprendemos muito sobre idiomas, sobre a internet e idiomas na internet ao trabalharmos neste relatório. Aqui está uma breve visão geral das percepções mais importantes da nossa jornada até agora.

Aprendizagem: A língua é um proxy do conhecimento e uma forma essencial de estar no mundo, não apenas uma ferramenta de comunicação. É por isso que o multilinguismo é tão importante: para que possamos honrar e afirmar a grande riqueza e as texturas de nossos muitos eus e de nossos diferentes mundos de modo mais forte.

Contexto: As pessoas conhecem seus mundos e se expressam em mais de 7.000 idiomas. Nossos idiomas podem ser orais (falados ou sinalizados), escritos, ou transmitidos por sons. .

Ainda assim, o suporte a idiomas nas principais plataformas de tecnologia e aplicativos só existe para uma fração desses 7.000 idiomas, estando somente cerca de 500 deles representados online em qualquer tipo de informação ou conhecimento. Algumas das línguas mais faladas no mundo não contam com muito suporte linguístico ou informações online. O suporte a idiomas mais rico, as informações mais extensas na internet (incluindo no Google Maps e Wikipédia) e a maioria dos sites estão em inglês.

Reflexão: A internet não é, nem de longe, tão multilíngue quanto imaginamos ou precisamos que seja.

Análise: A maioria das pessoas precisa usar o idioma colonial europeu mais próximo (inglês, espanhol, português, francês etc.) ou idioma regionalmente dominante (chinês, árabe etc.) para acessar a internet. Estruturas históricas e contínuas de poder e privilégio são intrínsecas à maneira como as línguas estão acessíveis (ou não) online.

Como podemos melhorar?: Contextos e ações para uma internet multilíngue

Na maioria dos casos, usar uma língua minoritária requer uma boa dose de perseverança, vontade e resiliência, uma vez que a experiência do usuário no uso destas é intercalada com falhas e dificuldades.
Claudia Soria

O objetivo de uma internet multilíngue inclusiva e representativa dos povos indígenas deve ponderar e levar em consideração os legados sociais e a experiência contínua de opressão colonial. Uma internet multilíngue não pode meramente aspirar a ser representativa, mas, considerando a história colonial, também deve buscar promover ambientes que aumentem a sobrevivência e o aprendizado de línguas indígenas para e pelos povos indígenas.
Jeffrey Ansloos and Ashley Caranto Morford

Jovens e crianças mapuche crescem lado a lado com a tecnologia e a internet. Ela é um espaço onde essas pessoas podem se aproximar do mapuzugun... As histórias do nosso povo têm que ser escritas, e precisam ser escritas ou faladas em mapuzugun... Nossas histórias não são necessariamente as de grandes heróis, como as celebradas pelos Estados coloniais e pós-coloniais. Nossa história é a história de cada mapuche que sobreviveu à adversidade e à violência. A mulher que teve que migrar para a cidade para ganhar um salário, as mulheres e os homens que voltaram da cidade, mas não havia mais lugar nos seus "lofs" e tiveram que voltar para as cidades com suas raízes cortadas, sem nenhum lugar para onde voltar. Essas experiências e as memórias de cada mapuche constituem nossa memória coletiva como povo.
Kimeltuwe project

Nesta seção de nosso resumo do Relatório do Estado das Línguas na Internet, reunimos as diferentes percepções dos nossos colaboradores, bem como da nossa Convenção sobre a Descolonização das Línguas da Internet de 2019 para entender os diferentes contextos, desafios e oportunidades relativos ao multilinguismo no mundo e online. Fazer quatro perguntas-chave a nós mesmos e uns aos outros pode nos apontar caminhos para imaginar e projetar uma internet muito mais multilíngue.

- **De quem são o poder e recursos?**
- **De quem são os valores e conhecimentos?**
- **De quem são as tecnologias e padrões?**
- **De quem são os projetos e imaginações?**

Poder e recursos de quem?

Contextos

Tanto a nossa análise estatística quanto as experiências vividas pelas pessoas deixam claro que a marginalização das línguas nos mundos físico e virtual não se baseia apenas no número de pessoas que as falam.

As línguas indígenas são faladas por cidadãos de mais de 6.000 nações indígenas globalmente, que foram os primeiros habitantes de grande parte do mundo até que a colonização e o genocídio destruíssem ou reduzissem seus povos e línguas. As línguas dominantes em alguns dos continentes com maior diversidade linguística, como Ásia e África, faladas e sinalizadas por milhões de pessoas, não são bem representadas ou, às vezes, não são representadas online. Se considerarmos a diáspora de pessoas que falam línguas como as muitas variedades diferentes de árabe, chinês, hindi, bengali, punjabi, tamil, urdu, indonésio (Bahasa Indonesia), malaio, suaíli, hausa e outras, em diferentes países e continentes, é evidente que são línguas marginalizadas online, mesmo que possam dominar outras em suas próprias regiões.

Essas formas de marginalização e exclusão digital não são acidentais, mas causadas por estruturas e processos históricos e contínuos de poder e privilégio. Elas também indicam que os recursos que vão para a infraestrutura linguística — por parte de editoras, universidades, governos e empresas de tecnologia — já estão direcionados para certas regiões (Europa e América do Norte) e idiomas (inglês e outras línguas da Europa Ocidental). Mesmo na Europa e na América do Norte, as comunidades indígenas, negras e outras comunidades marginalizadas têm dificuldade em preservar suas línguas ao longo das gerações.

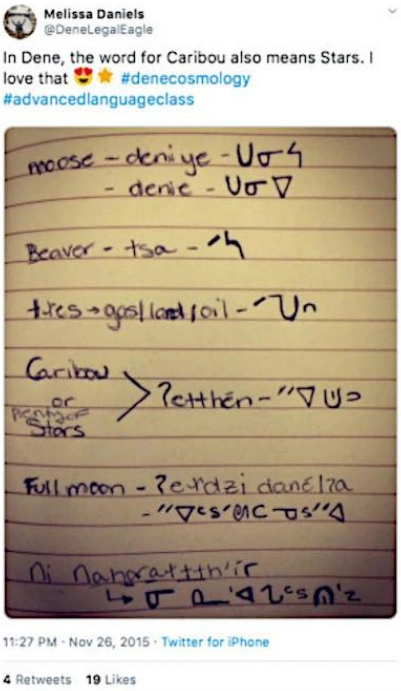
Particularmente, as forças da colonização e do capitalismo causam outros sistemas de discriminação e opressão, como racismo, patriarcado, homofobia, capacitismo, classismo e casteísmo, ao mesmo tempo que se conectam com eles. Isso significa que certas línguas — principalmente as coloniais europeias — são as mais proeminentes online, sejam ou não faladas pelo maior número de pessoas no mundo. Também significa que quando há informações e conhecimento em outras línguas mais marginalizadas, o conteúdo nessas línguas é limitado por quem tem acesso e poder para criá-lo ou impedir que outros produzam informações alternativas. Como exemplo, citamos a falta de conteúdo online feminista em cingalês ou a falta de conteúdo positivo para pessoas queer e com deficiência em bengali ou indonésio (Bahasa Indonesia).

Como a linguagem está no centro de quem somos, não conseguir nos expressar em nossas próprias línguas e na plenitude de nossos muitos eus é uma forma de violência. Mas as consequências dessa marginalização são violentas de outras maneiras. Como Uda coloca, “a falta de conteúdo digital feminista, favorável aos direitos humanos e respeitoso torna aqueles espaços online, onde a comunicação ocorre primordialmente em línguas locais hostis a

mulheres, pessoas queer e minorias. É óbvia a falta de mídias digitais que se oponham à narrativa dominante, que permanece contaminada com estereótipos negativos. Isso agrava o discurso de ódio e a violência sexual e de gênero online." Essas formas de violência se estendem aos indígenas, membros de castas oprimidas e comunidades religiosas minoritárias, a pessoas com deficiência e outras comunidades marginalizadas.

Ao mesmo tempo, essas comunidades estão usando a internet para lutar contra as diferentes formas de violência transgeracional contra elas e suas línguas. Jeffrey e Ashley analisaram cerca de 3.800 tweets com mais de 35 hashtags e 57 palavras-chave em línguas indígenas, em 60 grupos de línguas indígenas reconhecidos federalmente no Canadá. Eles descobriram que, por meio de hashtags no Twitter, os povos indígenas em todo o Canadá e em outras partes do mundo estão se conectando ativamente, participando de um renascimento das línguas indígenas e colaborando com seu florescimento.

Como eles dizem, "em um contexto social em que as línguas indígenas em todo o Canadá foram alvo de apagamento sob políticas coloniais de assimilação, as redes de hashtags do Twitter reuniram um contexto único e significativo para os povos indígenas compartilharem conhecimento sobre suas línguas. Nas várias redes incluídas em nosso estudo, há exemplos de recuperação e reconexão linguísticas para sobreviventes intergeracionais das políticas de assimilação colonial."



Ao dar permissão para citar este tweet, Melissa Daniels, do povo Dene, gostaria de reconhecer à professora de idiomas que ofereceu esse ensinamento, a anciã Dene e educadora Eileen Beaver.

Ações

- Reconhecer estruturas e processos de poder e privilégio em diferentes instituições e processos que suportam infraestruturas linguísticas.
- Garantir recursos de reparação para línguas e comunidades marginalizadas, incluindo aprendizagem e programação de linguagens.
- Viabilizar recursos para a criação e expansão de informações e conhecimentos de e para comunidades que vivenciam diversas formas de opressão e violência, com interseção em suas línguas e formas de escolha.

Valores e conhecimentos de quem?

Contextos

A história e a tecnologia da internet são baseadas em uma visão de mundo das epistemologias — ou formas de saber e fazer — ocidentais. Mais especificamente, a internet foi e continua sendo projetada e governada principalmente por homens brancos (e agora também homens não-brancos privilegiados, como os indianos que trabalham no Vale do Silício) privilegiados. Isso significa que os valores que estão mais frequentemente no cerne das arquiteturas e infraestruturas da internet são os valores do determinismo tecnológico — em que a tecnologia é vista como sendo a causa primária (e benéfica) de quaisquer mudanças sociais — e individualismo, em que o principal foco e força motriz é o indivíduo, não o coletivo.

Além disso, essa visão de mundo está enraizada na Era do Iluminismo, movimento do século XVIII no Norte Global que privilegia um certo tipo de ciência e tecnologia baseadas na racionalidade. O que nos esquecemos é que a matemática e as ciências floresceram em todo o Sul Global muito antes do século XVIII. Os primeiros sistemas de escrita e numeração, por exemplo, vieram da Mesopotâmia, atuais Irã e Iraque. O que é ainda mais problemático é nos esquecermos de que os recursos para o Iluminismo, esta "era de ouro" da ciência e tecnologia no Norte Global, vieram da Era dos Impérios no Sul Global, com seus movimentos em massa de colonização, escravidão, genocídio e extração de recursos da Ásia, África, América Latina e Ilhas do Caribe e do Pacífico. As bases para a natureza extrativista do capitalismo moderno estão enraizadas nessas histórias passadas de colonização e continuam a fazer parte do capital tecnológico.

Juntamente com os recursos materiais, o que foi destruído, ignorado ou minado nesses processos foram outras formas de conhecer, fazer e ser, isto é: epistemologias não-ocidentais, como conhecimentos indígenas, ou os conhecimentos dos povos de partes menos privilegiadas do mundo. A consequência mais devastadora disso para a linguagem — que é um proxy significativo para o conhecimento, como dissemos antes — foi a desvalorização completa das línguas da Europa não-ocidental e a destruição ativa ou negligência bem intencionada das formas orais e não-escritas de línguas. A tendência para o conteúdo escrito em um pequeno

conjunto de línguas privilegiadas continua a nos direcionar para um certo tipo de "conhecimento" escrito no mercado editorial e na academia, o que afeta a documentação e os dados subjacentes para processamento de linguagem natural ou alguns dos sistemas de linguagem automatizados que fazem parte das infraestruturas da internet, como o Google Tradutor.

Como Ana descreve, "apesar do fato de que pouco menos da metade das línguas do mundo são línguas que não têm um sistema de escrita e mantêm uma longa tradição oral, os idiomas com um sistema alfabético amplamente reconhecido e difundido dominam a web. A web reforça uma exclusão sistemática em que apenas as línguas escritas podem ser preservadas para o futuro".

Com a perda da linguagem, estamos perdendo mais do nosso futuro do que imaginamos. Estamos perdendo as formas de expressão em diferentes idiomas e visões de mundo e conhecimentos inteiros inerentes a essas línguas. Em um momento em que a humanidade está à beira do colapso planetário, as comunidades indígenas e seus conhecimentos estão protegendo nossa biodiversidade e preservando a vida como a conhecemos. Não é de surpreender que a perda de línguas está diretamente ligada à perda de biodiversidade e destruição de ecossistemas ao redor do mundo.

A internet pode ser uma infraestrutura fascinante para a preservação e expansão de diferentes formas de linguagem e conhecimento porque suas formas variadas de mídia podem imitar e representar línguas faladas, sinalizadas, e que vão além do texto. Mas é fundamental que essa promessa radical da internet não seja mais uma vez baseada em valores patriarcais e capitalistas coloniais. Como as comunidades de pessoas preservam e revitalizam suas línguas e identidades e compartilham seus conhecimentos em seus próprios termos? Por exemplo, em muitas comunidades indígenas, certos conhecimentos são sagrados e não devem ser compartilhados abertamente.

Um desses esforços liderados pela comunidade é o Papa Reo, uma tecnologia de reconhecimento voz para te reo māori (o idioma māori) de Aotearoa/Nova Zelândia. A comunidade māori criou e mantém a tecnologia e os dados para esta iniciativa e acredita que esta forma de soberania sobre os dados é fundamental para garantir que o conhecimento compartilhado por meio da linguagem seja usado para e por māoris, em vez de por empresas que buscam lucro. Curiosamente, mesmo reconhecendo o valor da tecnologia de código aberto, a equipe do Papa Reo decidiu não adicionar seus dados aos bancos de dados de código aberto, pois os recursos e privilégios da maioria das comunidades de código aberto foram negados à comunidade māori. Por outro lado, a Mukurtu é um exemplo de plataforma de código aberto construída com comunidades indígenas para gerenciar seus próprios dados de idioma.

Ações

- Criar infraestruturas linguísticas para a internet para o bem público, projetadas com valores coletivos e comunitários em seu cerne, que centralizem os conceitos feministas e indígenas de soberania e corporeidade, bem como colaborar com elas e compartilhá-las.
- Continuar desafiando e criticando as infraestruturas de tecnologia de linguagem que são opressoras por natureza: capitalistas, proprietárias, humanamente extrativas e ambientalmente destrutivas.
- Reconhecer que as tecnologias de linguagem livres e de código aberto também precisam prestar atenção a seus próprios privilégios relativos e honrar a maneira como as comunidades marginalizadas se autodeterminam, definem o que é "aberto" e o que desejam compartilhar com o mundo.

Tecnologias e padrões de quem?

Contextos

A indústria de tecnologia não é totalmente responsável pela atual falta de representação e suporte para a grande maioria dos idiomas do mundo. No entanto, a tecnologia do Norte Global é responsável por continuar mantendo e aumentando as desigualdades baseadas na linguagem e o colonialismo digital online.

As grandes empresas de tecnologia — que estão projetando e criando a maior parte das plataformas digitais, ferramentas, hardwares e softwares que usamos — podem ignorar a necessidade de criar uma internet verdadeiramente multilíngue, porque não consideram e não veem a maioria das nossas 7000 línguas faladas como parte essencial da infrestrutura da internet. Afinal, elas sabem que precisam fornecer suporte linguístico apenas quando isso fizer sentido para os negócios: seja para as línguas coloniais europeias, seja para as línguas do que chamam de "mercados emergentes". Na verdade, as línguas dominantes do sul e do sudeste asiático estão começando a ter um suporte melhor a idiomas nas grandes plataformas de tecnologia à medida que se tornam a maior base de clientes dessas empresas.

Ao mesmo tempo, algumas das tecnologias de linguagem mais difundidas que temos na internet são criadas e controladas por essas empresas, pois são elas que têm os recursos e a capacidade para isso. A Wikipédia é uma exceção notável porque é de código aberto e tem o apoio de suas comunidades de voluntários em todo o mundo. Em geral, as motivações proprietárias e com fins lucrativos não resultam nas ferramentas e tecnologias necessárias para um conteúdo profundo e com nuances em diferentes línguas marginalizadas de comunidades marginalizadas. Pior ainda, as tecnologias de linguagem que estão sendo construídas atualmente por essas empresas são sistemas automatizados de grande escala que dependem de

grandes quantidades de dados de linguagem de qualquer tipo de fonte — mesmo que essa fonte seja um discurso violento e cheio de ódio ou um texto escrito contra grupos marginalizados.

Como relatam Jeffrey e Ashley, "em toda as redes sobrevivência de aprendizagem e línguas indígena, o racismo foi identificado como um grande desafio social dentro da ecologia do Twitter. Existem várias maneiras específicas em que essas redes são ativamente direcionadas por textos e conteúdo multimídia incendiários, que algumas vezes constituem discurso de ódio, por uma vasta gama de usuários, que incluem pessoas reais e bots automatizados. No que se refere a usuários de bot, essas contas parecem seguir padrões semelhantes de disseminação de informações incorretas e, muitas vezes, distribuem textos sem sentido que refletem análises agregadas e geração de conteúdo".

Se uma empresa não se preocupa o suficiente com as consequências de sua falta de conhecimento e contexto do idioma local, isso pode causar danos incomensuráveis e violência real. Em Mianmar, onde o Facebook (ou Meta) é praticamente a internet, ativistas alertaram a empresa sobre o discurso de ódio por anos antes de ela formar uma equipe de língua birmanesa. Em 2015, o Facebook tinha quatro falantes de birmanês que revisavam o conteúdo para 7,3 milhões de usuários ativos em Mianmar. A que essa falta de atenção à linguagem e ao contexto levou? As Nações Unidas descobriram que o Facebook foi parte do genocídio contra os muçulmanos Rohingya em Mianmar, e a empresa está no centro de um caso contra o governo de Mianmar no Tribunal Internacional de Justiça.

Igualmente, o discurso de ódio na Índia contra muçulmanos, dalits e outras comunidades marginalizadas continuam contando com muito pouca moderação ativa, apesar de a Índia ser o maior mercado do Facebook e ter alguns dos idiomas mais falados do mundo. Na verdade, a empresa gasta 84% de sua "remessa global/cobertura linguística" para conter a desinformação nos Estados Unidos, que têm menos de 10% de seus usuários. Os 16% restantes são para o resto do mundo.

As empresas de tecnologia precisam reconhecer que a construção de tecnologias de linguagem demanda amplitude e profundidade em recursos, contexto sociopolítico e compromisso com uma experiência digital multilíngue segura e acolhedora.

Nossos colaboradores falam sobre o espectro de necessidades, desafios e oportunidades na criação de tais experiências. Particularmente, a falta de infraestrutura (de acesso à internet a dispositivos eficazes) e tecnologias para todas as línguas torna o uso digital de línguas marginalizadas cansativo, difícil, lento e impraticável. Aqui oferecemos alguns poucos exemplos poderosos.

As tecnologias de linguagem raramente são projetadas para línguas marginalizadas.

Donald conta como é raro para sua comunidade no Malaui ter acesso à internet e dispositivos para comunicação fácil em seus idiomas. Ele entrevistou 20 falantes de chindali, dez dos quais eram estudantes e dez anciãos da comunidade. Dos 20 entrevistados por ele, apenas cinco possuíam smartphones ou telefones básicos inteligentes e sete não possuíam nenhum aparelho. Apenas quatro dos 20 — todos estudantes — tinham laptops. Quanto à internet, apenas os universitários tinham acesso a ela por meio de suas faculdades ou planos de dados pessoais.

Uma vez online, é raro que a maioria das pessoas no mundo consiga acessar teclados em seus próprios idiomas. A maioria das comunidades precisa adaptar um teclado projetado principalmente para idiomas europeus e colar caracteres de seus próprios idiomas no teclado. Isso é difícil o suficiente para o teclado de um computador pessoal, mas impossível para um teclado de telefone pequeno. Também é ainda mais difícil para línguas que são principalmente orais e não têm um sistema de redação padronizado.

Como Ana diz sobre sua língua, "os teclados não têm os símbolos zapotecas corretos para representar os sons e tons de nossas línguas. Os esforços para que as línguas indígenas fossem escritas têm sido discutidos por anos para tentar chegar a um consenso sobre um formato padronizado como o latim, um formato um tanto quanto impositivo e influenciado pelo Ocidente, e até certo ponto aceito e exigido por um setor de falantes". Isso é ainda mais difícil para linguagens como arrernte, que combinam voz e gestos, como Joel e Caddie nos contaram por meio do projeto Indigemoji.

Se você pertence a uma comunidade que já se sente insegura no mundo e o acesso à internet não foi desenhado no seu próprio idioma, é improvável que fique à vontade para produzir e tornar visíveis conteúdos críticos e relevantes para e pela sua comunidade. Paska, em seu artigo sobre conteúdo LGBTQIA+ em indonésio (Bahasa Indonesia), sugere que "muitos indivíduos LGBTQIA+ na Indonésia ainda não estão familiarizados com o aspecto técnico de um site, nem têm conhecimento suficiente sobre como funciona um mecanismo de busca".

Comunidades marginalizadas em regiões privilegiadas do Norte Global também enfrentam esses desafios de não ter conteúdo que afirme e, às vezes, salve vidas em seus próprios idiomas. Jeffrey and Ashley explicam: "Um dos principais desafios [...] se encontra nas limitações técnicas das tecnologias de tradução atuais para línguas indígenas no contexto canadense. Línguas indígenas como hul'qumi'num, sḵwx̱wú7mesh (squamish), lewkungen e neheyawewin (cree) estão sendo tecnicamente reconhecidas de maneira errônea como alemão, estoniano, finlandês, vietnamita e francês pela tecnologia de tradução do Twitter."

De modo geral, Uda confirma: "A criação de conteúdo digital em idiomas locais continua sendo um desafio devido à indisponibilidade de ferramentas e complexidade daquelas disponíveis. A criação de conteúdo em um idioma local exige ferramentas e habilidades especiais. Esse desafio contribui para a falta de conteúdo progressista nas línguas locais."

As tecnologias de linguagem são principalmente projetadas em uma abordagem de cima para baixo, priorizando o lucro sobre o patrimônio e a segurança. Claudia descreve a abordagem pela maioria das empresas de tecnologia de forma muito clara quando diz: "[o] fornecimento de tecnologia e mídia é feito de cima para baixo por grandes empresas, com pouco ou nenhum envolvimento das comunidades de falantes. Nesse caso, pode-se identificar também uma abordagem paternalista: como muito pouco está disponível, tudo o que é fornecido deve ser bom e bem-vindo por definição. Muitas vezes as empresas oferecem soluções prontas, sem levar em conta as reais necessidades, desejos e expectativas dos falantes de línguas minoritárias. É como se a suposição fosse que essas pessoas deveriam ser gratas por qualquer produto ou oportunidade que lhes seja dado, não importando se é realmente interessante ou relevante para a vida delas. Exceções notáveis a esse comportamento são exemplificadas por van Esch et al. (2019), que enfatiza repetidamente a necessidade de estreita colaboração com falantes de línguas para planejar o desenvolvimento de aplicativos de Processamento de Linguagem Natural."

Essa abordagem raramente compreende os contextos em que vivem as comunidades linguísticas marginalizadas e as diferenças de design necessárias para colocar seus idiomas online de uma forma rica e com nuances. Quando Emna entrevistou Gamil, do Sudão, ele falou em árabe sudanês para dizer: "O Sudão é um país com muitas tribos e grande variedade de tradições e costumes. Então o Norte fala o dialeto árabe (o mesmo que estou usando), e isso por causa do colonialismo, é claro. Já no Oriente e no Ocidente, as tribos falam uma língua local diferente que somente elas conseguem falar e entender. É muito raro as pessoas do Norte encontrarem alguém que as entenda, a não ser que tenha vivido lá e interagido com elas. A origem da nossa língua é o idioma cushítico. As civilizações cushítica e núbia são nossa referência. Quando a Grande Barragem afundou, perdemos a identidade cushítica e não encontramos um dicionário para decifrar o idioma e traduzi-lo para outros. A língua núbia, porém, é conhecida e traduzida. O idioma núbio existe até mesmo em telefones celulares Huawei como uma das opções de idiomas nas configurações móveis. As línguas do Oriente e do Ocidente não são escritas (ou talvez sejam, não sei, preciso verificar). No Norte, falamos árabe. Somos africanos e falantes de árabe, não árabes."

As entrevistas de Emna a levaram a dizer: "A web precisa de todos os usuários: os que escrevem e os que não. No entanto, essa mudança não é estritamente responsabilidade dos usuários; as empresas que projetam e desenvolvem software também devem se responsabilizar pelo projeto futuro da web. Parece que todo mundo está falando hoje em ser inclusivo e lutar contra o racismo, a discriminação e outras formas de colonialismo. Entretanto, para fazer uma mudança na web, as empresas devem discutir como elas perpetuam a exclusão. Web designers, engenheiros de tecnologias da web e proprietários de empresas de tecnologia também devem contribuir para tornar seu software acessível a todos os usuários."

Uma forma de analisar criticamente essas diferentes formas de exclusão é reconhecer que a própria base da tecnologia digital — o código em si — é escrito principalmente em inglês. Há muito poucas linguagens de programação baseadas em outras línguas, o que significa que os próprios tecnólogos precisa saber inglês bem o suficiente para codificar nele. A linguagem de programação Qalb, baseada em sintaxe e caligrafia árabes, é um dos poucos exemplos tentando quebrar essa tendência. Em geral, porém, a indústria de tecnologia precisa entender como o privilégio de idioma já está codificado em cada tecnologia e na maioria dos tecnólogos.

Como Jeffrey e Ashley afirmam, "nosso estudo deixa claro que a promoção de línguas indígenas na internet deve estar enraizada em uma análise crítica da própria tecnologia, bem como nos processos sociais que envolvem sua utilização".

Ações

- Reconhecer que a tecnologia e o conteúdo multilíngues, projetados e criados pelos próprios falantes de línguas, é um direito humano fundamental que precisa ser priorizado por empresas de tecnologia e organizações de padronização e apoiado por instituições globais como a UNESCO.
- Construir um modelo de governança global liderado pela comunidade de infraestruturas linguísticas que funcione em parcerias de confiança e respeito com empresas de tecnologia e outras instituições.
- Centralizar o consentimento e a ética da comunidade na criação de dados e tecnologia linguísticos, garantindo que as comunidades linguísticas tenham poder e segurança sobre o que e como compartilham, especialmente aqueles ainda mais marginalizados por diferentes sistemas interseccionados de opressão e discriminação.

Projetos e imaginações de quem?

Pelos números e histórias que compartilhamos, sabemos que infraestruturas linguísticas significativas e eficazes só são possíveis quando centramos as necessidades, projetos e imaginações da própria comunidade linguística. As línguas marginalizadas florescerão e se expandirão apenas quando a maioria minorizada do mundo fizer parte da construção da tecnologia.

Nossos colaboradores têm uma série de sugestões para garantir isso, como, por exemplo, que empresas de tecnologia contratem tecnólogos e outros especialistas de comunidades linguísticas marginalizadas, bem como oferecer recursos a essas comunidades de forma responsável para esse trabalho. Eles recomendam que haja estruturas e processos plurais baseados em contextos de linguagem, em vez de uma abordagem única e estereotipada. Como Claudia coloca, "os falantes de línguas minoritárias não precisam de soluções prontas: suas necessidades e exigências precisas devem ser ouvidas e acomodadas em produtos feitos sob

medida para atendê-las. Os contextos sociolinguísticos das várias línguas minoritárias podem diferir muito, assim como as soluções oferecidas."

Nossos colaboradores também descrevem as responsabilidades e abordagens de suas próprias comunidades para preservar e expandir seus idiomas, usando as tecnologias disponíveis. Ishan nos diz: "Como pessoa queer e com deficiência, acredito que devemos fazer o melhor uso dos nossos recursos socioculturais e econômicos disponíveis para visibilizar nossas experiências, aspirações e demandas e fazer com que sejam ouvidas. Nós, usuários da internet, devemos assumir a responsabilidade de torná-la inclusiva e acessível."

Ana explica como as plataformas com melhores infraestruturas orais e visuais são mais usadas por suas comunidades do que as que dependem de texto. "Hoje, no caso das línguas zapotecas, a oralidade é mais usada que a linguagem escrita no ciberespaço. Plataformas sociais como YouTube, Facebook, WhatsApp e Instagram se apresentam como recursos disponíveis e amigáveis para línguas orais. Essas plataformas permitem que os usuários façam upload de conteúdos visuais que enriquecem a mensagem da forma como desejam que seja entregue, sem cair no rigor da redação. São exatamente essas plataformas que têm mais usuários globalmente e estão sendo usadas por comunidades indígenas. No caso das comunidades da Serra Zapoteca, a maioria dos usuários se conecta à internet através do Facebook, pelo qual transmitem festivais tradicionais, danças, música, narrações de eventos importantes e anúncios para migrantes e comunidades locais. Antes da Covid-19, o Facebook já desempenhava um papel importante no luto pelas famílias de migrantes, pois funerais e rituais eram transmitidos por meio da plataforma. Essa mesma plataforma é usada por algumas comunidades zapotecas para retransmitir a programação de rádio, fazendo uma ponte importante entre um meio de comunicação analógico amplamente utilizado em áreas rurais, como o rádio, com um espaço universal e onipresente, como a internet."

Para aqueles que podem usar formas de linguagem escrita, as hashtags se tornaram uma maneira estimulante para as comunidades se conectarem digitalmente para aprender e divulgar suas próprias línguas e inspirar outras comunidades também. Como Jeffrey e Ashley descrevem: "Em um contexto social em que as línguas indígenas em todo o Canadá foram alvo de eliminação sob as políticas coloniais de assimilação, as redes de hashtags do Twitter reuniram um contexto único e significativo para os povos indígenas compartilharem conhecimento sobre as línguas indígenas... Por exemplo, uma rede de hashtags de alunos de idioma gwichin inspirou os alunos de anishnaabemowin (ojibway) a criar sua própria rede de hashtags. Da mesma forma, uma rede da língua neheyawewin (cree) no Twitter inspirou os alunos de hul'qumi'num a começar sua própria Palavra do Dia no Twitter."

O uso amplo e criativo dessas plataformas proprietárias por comunidades marginalizadas é um motivo considerável para as empresas de tecnologia trabalharem com elas, em vez de contra elas.

Ao mesmo tempo, Claudia adverte que os ativistas de línguas de comunidades marginalizadas precisam estar mais bem informados e coordenados para não perderem energia e recursos no processo. “Embora louváveis, essas iniciativas [dos ativistas] tendem a sofrer de falta de coordenação, pouco planejamento e ainda menos capacidade de ser descobertas. Isso leva a um problema muito sério para as comunidades onde os recursos não são ilimitados: a duplicação de esforços. Em ambos os casos, o principal problema reside no conhecimento limitado do que já está disponível e do que é necessário. A fim de descolonizar a tecnologia da linguagem para as línguas minoritárias, é importante obter uma imagem mais clara de até que ponto as línguas minoritárias são usadas na mídia digital, com que frequência e para que fins. É igualmente importante saber sobre os obstáculos que os falantes de línguas minoritárias enfrentam quando (se) tentam usar essas línguas: enfrentam dificuldades técnicas? Estão bloqueados por algum tipo de paranóia auto-induzida? Como escrever em uma língua minoritária é uma espécie de exposição ao mundo exterior, as pessoas evitam fazê-lo por medo de serem ridicularizadas ou estigmatizadas? Da mesma forma, pouco se sabe sobre o desejo dos falantes de línguas minoritárias em relação às oportunidades digitais: o que eles querem ou esperam que seja disponibilizado?”

Este relatório foi uma tentativa de compreender esses desafios e os caminhos a seguir. Fazer as mesmas coisas repetidamente para idiomas diferentes não funciona. Por exemplo, construir um aplicativo em inglês e presumir que ele funcionará de maneira razoavelmente parecida em indonésio (Bahasa Indonesia) é muito problemático. Melhorar a internet é mudar a dinâmica de poder entre as pessoas, não apenas corrigir problemas técnicos. E o multilinguismo na internet é um conjunto de questões sociotécnicas e políticas complexas. Precisamos colocar em primeiro lugar as necessidades das comunidades linguísticas, não as tecnologias da internet. Trabalhar assim tornará a tecnologia linguística mais eficaz e útil.

E o que é ainda mais fundamental: sabemos que são os projetos e a imaginação do que chamamos de “maioria minorizada” do mundo que vão mudar nossa infraestrutura linguística para melhor. “O [Indigemoji] chega em um momento crítico de rápida absorção de tecnologia e nova conectividade na Austrália Central. O projeto convida a população local a imaginar o que poderia fazer com essas novas plataformas. Como fazer com que não sejam mais uma força colonizadora? E como podemos incorporar nossas línguas e cultura neles, para torná-los nossos?”

Ações

- Centralizar a tecnologia da linguagem nos contextos, necessidades, projetos e imaginações de comunidades linguísticas locais, mas globalmente conectadas, em vez de focar na tecnologia que presume que uma linguagem serve para todos.
- Ser criativos ao usar a mais ampla gama de tecnologias da internet para explorar a mais ampla gama de linguagens corporificadas (oral, visual, gestual, texto etc.) para que

diferentes formas de conhecimento possam ser expressas e compartilhadas de maneira fácil e acessível.
- Aprender com nossas nações indígenas a projetar tecnologias de linguagem honrando memórias coletivas e comunitárias ao mesmo tempo que se planeja o que está por vir. Vamos voltar para os nossos futuros.

Por fim, o que você pode fazer?

Se alguém não fala inglês tão bem quanto você, não significa que a pessoa seja ignorante. Isso significa que ela fala melhor em uma das outras 7.000 línguas do mundo.

Todos nós, com muitas habilidades e experiências diferentes, precisamos trabalhar juntos para criar e expandir uma internet verdadeiramente multilíngue. Também precisamos garantir que as informações e conhecimentos que compartilhamos nessas muitas línguas não causem males, mas, em vez disso, levem ao bem coletivo do nosso mundo. Precisamos do que chamamos de 'solidariedade em ação'.

Se você estiver na área de tecnologia:

- Reconheça como as políticas da sua empresa estão contribuindo (ou não) para a multilinguagem da internet e para o aprofundamento do conhecimento humano compartilhado.
- Dê destaque ao trabalho de internacionalização e localização de idiomas (marginalizados) em suas estratégias em vez de tratar essa questão como supérflua. Faça isso em parceria com a comunidade, em vez de usar uma abordagem de cima para baixo e descontextualizada.
- Aceite as críticas bem embasadas de grandes modelos de linguagem e tecnologias de linguagem automatizadas, e os extensos danos que eles podem causar sem uma supervisão humana cuidadosa.
- Crie processos humanos de atenção/cuidado à contextualização de contextualização, curadoria e moderação de todo o trabalho linguístico, usando conjuntos de dados menores administrados pela comunidade.
- Trabalhe respeitosamente com as comunidades, especialmente aquelas mais marginalizadas e que provavelmente serão as mais prejudicadas por qualquer falta de cuidado e atenção.
- Recorra a quem pode lhe dar tempo e experiência nessas comunidades linguísticas.

Se você trabalhar com tecnologia em uma organização de padrões de tecnologia:

- Reconheça como os padrões de linguagem precisam estar muito bem contextualizados.

- Construa relacionamentos e processos melhores com comunidades linguísticas marginalizadas para que mais padrões possam ser feitos em parceria com comunidades ou, ainda melhor, conduzidos por elas.
- Convide intencionalmente mais membros de comunidades linguísticas marginalizadas para fazer parte da governança e dê a eles os recursos necessários para participarem plenamente.

Se você atuar no governo:

- Reconheça que o conteúdo nos idiomas de seus cidadãos precisa ser acessível a todos, não apenas para uma minoria privilegiada.
- Apoie a expansão do conteúdo nos idiomas de e para aqueles que são marginalizados ou discriminados em suas regiões.
- Apoie a preservação e digitalização de línguas marginalizadas em sua região, não apenas das línguas dominantes.
- Se você estiver na área de software livre e aberto e de conhecimento aberto:
- Reconheça que a tecnologia e o conhecimento gratuitos e de código aberto também têm seus próprios desequilíbrios e limitações de poder, embora se destinem ao bem coletivo.
- Respeite os limites que as comunidades marginalizadas estabelecem em seu compartilhamento de conhecimentos, especialmente devido às formas como seus conhecimentos foram historicamente explorados e mercantilizados no passado.
- Trabalhe com comunidades linguísticas marginalizadas para criar as tecnologias e conhecimentos de que precisam, não aqueles dos quais você acha que precisam.

Se você estiver em uma instituição GLAM (galerias, bibliotecas, arquivos, museus e memória):

- Reconheça que a linguagem está no centro dos conhecimentos e culturas de que você está fazendo a curadoria, preservando e exibindo.
- Trabalhe com comunidades linguísticas marginalizadas para garantir que suas histórias e línguas sejam afirmadas, reconhecidas e amplificadas como desejam, com formas de identificar sua proveniência (ou a propriedade e localização dos materiais). Isso inclui o direito das comunidades marginalizadas de optar por não ter certos conhecimentos e materiais compartilhados publicamente. Isso é essencial porque muitas instituições GLAM, especialmente no Norte Global, estão enraizadas em histórias complexas de colonização e capitalismo.
- Certifique-se de que os materiais linguísticos armazenados em suas coleções sejam de acesso livre e fácil às comunidades marginalizadas e seus aliados, para que possamos construir uma infraestrutura linguística coletiva juntos.

Se você estiver na área de educação:

- Reconheça como nossa educação privilegia fontes baseadas em texto e certos idiomas.
- Expanda suas formas de ensino e aprendizagem para incluir vários idiomas e as diferentes formas de linguagem e conhecimento que eles incorporam.
- Leia, ouça e cite trabalhos de tradução sempre que possível e incentive outros a fazer o mesmo.

Se você estiver no mercado editorial:

- Reconheça o quanto a maioria das publicações no mundo atual privilegia as línguas coloniais europeias.
- Expanda o número de idiomas em que você publica e faça digitalizações em todos esses idiomas.
- Publique mais livros e materiais multilíngues.
- Experimente formas multimodais de publicação, de modo que diferentes formas de linguagem (oral, visual e textual) possam ser compartilhadas simultaneamente com mais facilidade.
- Honre e reconheça seus tradutores.

Se você estiver na área de filantropia:

- Reconheça que a linguagem está no cerne da expertise, experiência e conhecimento humanos de todos os tipos, não importa o que você financia.
- Ofereça tradução simultânea ou consecutiva em vários idiomas em todos os diferentes eventos e convenções globais e regionais que você apoia.
- Apoie a produção, preservação e digitalização de materiais nos idiomas das comunidades que você apoia e certifique-se de que seus próprios materiais estejam nos idiomas dessas comunidades.

Se você estiver em uma comunidade linguística marginalizada:

- Reconheça que você não está sozinho.
- Saiba que é um direito da sua comunidade decidir quais conhecimentos vocês gostariam de compartilhar com o mundo e de que forma.
- Trabalhe com os mais velhos, acadêmicos e gerações mais jovens em sua comunidade, bem como amigos de outras comunidades, para coletar e compartilhar esse conhecimento.
- Se você quiser se conectar com outras pessoas que fazem um trabalho semelhante, entre em contato conosco!

Se você ama idiomas e está pensando no que pode fazer:

- Reconheça que idiomas estão no cerne do que nós somos e do que fazemos, e são fundamentais para diferentes conhecimentos e culturas – incluindo a sua!
- Tenha conversas com sua família, amigos, e comunidades para destacar como e porque o inglês e alguns outros idiomas dominam o acesso e o conteúdo da internet, e como mudar isso juntos.
- Procure ativamente, leia, ouça e compartilhe contribuições de comunidades linguísticas marginalizadas (inclusive este relatório!)
- Se você quiser receber atualizações sobre a nossa iniciativa, nos siga nas redes sociais!

Gratidão

Expressamos todo o nosso amor, respeito e solidariedade às muitas comunidades marginalizadas ao redor do mundo (indígenas e outras) que mantêm a língua no centro de suas identidades e modos de ser. Seus esforços para preservar, revitalizar e expandir essas línguas e formas de expressão de maneiras significativas nos inspiram a imaginar uma internet mais multilíngue e plural, na qual possamos ser o mais completos e ricos que conseguimos. Também agradecemos profundamente a todos os acadêmicos e tecnólogos da comunidade e instituições que amam os idiomas como nós, e que trabalham arduamente todos os dias para tornar a internet tão multilíngue quanto nosso mundo físico.

Aos nossos muitos colaboradores, tradutores e comunidades pelo mundo (especialmente aqueles que se juntaram à nossa conversa sobre Descolonizar as Línguas da Internet em 2019): obrigado por tudo o que vocês fazem e são no mundo, e por terem sido pacientes conosco enquanto tentávamos sobreviver nos últimos dois anos! Agradecemos especialmente à nossa ilustradora por suas visualizações criativas dos ensaios, e ao nosso animador, que nos presenteou com a animação dessas ilustrações.

Muita gratidão a todos os nossos amigos e comunidade que revisaram nosso trabalho de diferentes perspectivas e em diferentes idiomas. Todos os erros são nossos, mas seu apoio e solidariedade ajudaram a tornar este trabalho em andamento muito melhor. E, finalmente, agradecemos uns aos outros e às nossas famílias de sangue e famílias escolhidas: não poderíamos ter sobrevivido aos últimos anos (especialmente 2019–2021) sem nos abraçarmos, mesmo que apenas virtualmente. Amor e confiança são a melhor linguagem de todas.

Definições

Existem muitas maneiras diferentes de definir os diferentes aspectos da linguagem e das histórias que discutimos. Nem todas essas definições concordam umas com as outras! Usamos certos termos e frases de maneiras específicas ao longo deste relatório. Estas são nossas definições dessas palavras e frases-chave.

- **Línguas dominantes**: línguas que são as faladas pela maioria da população em uma determinada área, ou que dominam por meio de formas específicas de poder e validação histórica, por meio de forças legais, políticas ou culturais. Por exemplo, o hindi é uma língua dominante no sul da Ásia, em comparação com muitas outras línguas, especialmente considerando que o próprio hindi é uma família de línguas ou o que alguns chamam de "dialetos". Igualmente, o mandarim é uma língua dominante na China, por política governamental, em comparação com outras formas de chinês e também com outras línguas indígenas da região. Algumas línguas dominantes também são línguas "oficiais" ou "nacionais" em uma região ou país.

- **Línguas coloniais europeias**: Línguas da Europa Ocidental que se espalharam pela África, Ásia, Américas, Caribe e Ilhas do Pacífico através dos processos de colonização por empresas e governos da Europa Ocidental, a partir do século XVI. Elas incluem inglês, espanhol, francês, português, holandês e alemão. É importante notar que essas línguas também foram línguas "colonizadoras" para os povos indígenas da América do Norte, não apenas da América Latina (América Central e do Sul).

- **Sul Global e Norte Global:** O termo "Sul Global" refere-se às regiões da África, Ásia, América Latina e Caribe e Ilhas do Pacífico que foram colonizadas por países da Europa Ocidental. Não é um termo geográfico. Em vez disso, pretende refletir as condições socioeconômicas e políticas históricas e contínuas que caracterizam esses países e regiões, e os distingue dos países privilegiados da Europa e da América do Norte, ou "Norte Global". Foi criado e ampliado por acadêmicos e ativistas do Sul Global para ir além do que eles consideravam ser a natureza pejorativa e indesejável de termos como nações "menos desenvolvidas" e "em desenvolvimento" e "Terceiro Mundo". Como a colonização levou ao genocídio ou dizimação de muitas nações indígenas no Norte Global, e como alguns indivíduos e comunidades no Sul Global se beneficiaram e participaram da colonização de seus próprios povos, às vezes dizemos que existe um Sul Global no Norte Global e um Norte Global no Sul Global. Essas estruturas e processos afetam o status dos idiomas também nessas regiões (ver o termo "maioria minorizada").

- **Línguas indígenas**: As línguas faladas pelas nações indígenas de uma determinada região ou lugar são línguas indígenas. Os povos indígenas são vistos como os "povos originários" ou primeiros habitantes de lugares em todo o mundo que foram posteriormente colonizados e ocupados por um grupo cultural diferente. A maioria das mais de 7.000 línguas no mundo é falada por comunidades indígenas.

- **Línguas e dialetos**: Consideramos qualquer sistema estruturado de expressão entre humanos, seja por voz, som, sinal, gesto ou escrita, como língua. Alguns linguistas definem um "dialeto" para descrever o que soa como diferentes variedades da mesma língua, mas que pode ser "mutuamente inteligível" — compreendido por todos os falantes dessas diferentes variedades que podem, então, falar uns com os outros. No entanto, como a diferença entre a definição mais comum de "língua" e "dialeto" é uma escolha política em vez de linguística, baseada em processos históricos de poder e privilégio, raramente usamos o termo 'dialeto' neste relatório. Preferimos usar o termo "família de línguas", mostrando que há muitas línguas que podem ter histórias semelhantes, mas também têm características diferentes, como as famílias de línguas árabe, chinês ou hindi.

- **Idiomas locais**: Neste relatório, definimos os idiomas locais como os idiomas falados pelo maior número de pessoas em um país ou região.

- **Idiomas marginalizados**: Neste relatório, idiomas marginalizados são aqueles que não são proeminentes na internet em relação a suporte ou conteúdo de idioma, ou seja, em termos de informações e conhecimentos nesse idioma. Essas línguas são marginalizadas por estruturas e processos históricos e contínuos de poder e privilégio, incluindo colonização e capitalismo, e não pela população ou pelo número de falantes. Algumas línguas marginalizadas já estão ameaçadas de extinção no mundo (como muitas línguas indígenas). Mas algumas línguas marginalizadas são faladas por uma maioria significativa de povos em sua região ou no mundo, e ainda assim sub-representadas online (como punjabi e tamil, ou hausa e zulu, como apenas alguns exemplos de muitas línguas dominantes na Ásia e na África).

- **Línguas minoritárias e majoritárias**: as línguas minoritárias são aquelas faladas por uma minoria (em termos de números) da população, em qualquer território ou região descrita, enquanto a língua majoritária é falada pela maioria (em número) dessa população.

- **Maioria minorizada do mundo**: Estruturas históricas e contínuas de poder e privilégio resultam na discriminação e opressão de muitas comunidades e povos diferentes, em todo o mundo. Essas formas de poder e privilégio costumam se interligar e se cruzar, de modo que algumas comunidades são prejudicadas ou oprimidas de várias maneiras: por exemplo, por gênero, raça, sexualidade, classe, casta, religião, região, deficiência e, claro, pelo idioma. Seja online ou no mundo físico, essas comunidades constituem a maioria do mundo em população ou número, mas muitas vezes não estão em posições de poder e, portanto, são tratadas como uma minoria. Em outras palavras, eles são a "maioria minorizada" do mundo.

[Mais sobre como citar e usar este relatório ↪]

[Mais sobre nossos recursos e inspirações ↪]

Este trabalho está licenciado sob Atribuição-NãoComercial 4.0 Internacional (CC BY-NC 4.0), com exceção de algumas partes do conteúdo. Mais informação.