Explorando de forma confiável a presença das línguas na Internet

A Internet é um recurso valioso para os linguistas porque oferece um espaço amplo e de fácil acesso em que se pode observar a evolução das línguas ao longo do tempo. Daniel Pimienta, diretor do Observatório da Diversidade Linguística e Cultural na Internet (OBDILCI), desenvolveu um método para medir a presença das línguas na Web, que foi consideravelmente melhorado nos últimos anos. Este método permitiu o desenvolvimento de uma base de dados abrangente que poderia apoiar a investigação linguística, as políticas públicas relacionadas com as línguas e as estratégias de comércio eletrônico.

O advento da Internet e a sua utilização generalizada abriram novos e interessantes caminhos para o estudo das línguas. Medir a presença online de diferentes línguas oferece informações valiosas sobre seu uso e desenvolvimento futuro.

Estimativas fiáveis do uso das línguas na Internet poderiam, em última análise, orientar o desenvolvimento de políticas públicas destinadas a influenciar a sua presença no ciberespaço. O Observatório da Diversidade Linguística e Cultural na Internet, instituto de investigação fundado em 1996, é especializado no desenvolvimento de métodos eficazes para medir a presença e evolução das línguas na Internet.

Embora os algoritmos de reconhecimento de línguas – ferramentas informáticas capazes de identificar línguas escritas – possam parecer meios adequados para determinar a prevalência de línguas online, a Web tornou-se agora tão vasta que é impraticável aplicar estas ferramentas a todos os conteúdos online. Alguns estudos utilizaram estes algoritmos para analisar subconjuntos da Web, mas esta abordagem experimental provou ser ineficaz, levando a resultados tendenciosos e muitas vezes não confiáveis.

Até recentemente, a fonte mais popular de estatísticas de uso de línguas on-line (W3Techs) contou com algoritmos para analisar os sites classificados como mais visitados. Embora estas estatísticas forneçam informações interessantes, podem não refletir com precisão a presença de línguas online devido à falta de consideração da natureza muitas vezes multilíngue dos websites, o que leva a distorções significativas.

Em 2017, o Observatório da Diversidade Linguística e Cultural na Internet desenhou uma nova abordagem que poderia ajudar a acompanhar melhor o progresso e a prevalência das línguas online. Utilizando esta abordagem, Daniel Pimienta e os seus colegas conseguiram identificar indicadores significativos relativamente à presença de 343 línguas na Internet.

INDICADORES DE PRESENÇA LINGUÍSTICA NA INTERNET

Já em 1998, os investigadores do observatório introduziram uma abordagem para estudar a presença de sete línguas online, que se baseava em dados recolhidos por motores de busca como AltaVista e Google. No entanto, em 2007, começaram a perceber que os resultados dos motores de busca estavam a tornar-se pouco fiáveis e, por isso, procuraram métodos alternativos.

A nova abordagem introduzida em 2017 abordou os fortes enviesamentos associados a esforços anteriores nesta área de investigação. Inicialmente, os investigadores aplicaram esta abordagem a 138 línguas com mais de 5 milhões de falantes nativos cada uma, mas recentemente conseguiram estendê-la a 343 línguas, com mais de um milhão de falantes nativos cada.

Utilizando o método proposto, Pimienta e colegas compilaram um conjunto de indicadores da presença online destas 343 línguas. Estes indicadores foram divididos em três grandes categorias, nomeadamente indicadores intermédios, macro e avançados.

Os indicadores intermédios, todos expressos em percentagens, incluem os utilizadores da Internet (ou seja, falantes de uma determinada língua ligados à Internet), a utilização de serviços ou aplicações específicas da Internet e o tráfego comunicado em sites. Incluem também uma aproximação do nível de suporte digital das línguas e os chamados índices, que são avaliações de países baseadas em parâmetros da sociedade da informação, transformadas por ponderação em avaliações linguísticas.

O segundo conjunto de indicadores, denominado indicadores macro ou resultados de modelo, inclui falantes conectados (ou seja, a percentagem de falantes de primeira e segunda língua em todo o mundo que estão ligados à Internet), percentagem de conteúdo da web em cada língua, produtividade de conteúdo (a proporção de conteúdo para usuários da Internet) e presença virtual (a proporção de conteúdo da web para falantes).

Finalmente, os indicadores mais avançados identificados pelos investigadores incluem a cibergeografia das localizações linguísticas ou, por outras palavras, a distribuição das línguas na web em grupos geográficos (ou seja, línguas europeias, asiáticas, árabes, americanas e africanas) e um indicador de ciberglobalização. O valor deste último indicador, calculado a partir de alguns outros indicadores, expressa essencialmente as "vantagens estratégicas" de uma determinada língua na Internet.

UM NOVO MÉTODO PARA ESTUDAR A PRESENÇA DAS LÍNGUAS NA INTERNET

A nova abordagem do Observatório consiste em estimar indiretamente a quantidade relativa de conteúdo web por língua. Para isso, leva em consideração fatores cruciais que muitas vezes são ignorados na descrição da presença de uma língua na Internet, mas que devem ser levados em consideração para evitar erros ou vieses.

Em primeiro lugar, a equipe considera a provável existência de uma "lei econômica" relacionada com a Internet, que liga a oferta (ou seja, o conteúdo da Web disponível numa língua) à procura (ou seja, o número de falantes desta língua ligados à rede). Estudos anteriores demonstraram que quanto mais falantes de uma determinada língua estão ligados à Internet, mais páginas Web existem nessa língua.

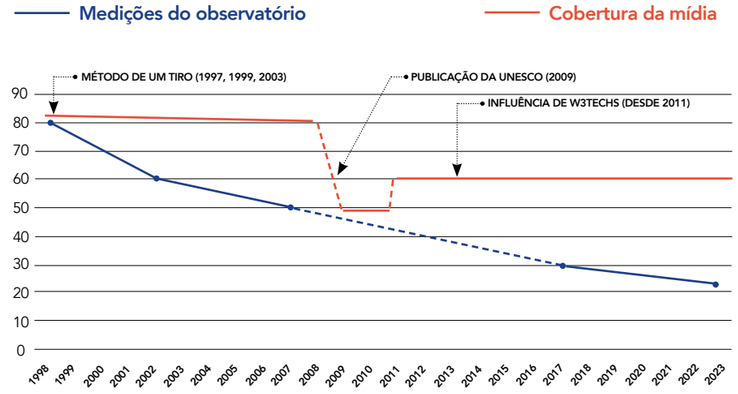
Além disso, pesquisas anteriores sugerem que os utilizadores da Internet preferem muitas vezes comunicar na sua língua materna quando o conteúdo que procuram está disponível nessa língua, mas estão dispostos a utilizar a(s) sua(s) segunda(s) língua(s) caso contrário. Em alguns casos, os utilizadores da Internet também podem criar conteúdos na sua segunda língua por razões econômicas e utilizar serviços de tradução para o fazer.

A presença online de uma língua também está ligada à quantidade de tráfego da Internet para diferentes sites, ao número de assinaturas de redes sociais e ao progresso dos diferentes países em termos de serviços online para os cidadãos. Os indicadores de presença na Internet criados pelos pesquisadores levam coletivamente em consideração todos esses fatores, criando uma imagem mais detalhada de quanto e como existem diferentes línguas na Web.

UM BANCO DE DADOS FACILMENTE ACESSÍVEL

Usando esta abordagem, Pimienta e os seus colegas decidiram calcular indicadores de presença online para línguas faladas nativamente por mais de um milhão de pessoas em todo o mundo. Isto permitiu-lhes construir uma base de dados abrangente que resume a presença destas línguas online, que o observatório planeja atualizar todos os anos.

POSIÇÃO DO OBSERVATÓRIO SOBRE PORCENTAGENS DE PÁGINAS EM INGLÊS AO LONGO DO TEMPO
mito vs. realidade



A percentagem de páginas web em língua inglesa caiu consideravelmente ao longo do tempo, de acordo com o medições do observatório.

Entre 2011 e 2022, a W3Techs apresentou estatísticas que sugerem que o inglês tem uma presença online estável com mais de 50% do conteúdo, mas a análise do observatório sugere que não é esse o caso.

A base de dados do observatório está livremente disponível para linguistas explorando a presença de diferentes línguas online.

Os valores obtidos são muito interessantes, porque muitas vezes não correspondem aos valores produzidos por outros esforços de medição linguística. Por exemplo, entre 2011 e 2022, o site W3Techs apresentou estatísticas que sugerem que o inglês representa de forma estável a maior parte do conteúdo online (mais de 50%), enquanto a análise do Observatório indica que este não é o caso (o inglês abarca cerca de 20% do conteúdo na Internet hoje).

Pimienta e seus colegas descobriram que os línguas com mais conteúdo da web são o inglês e o chinês. Estima-se que cada um desses línguas represente entre 16 e 26% de todo o conteúdo da web, seguido pelo espanhol (entre 7 e 9%), o árabe, o hindi, o russo, o francês e o português (3 a 4%), o japonês, o alemão e o malaio (2 a 3%), seguido pelo bengali (1,5 a 2,3%) e o turco, o vietnamita, o italiano, o coreano e o persa (0,8 a 1,2%). No geral, as conclusões da equipe sugerem que o conteúdo da web é cada vez mais multilíngue, com a prevalência do inglês a diminuir gradualmente de 80% em 1998 para 20% hoje. Permanece claro que isto não significa que a quantidade de conteúdo em inglês tenha diminuído ao longo do tempo, mas sim que o conteúdo online em muitas outras línguas aumentou, reduzindo assim a percentagem de conteúdo em inglês.

Coletivamente, as 215 línguas restantes representam de 18 a 26% do conteúdo da Web. Ao examinar vários indicadores da presença linguística online, a base de dados dos investigadores oferece vários outros conhecimentos que poderiam informar a investigação linguística, as políticas públicas e as estratégias de comércio eletrônico.

Por exemplo, Pimienta e seus colegas descobriram que os falantes de norueguês são os mais conectados à Internet, com impressionantes 98,8% de falantes conectados, seguidos pelos falantes de dinamarquês (98,7%), alemão suíço (94,1%), catalão (94,5%) e finlandês (92,8%). Além disso, as pessoas de língua japonesa parecem ser as mais presentes virtualmente, gerando, proporcionalmente ao seu número, mais conteúdo do que os falantes de outras línguas.

Indicadores relacionados à cibergeografia das línguas também forneceram informações valiosas. Por exemplo, mostraram que embora o número de línguas faladas por mais de um milhão de pessoas em África seja maior do que em cada uma das outras regiões geográficas, os falantes de línguas africanas estão menos ligados à Internet; no entanto, a tendência recente é finalmente no sentido do crescimento das línguas africanas.

Finalmente, o indicador da ciberglobalização enfatiza a vantagem estratégica das línguas inglesa e francesa. Em outras palavras, falar um desses dois línguas abrirá mais oportunidades futuras para os usuários da Internet.

As pessoas que falam japonês parecem ser os mais presentes virtualmente, gerando proporcionalmente mais conteúdo do que os falantes ligados de outras línguas.

EXPLORANDO DE FORMA CONFIÁVEL A PRESENÇA DAS LÍNGUAS NA INTERNET

Um trabalho recente do Observatório da Diversidade Linguística e Cultural na Internet permitiu produzir novos dados fiáveis que expressam a presença de diferentes línguas online. À medida que o uso da Internet continua a crescer globalmente, estes dados podem esclarecer como as línguas utilizadas pelos utilizadores da Internet estão a progredir na web.

Muitas estatísticas existentes sobre a presença de línguas na Internet provaram ser enganosas, não conseguindo representar adequadamente as línguas presentes na Web. No entanto, algumas destas estatísticas têm sido amplamente utilizadas em estudos linguísticos e divulgadas nos meios de comunicação, levando a mais confusão e desinformação.

A base de dados online do Observatório é acessível ao público e pode ser utilizada por linguistas que explorem a presença de diferentes línguas na Internet. Ao mesmo tempo, poderá também servir de referência para decisores de políticas públicas, estrategas de comércio eletrónico e fornecedores de serviços de Internet, ajudando-os a compreender melhor, a partir da sua própria perspetiva, a utilização de diferentes línguas na Internet.

Por trás da pesquisa

Dr Daniel Pimienta

E: pimienta@funredes.org **W:** funredes.org/pimienta **W:** obdilci.org
www.linkedin.com/in/pimienta

Detalhe

Biografia
Daniel Pimienta estudou matemática aplicada e ciências da computação em Nice (França). Foi arquiteto de sistemas de telecomunicações na IBM, e posteriormente criou FUNREDES, uma ONG em pesquisa-ação pioneira em TICpD. Foi Secretária Geral da Rede Global para a Diversidade Linguística (MAAYA). Atualmente é responsável por OBDILCI, produzindo indicadores sobre a presença das línguas na Internet.

Financiamento
Organização Internacional da Francofonia (OIF), Instituto Guimarães Rosa (IGR) do Ministério das Relações Exteriores (MRE) do Brasil, Instituto Internacional da Língua Portuguesa (IILP) da Comunidade dos Países de Língua Portuguesa (CPLP), Delegação Geral para a Língua Francesa e as Línguas da França (DGLFLF).

Collaborators
Cátedra UNESCO de Políticas Linguísticas para o Multilinguismo (UCLPM)

Referências

- Pimienta, D, (2022) Recurso: Indicadores sobre a presença de línguas na Internet. Anais de SIGUL2022 @ LREC2022, 89-91. aclanthology.org/2022.sigul-1.11 - versão em português
- Pimienta, D, Blanco, A, Müller de Oliveira, G, (2023) O método por trás da produção inédita de indicadores da presença de línguas na Internet. Frontiers in Research Metrics and Analytics, 8. doi.org/10.3389/frma.2023.1149347 - versão em português
- Pimienta, D, Prado, D, Blanco, A, (2009) Twelve years of measuring linguistic diversity on the Internet: balance and perspectives. UNESCO publications for the World Summit on the Information Society. CI-2009/WS/1. unesdoc.unesco.org/ark:/48223/pf0000187016

Objetivos de pesquisa

O Daniel Pimienta estuda a presença de línguas na Internet.

Resposta pessoal

O que o inspirou a realizar esta pesquisa?

"Em 1995, durante a Cimeira da Francofonia em Cotonou (Benin), o presidente francês Chirac apresentou a Internet como um espaço 100% de língua inglesa. Naquela época, eu era uma espécie de evangelista da Internet na Fundação de Redes e Desenvolvimento (FUNREDES) e achava que essa afirmação era falsa e não baseada em dados comprovados. Em resposta, decidi lançar um esforço de investigação destinado a medir a presença das línguas na Internet. Este projeto amadureceu em 1998 com a ajuda da União Latina e produziu uma série de resultados até 2007. Percorreu um caminho difícil depois de 2007 e até 2017, quando conseguiu relançar-se numa abordagem nova e promissora que permitiu ampliar o número de línguas processadas."

Como a base de dados de indicadores linguísticos que você compilou poderia informar políticas públicas destinadas a fortalecer a presença de línguas online?

"Hoje, as estratégias para fortalecer as línguas devem centrar-se principalmente no ciberespaço devido ao seu poderoso impacto global. Qualquer que seja a política que desenvolva, em qualquer área, necessita de indicadores significativos, fiáveis e duradouros para definir a sua estratégia e ser capaz de avaliar frequentemente os impactos das suas ações, a fim de as adaptar em conformidade. Os indicadores destas políticas linguísticas no ciberespaço têm sido caracterizados há demasiado tempo por dados tendenciosos, sobrestimando largamente a realidade da presença do inglês e subestimando a parte notável do multilinguismo na Internet, desmotivando assim os esforços de produção de conteúdo local. Esta desinformação deve acabar e estamos felizes por poder contribuir desta forma e fortalecer todas as ações a favor do multilinguismo na rede."