STATE OF THE INTERNET'S LANGUAGES

Summary Report · 2022

STATE OF THE INTERNET'S LANGUAGES REPORT

Summary Report

Why this report? Who are we?

Dictionaries and grammars tell us that language is a structured way of expressing information, mostly between humans. But language is so much more than that: it is the foundational heritage we offer each other, most often given to us by our ancestors, and if we are lucky, passed on to those who come after us. When we think, when we speak, when we hear, when we imagine... we are using language for ourselves and each other. It is at the core of who we are, and how we are, in the world. It is what helps us tell stories and share what we know about ourselves and each other. What language do you speak? Do you dream in that language? Do you think in a different language than the one you speak at work? Is the music you love in a language you don't always understand?

Every language is a system of being, doing, and communicating in the world. And most importantly, of knowing and imagining. Every one of our languages is a system of knowledge in itself: our language is a fundamental way in which we make sense of our world and explain it to others. **Our languages can be oral (spoken and signed), written, or transmitted through sounds** made by a whistle or a drum! In all these forms, **language is a proxy for knowledge**. In other words, language is the most obvious way in which we express what we think, believe, and know.

Now consider the languages you speak, think, dream or write in. Of those languages, how many are you able to fully share and communicate in digital spaces? What is your experience of using language or languages online? Does the hardware you use have the characters of your language? Do you need to modify keyboards to use it with your language? When you look for information using a search engine, do the results come back in the language you want? Have you had to learn a different language from your own to access and contribute to the internet? If your answer to any or many of these questions is "no", then you are amongst the privileged few people in the world who are able to use the internet in your own language easily. And your language might just be... English.

The internet and its different digital spaces provide one of the most critical infrastructures of knowledge, communication and action today. Yet with over 7000 languages in the world (including spoken and signed languages), **how many of these languages can we fully experience online? What would a truly multilingual internet look, feel and sound like?**

This report is one way in which we try to answer this question. We are a collaborative of three organizations: Whose Knowledge?, Oxford Internet Institute, and The Centre for Internet and Society (India). We came together to offer different insights, experiences, and analyses on the state of languages on the internet, and in partnership with others who care about these issues, we hope to create a more multilingual internet, digital technologies and practices.

This report intends to do three things:

- **Map the current status of languages on the internet**: we are trying to understand which languages are currently represented on the internet and how. We do this through quantitative data (looking at numbers across different digital platforms, tools and spaces), as well as qualitative data (learning from people's own stories and experiences with languages online).
- **Raise awareness of the challenges and opportunities in making the internet more multilingual**: Creating and managing the technologies, content, and communities for the world's languages has significant challenges, and also exciting possibilities and opportunities. This report will lay out some of these challenges and possibilities.
- **Advance an agenda for action:** With these insights and awareness, we offer some ways in which we - and many others who work on these issues across the world — would like to plan and act to ensure a more multilingual internet.

What is this report and what is it not?

This report is a work-in-progress or a work-in-process!

So many different individuals, communities, and institutions have been working on different aspects of languages for a very long time, and different aspects of languages online, more recently. We are inspired by them, but this report is not meant to be a complete, exhaustive survey of every one of them and their work. We also don't know everyone working on languages and the internet, although we have tried to include most of those we do know and are inspired by, in some way or the other, by including them in our Resources and Gratitude sections.

We are limited by the data we could gather, and we've discussed some of those constraints in our Numbers section. We welcome comments and suggestions for improving and updating the information we have offered here, and we would love to hear from those who are already working on these issues and would like to be included in future updates of this report.

We've done our best to write this report in an as accessible a style as possible. We want multiple generations and communities of people to join us in our work, and do not want jargon or "academic" language to be a barrier to reading and reflecting. We also want it to be translated into as many languages as possible (translators: reach out to us!), and while we wrote some of this report first in English, we do not want English to be a barrier for either reflection or action.

We hope this report will serve as a "baseline" for further research, discussion and action on these issues, while building on the many efforts of the past.

Who are we, and why did we come together to do this report?

Three organizations came together to do the research for this report: The Centre for Internet and Society, Oxford Internet Institute, and Whose Knowledge?. All of us are interested in the implications of the internet and digital technologies from different perspectives of research, policy and advocacy.

For the past few years, we've been working in our own ways to understand knowledge inequalities and injustice on the internet: who contributes to the content online and how? We soon realized that there was very little data on knowledge in different languages on the internet. Then we wanted to find out more: what is the extent to which the world's languages are on the internet right now? How multilingual is the internet? Our exploration was limited only to a few areas in which we could find useful public and open information, but we hope it will be another contribution for all of us who are striving for a multilingual internet.

A brief note on Covid-19 and this report: We started working on this report in 2019 before Covid-19, but most of the work of analysis, interviews, and writing happened during the global pandemic that has changed our lives individually and collectively. Everyone who contributed to this report has been affected in some way, and it took much longer than we anticipated for us to share it with the world. But Covid-19 also helped us remember how interconnected we are, how essential it is for us to be able to convey complex ideas in different languages, and how critical it is to have resilient and accessible (digital) infrastructure that is truly multilingual.

How to read this report?

This report is what we're calling "digital first", i.e. the best way to read, listen, and learn from it is through this website, because the report has a few different layers and levels. Our report brings together Numbers and Stories. Learning about the state of languages online from a statistical perspective gives us an overview of the issues and helps us understand the different contexts people are experiencing. But people's experiences of language on the internet from around the world, in different contexts, help us learn more deeply about how easy or difficult it is for people to use the internet in their languages. With both stories and numbers, we can begin to address some of the underlying contexts, challenges and opportunities.

This is why there are three main layers to this report:

- The summary of what's in the State of the Languages Report and how we created it (what you are reading right now!)
- Numbers that analyze a few critical language issues in some of the digital platforms, apps and devices that we use every day. Our friends at Oxford Internet Institute led this work, and you will find their fascinating data visualizations and analyses here. Please note that this analysis is limited to the data we could access, from open and publicly available datasets and materials. Other methodological constraints are discussed in more detail in these essays, but most importantly: it's hard to find a single and consistent way to identify languages, and similarly, difficult to estimate how many people use specific languages, especially because languages and their use are dynamic, changing over time.
- Stories that give us a deeper understanding of how different people and communities around the world experience the internet in their own languages, and in many cases, how difficult it is right now to find information they need in their own languages. We invited these stories in written and spoken forms, so you will find textual essays, as well as audio and video interviews. Our friends at The Centre for Internet and Society led this work, pulling together this rich tapestry of language experiences from around the world. We have contributions about Indigenous languages like Chindali, Cree, Ojibway, Mapuzugun, Zapotec, and Arrernte from Africa, the Americas, and Australia; minority languages like Breton, Basque, Sardinian, and Karelian in Europe; as well as regionally and globally dominant languages like Bengali, Indonesian (Bahasa Indonesia) and Sinhala in Asia, and different forms of Arabic across North Africa.

Most importantly, our contributors have written or spoken in their own languages as well as in English, and our summary is also written and spoken in different languages, so we hope you will enjoy reading or listening in more languages than just one!

We have also done our best to bring these contributions to life in visual form, with imaginative illustrations and animation that carefully bring together different social and technical aspects of language. As with everything else in our report, these were also collaboratively developed by our illustrator in conversation with our contributors.

How multilingual is the internet?

The internet is not yet (and sadly, nowhere near) as multilingual as we are in real life. We try to understand why by looking at both some numbers and experiences of people across the world. Here we give you only a brief summary and analysis of the richness and depth of the work done by our contributors from across the world, so please look at their essays for more detail and inspiration.

We first look at the contexts in which people are using the internet from around the world, in different languages. We look at the ways in which information and knowledge are distributed, or not, across different languages and geographies. Then we look more closely at the major platforms and applications that we use to create content, communicate, and share information online, and how many languages each supports. We look in detail at Google Maps and Wikipedia as two multilingual spaces of content that we all use in our every day, and how they work across different languages.

Along the way, we share stories and experiences from people accessing and contributing to knowledge on the internet in their own languages. As we've learned, most of our contributors find themselves needing to switch from their first language of choice to another language in order to access and contribute to the issues they value.

Language context: geographical and digital knowledge inequities

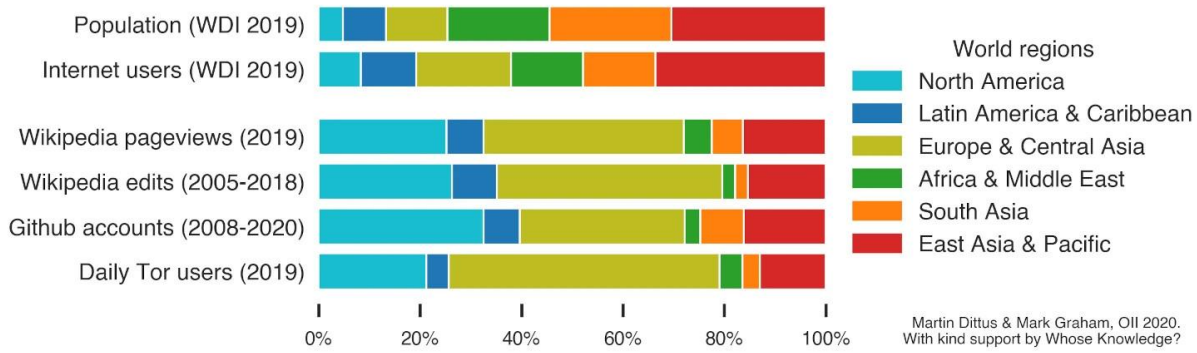
Languages with an oral tradition do not fit in the web we have today.
Ana Alonso

It seems to us that these platforms, in general, perpetuate the colonialist idea that there are languages with greater value and capacity to communicate, which is a detrimental view of minority languages like Mapuzugun.
Kimeltuwe project

We know that over 60 % of the world is now estimated to be connected digitally, most of us via our phones and mobile devices. Of all those who are online, three-fourths of us are from the Global South: from Asia, Africa, Latin America, the Caribbean and the Pacific Islands. But how meaningful and equitable is our access? Are we able to create and produce public online knowledge to the same extent we consume it?

In Martin and Mark's survey of the global population compared to the number of internet users, we find that certain populations can access the internet much more meaningfully than others, including in very well known digital spaces. For instance, even when most of us who are online are from the Global South, we are not able to access the internet as creators and producers of

knowledge, only as consumers. Most Wikipedia edits, the majority of accounts on Github (a repository for code), and the greatest number of Tor users (a secure browser), are from Europe and North America.

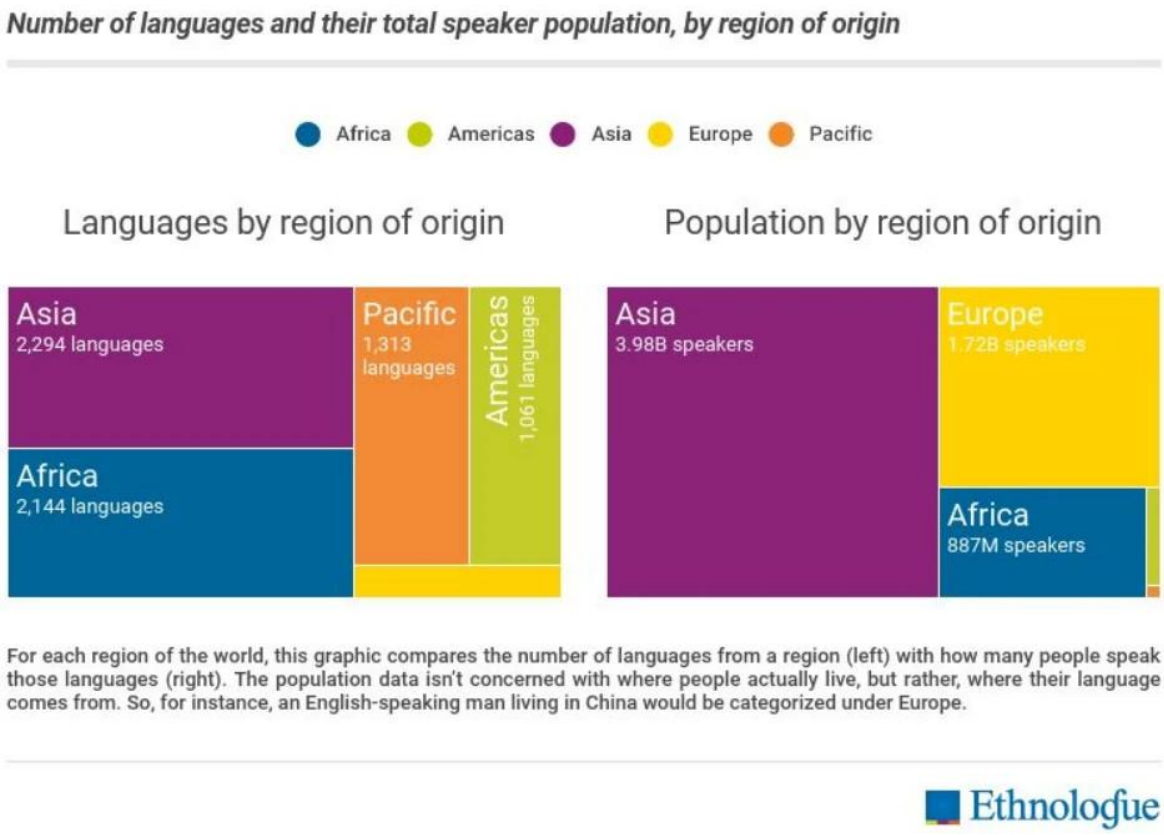


Measures of digital participation by world regions. (Data: World Bank 2019, Wikimedia Foundation 2019, Wikipedia 2018, Github 2020, Tor 2019)

And what does this uneven access mean for languages? Are we all able to access the internet in our own languages? Are we able to create content and information in our own languages?

As other projections show us, over 75 % of those who access the Internet do so in only 10 languages — most of those languages have a European colonial history (English, French, German, Portuguese, Spanish...) or are dominant in specific regions where other languages struggle to remain relevant (Chinese, Arabic, Russian...). In 2020, it was estimated that 25.9 % of all internet users came online in English, while 19.4 % accessed the internet in Chinese. China is the country with the most internet users worldwide, and it's worth remembering that what we call "Chinese" is not a single language, but a family of many different languages, with Mandarin Chinese being the most dominant.

Interestingly, the languages of the internet currently may be mostly European in origin, but Europe has the fewest languages in the world, compared to all the other continents. Of the 7000+ languages in the world, over 4000 come from Asia and Africa (over 2000 languages each), and the Pacific Islands and the Americas have over 1000 languages each. Papua New Guinea and Indonesia are the countries with the most languages, with over 800 in PNG and over 700 in Indonesia.



The number of languages and the total population of speakers by region, across the world. Source: Ethnologue

Many languages from South Asia (Hindi, Bengali, Urdu...) are in the top 10 languages of the world by the number of people who speak it as a first (or native) language, but these are not the languages in which people from South Asia are able to access the Internet. And of course, as we learn from Ishan, whose first language is Bengali, even if you can access digital knowledge in your chosen language, the kind of information you seek may not exist in it; in Ishan's case, content on disability and sexuality rights. The situation in South East Asia — which has some of the highest internet usage in the world and some of the greatest diversity of languages — is similar. Paska finds exactly the same issues with content on sexuality rights in Indonesian (Bahasa Indonesia) as Ishan finds in Bengali.

We also know that of over 7000 languages in the world, only about 4000 of them have written systems or scripts. However, most of these scripts were not developed by the speakers of the languages themselves, but as part of the many colonizing processes across the world. Simply having a script doesn't mean that it is understood or used widely. Most of the languages in the world are transmitted in spoken or signed form, and not through writing. Even in languages that have written forms, publishing has been skewed towards the European colonial languages,

and to a much lesser extent, towards regional dominant languages. In 2010, Google estimated that there were about 130 million books ever published, and a significant proportion of those were in about 480 languages. Most well-established academic journals in science or social science are in English. The most translated book in the world is the Bible (into over 3000 languages), and the most translated document in the world is the United Nations' Universal Declaration of Human Rights (into over 500 languages).

Why does this matter? Because digital language technologies rely on automated processing of published materials in different languages in order to improve their language support and content. So when text publishing across the world is itself biased towards certain languages — and cannot include any non-written languages — it deepens the language inequities we experience. And of course, non-text based languages — those based on sign, sound, gesture, movement — are completely missing from the publishing industry, and therefore often from digital language technologies.

For instance, as Ana tells us, "the web is not designed to respond to users of languages with an oral tradition only." In this domination of written languages on the internet, it is hard to find content from oral and visual language traditions. We cannot easily search for gestures, signs and whistles, for example. In our interview with Joel and Caddie, they tell us about Australia's first set of Indigenous emojis made on Arrernte land in Mparntwe/Alice Springs, and how the physical gesture is often combined with the spoken word to make meaning in Arrernte. Emna says the same about Tunisia, and the different languages spoken by her people: "when it comes to preserving a language, we shouldn't focus only on writing, we need to do it in oral forms, gestures, signs, whistles etc. and that cannot be captured fully in writing."

Digital technologies offer us such possibilities for representing the plurality of language forms that are based around text, sound, gesture, and more. They can also help us preserve and bring back to life the languages that are in danger of dying: over 40 % of all languages. Every month, two Indigenous languages and the knowledges they express die, and are lost to us.

Why is it that these different language contexts are not better represented online?

In her essay, Claudia offers us three dimensions to understand the relationship between languages and technology: availability, usability, and how technology is developed. As we find throughout this report, what Claudia calls "majority languages" (and we find are mostly European colonial languages or regionally dominant languages) have a whole range of media, services, interfaces and apps available, while other languages have far less availability, including in terms of infrastructure like keyboards, machine translation, or speech recognition. Tech companies also spend far more time and resources on the usability of these majority languages, because this is where they see the most profit. Finally, she finds that most language technology is either developed through top-down processes with little collaboration with the

language communities, or the few efforts to work with communities are poorly planned and coordinated.

These concerns and challenges of context also give us ways forward to create a more multilingual internet, and we will come back to these possibilities later.

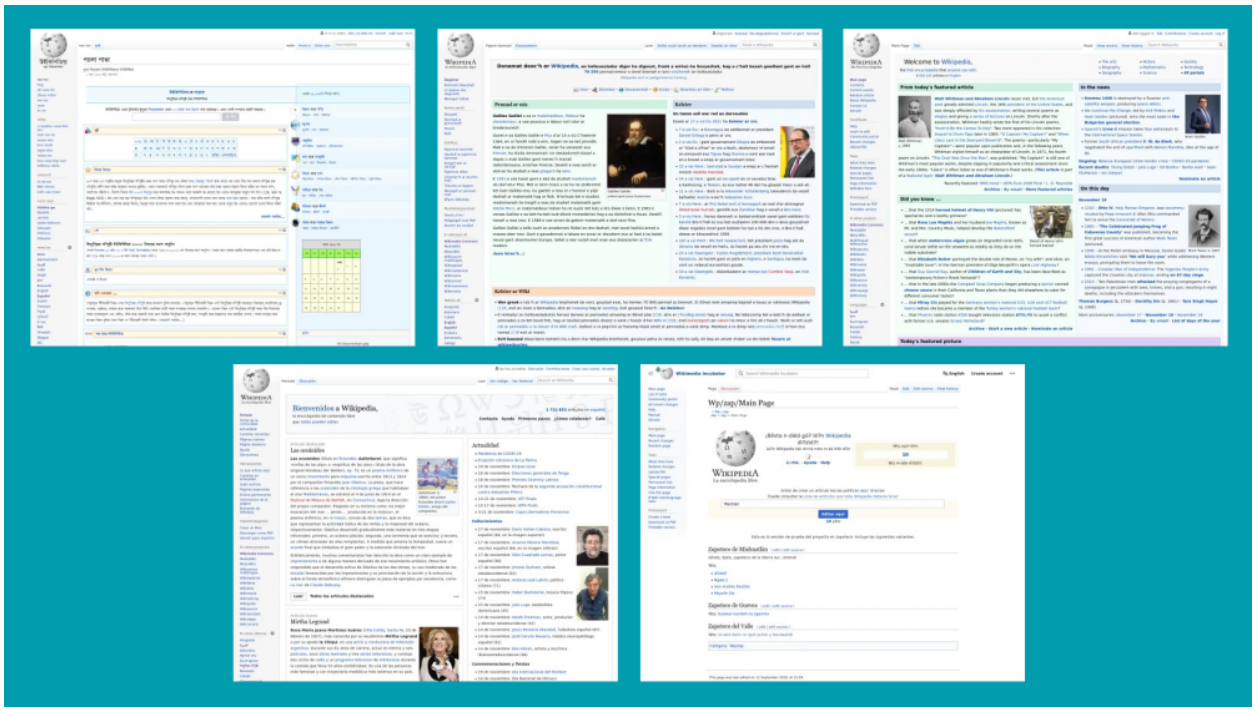
Language support: platforms and messaging apps

When you write the word "good morning", before you finish typing, the phone or computer will have already suggested the word. When I am writing the same word in Chindali ("mwalamusha"), I have to type the whole word and this takes a lot of time; and it will be underlined because the computer or phone cannot recognize the word.
Donald Flywell Malanga

Keyboards with Sinhala and Tamil letters are rare. Our parents printed tiny Sinhala letters, cut them out, and taped them onto the keys beside the original English characters. Though numerous Sinhala fonts have been developed, none work as well as Unicode fonts.
Uda Deshapriya

If popular apps and key software interfaces are not provided in Breton soon, unable to compete with French apps, the language will inevitably appear less appealing to the younger generations.
Claudia Soria

We went deeper to understand how the internet is not yet as multilingual as the world we live in. We looked at what kind of language support — i.e. user interfaces in different languages — major digital platforms and applications provide us for communicating, and creating and sharing content in our languages.

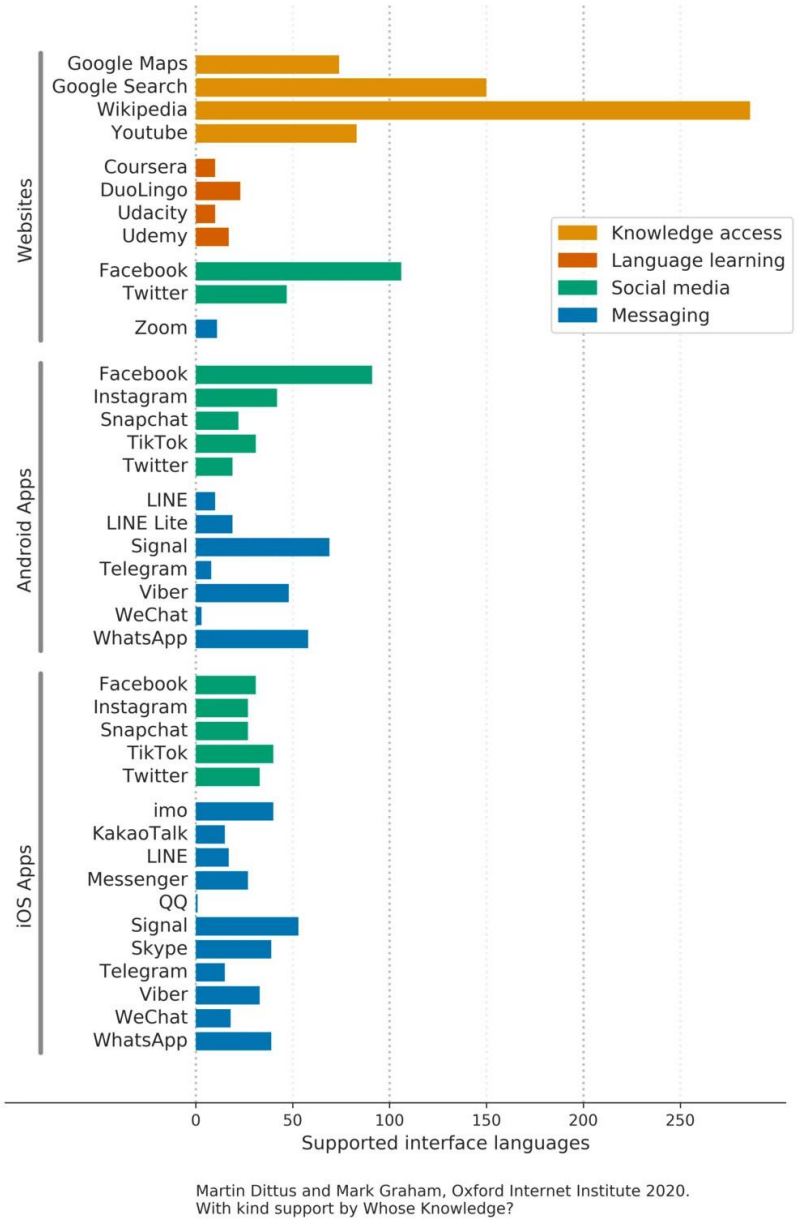


Wikipedia interface in multiple languages.

Martin and Mark analyzed the language support for 11 websites, 12 Android apps, and 16 iOS apps. They chose widely used platforms that specialize in the collection and sharing of knowledge, in particular those that seek to have a global presence and audience across the world. They also had to rely on publicly available data for these platforms and apps, and focussed on written text interfaces and content, rather than voice.

The platforms were grouped into four broad categories (recognizing that they overlap):

- **Knowledge access** (knowledge and information platforms, including search engines): Google Maps, Google Search, Wikipedia, YouTube.
- **Language learning** (self-guided language learning platforms): DuoLingo, and the education platforms Coursera, Udacity, Udemy.
- **Social media** (public-facing social media platforms): Facebook, Instagram, Snapchat, TikTok, Twitter.
- **Messaging** (private and group messaging): imo, KakaoTalk, LINE, LINE Lite, Messenger, QQ, Signal, Skype, Telegram, Viber, WeChat, WhatsApp, Zoom.



The number of supported interface languages for each platform, by platform category.

We found that text-based language support is highly unequally distributed across different digital platforms. Major web platforms like Wikipedia, Google Search, and Facebook offer the most language support currently. Interestingly, Wikipedia (which is a non-profit, and edited by volunteers from around the world) is the most comprehensively translated platform by far. Wikipedia supports over 400 languages with a basic user interface, with Wikipedias in around 300 languages having at least 100 articles. Google Search supports 150 languages, while Facebook supports 70-100 languages. Signal leads the messaging apps with almost 70 languages supported on Android and 50 on iOS. On the other hand, most of the platforms only focus their language support on a small number of widely spoken languages, leaving the majority of languages unsupported. The messenger app QQ, for example, only supports Chinese.

The small number of languages that tend to be supported by most of the platforms surveyed include European languages such as English, Spanish, Portuguese, and French, and certain Asian languages such as Mandarin Chinese, Indonesian, Japanese, and Korean. Major languages like Arabic and Malay are less well-supported, and other languages spoken by tens to hundreds of millions are not represented well in terms of interface support.

What does this lack of language support mean for most people across the world? In 2021, we estimate about 7.9 billion humans on our planet, most of whom live in Asia (nearly 4.7 billion) and Africa (nearly 1.4 billion). Yet most of the world is not served by the internet's languages:

- *People speaking African languages:* the vast majority of African languages are not supported as an interface language by any of the platforms surveyed, and as a result more than 90 % of Africans need to switch to a second language in order to use the platform – for many this may mean a European-colonial language or a more dominant language in their region.
- *People speaking South Asian Languages:* In South Asia, almost half of the platforms surveyed do not offer interface support for any regional language, and even major South Asian languages such as Hindi and Bengali, spoken by hundreds of millions of people, are less widely supported than other languages.
- *People speaking South-East Asian languages:* Support for South-East Asian languages is similarly mixed. While Indonesian, Vietnamese, and Thai tend to be very well-supported by the platforms we surveyed, most other South-East Asian languages are not supported by most of the platforms we surveyed.

Martin and Mark's findings are heightened by the everyday realities of those living in these regions of the world. Donald from Malawi found, for instance, that when he asked speakers of Chindali, an endangered Bantu language, how they communicated via their phones, they all described how time-intensive and laborious it is in Chindali, as most of their phones were embedded with language support in English, French, or Arabic and did not recognize Chindali. These technological challenges are, of course, in addition to economic and social constraints that limit the ability of Chindali speakers to buy a smartphone or data plans. Even for those using the national language of Malawi, Chichewa, the lack of language support makes it difficult: "Why should I buy an expensive phone or waste my time to go and access the internet when the language is English which I cannot understand?"

In fact, the lack of language support for most African languages was sharply noted in 2018, when Twitter first recognized Swahili, a language spoken by over 50-150 million people across East Africa and beyond (as either first or second languages). Until then, Swahili and most other African languages were referred to as Indonesian on the platform. The recognition of Swahili words and translation support was not initiated by the tech company either; it was, in fact, the result of a campaign by Swahili speaking users of Twitter.

The situation is not much better in Latin America for Indigenous languages. In the interview with the Kimeltuwe project working on Mapuzugun, spoken by the Mapuche peoples across present day Chile and Argentina, they pointed out that, "It would be great to be able to post in Mapuzugun on platforms like YouTube or Facebook. Not even in terms of the interface being translated, but in terms of just being able to tag, in the menus that are available, the language as being Mapuzugun. For example, when uploading a video to YouTube or Facebook, you can't add a transcript in Mapuzugun since it does not appear in the list of predetermined languages. So, if you want to upload a transcript in Mapuzugun, you must say that it is in Spanish or English."

Martin and Mark did not analyze language support on specific devices, such as mobile phones, but we know that digital keyboards are one of the few critical spaces in which linguists and technologists have made the most progress. Gboard, Google's smartphone keyboard for the Android operating system, for instance, supports over 900 language varieties and is based on significant work with different language communities and scholars. Yet we can access a smartphone keyboard with these capabilities, only by affording a relatively high-end smartphone.

At the same time, Uda's experience with Sinhala — a language spoken by over 20 million people in Sri Lanka as either a first or second language — shows that it's still difficult to create content in a language whose script is not easily understood by some of the technologists working on language support, especially if the form is very different from the Latin script of western European languages. She says, "The key issue with Unicode Sinhala relates to the order in which different characters need to be inserted to create a letter. This order requires you to follow consonants with diacritics. This thinking follows rules in European Languages based on Latin script. However, in Sinhalese, sometimes diacritics precede the consonant."

Unicode is the technology standard for coding the text that is expressed in a language's writing system or script. Version 13 has 143,859 characters for over 30 writing systems in use today, since the same writing system can be used for more than one language (for example, the Latin script for most Western European languages, the Han script for Japanese, Chinese and Korean languages, and the Devanagiri script for different South Asian languages). It also has characters for historical scripts of languages no longer in use. The Unicode Consortium (a non-profit based in California) also decides on emojis — the symbols that we use everyday through different interfaces.

Martin and Mark's survey of language support and the experiences of other contributors from across the world, all add more detail to this brief description of the limited and uneven technical support for most languages on platforms and apps right now. Do read them!

Language content: accessibility and production

Feminist content is particularly inaccessible in local languages. Women's Development Foundation is a rural women's group who have been working on women's rights issues since 1983. But it was only in 2019 that we started sharing Sinhala language feminist content on socio-political and economic issues online.
Uda Deshapriya

Unfortunately, it was —and still is— so difficult to find educational and positive queer content in Bahasa Indonesia on the internet... If we search for "LGBT" or "homoseksualitas" (homosexuality) on Google —the largest and most popular search engine— we will find so many results containing the word "penyimpangan" (deviation), "dosa" (sin), and "penyakit" (disease).
Paska Darmawan

The information on the intersection of queerness and disability (or even its absence) that is available in Bengali on the internet is to a great extent shaped by and in turn instrumental in shaping homophobia and ableism.
Ishan Chakraborty

We wanted to understand content on the internet by analyzing what version of the world and whose knowledge we are experiencing, once we are online. After all, over 63 % of all websites use English as their primary language of content.

In their essays and interviews, our contributors discuss the different constellations of historical, sociopolitical, economic and technological challenges in meaningfully accessing the internet in their languages. Most significantly, they all address the challenges in finding relevant content on the internet in their languages, and in creating content that is meaningful to them in these languages. In other words, it isn't enough for us to be able to access information and knowledge created for us in other languages by those who may not understand our contexts and experiences, and worse, be hostile to them. We need to be able to produce meaningful knowledge for ourselves and our communities, or, at the very least, be able to support the production and expansion of this content in all our different languages.

This is particularly true for those who have accessibility issues, and for those who experience different interlocking forms of marginalization and exclusion.

As Joel described to us in the interview about the Indigemoji project, it began with a frustrated tweet from his car in which he simply pulled over to the side of the road one day, and started putting Arrernte words against emojis to describe their meaning. For decades after emojis were first introduced to the internet, First Nations or Indigenous peoples had unsuccessfully petitioned for their use to express their oral and visual languages, like Arrernte. As we said earlier, the Unicode Consortium deliberates upon public requests for new emojis, and petitions

such as an emoji for the Australian Aboriginal Flag were rejected. For Joel, Caddie and others, the Indigemoji project became a multi-generational effort to push past these multiple forms of marginalization physically and virtually, and create their own content in the ways that were meaningful to their Indigenous identities and language.

A tweet featuring a list of emojis with Arrernte words next to them. Source: Indigemoji

It's important to remember that Indigenous languages are "minority" languages in our world today, because of histories of mass genocide through colonization in which Indigenous Nations were destroyed or became minority populations after being the primary inhabitants of a particular region or land. And these processes of colonization also affect dominant languages, spoken by millions worldwide.

Ishan is a visually disabled queer academic for whom being online is itself an effort against all odds. Then he struggles to find relevant information in Bengali around disability, around queerness, and even more so, around the intersections of these issues. This leads to what he calls 'marginality within marginality': "on the one hand, the homophobic and ableist attitudes of the society, and on the other hand, the internalized homophobia and/or the ableism of the individuals (queer and/or disabled) — together, these complementary conditions perpetuate the mechanism of marginalization. The societal location of a queer-disabled individual may be described as 'marginality within marginality.'"

In other words, critical processes of access and information even in a dominant language like Bengali — spoken by around 300 million around the world — is missing from the internet.

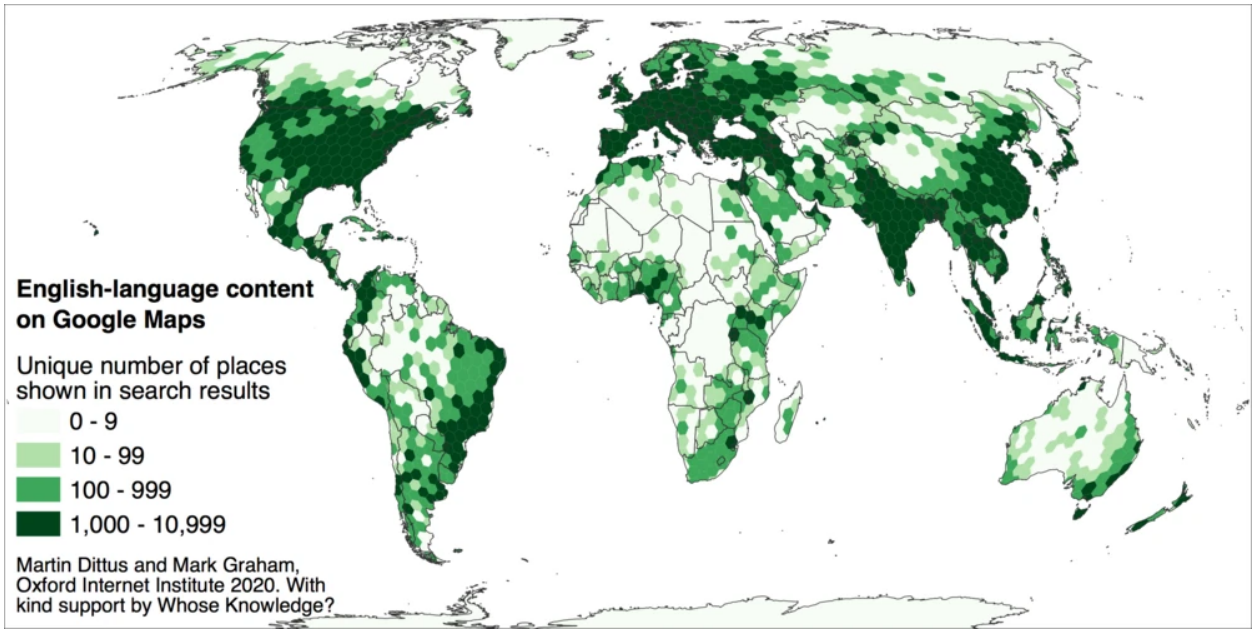
Martin and Mark decided to go deeper into analyzing the extent and kind of content in different languages, on two different kinds of information and knowledge platforms: Google Maps and Wikipedia

Google Maps

Can we access Google Maps in all our many different languages? Does the language we use change the version of the world we see through Google Maps?

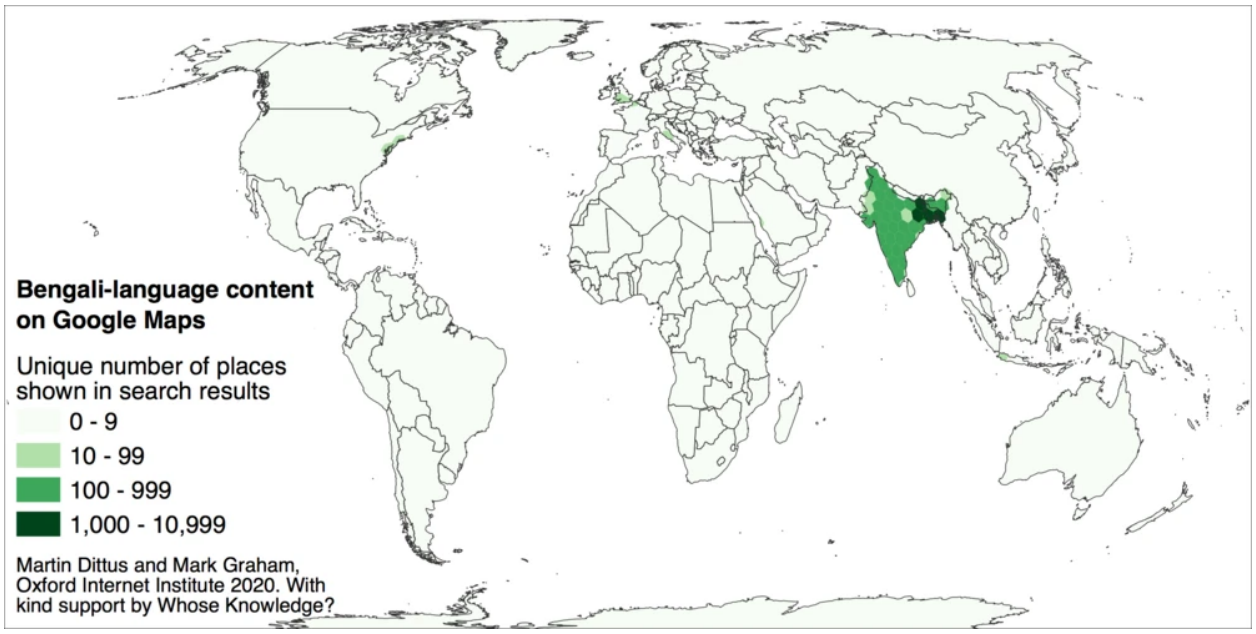
To answer these questions, Martin and Mark collected data about Google Maps' global content coverage in the 10 most widely spoken languages: English, Mandarin Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Portuguese, and Indonesian (Bahasa Indonesia). They collected tens of millions of individual search results in these languages, and across these identified and mapped around three million unique places (venues and other locations).

Unsurprisingly, the maps have the most content when we access Google Maps in English. Google's English-language map covers the world, although it is much more dense (i.e. it has more information) in the Global North, with a focus on Europe and North America. It also covers South Asia and parts of South-East Asia, as well as large parts of Latin America relatively well. In comparison, though, many parts of Africa are comparatively sparse in content.



The information density of Google Maps for English speakers. Darker shading indicates where search results include a greater number of places.

Compared to the relatively well-covered English map, we find that Bengali (which is Ishan's first language) lies at the other extreme — its coverage is mostly restricted to South Asia, especially India and Bangladesh, and Google Maps has little to no content for Bengali speakers in most of the rest of the world. In order to discover additional content, and navigate in places beyond India and Bangladesh, Bengali speakers need to switch to a second language such as English. This is also true for Google Maps in Hindi (the third most spoken language in the world, after English and Mandarin Chinese).



The information density of Google Maps for Bengali speakers. Darker shading indicates where search results include a greater number of places.

Much more on Google Maps in different languages in Martin and Mark's detailed essay.

Wikipedia

As Martin and Mark's platform survey showed us, Wikipedia is at the forefront of language support on the internet, with its user interface translated into more languages than any of the commercial platforms we looked at, including Google and Facebook.

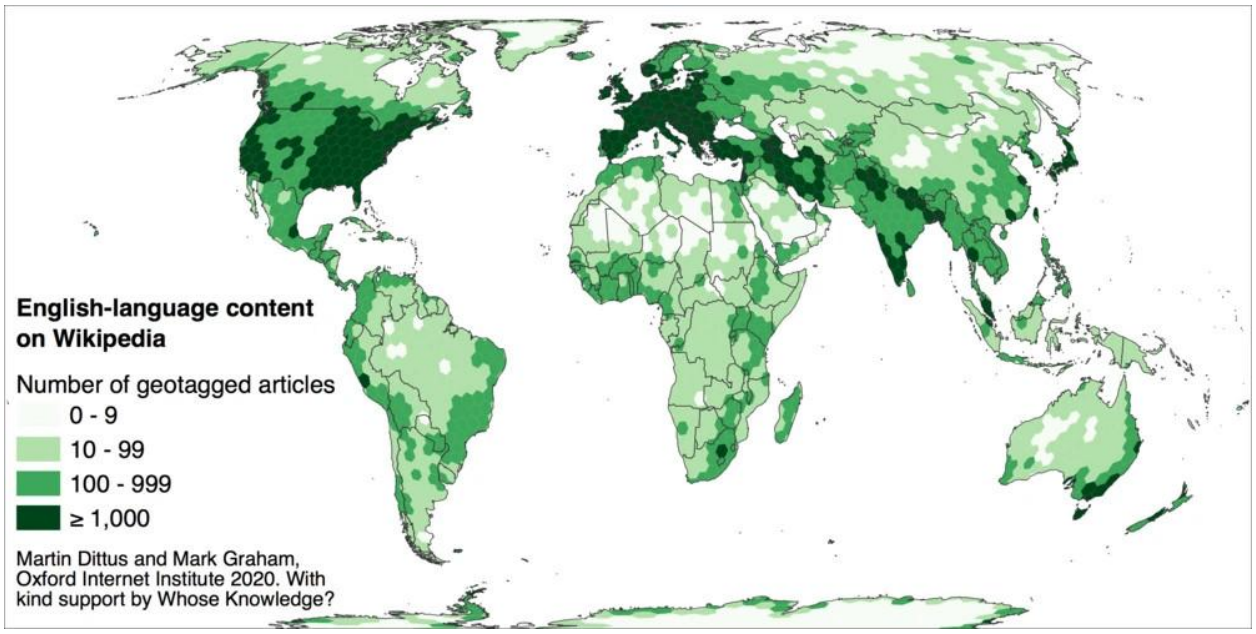
In terms of actual content — information and knowledge in Wikipedia articles — Wikipedia has more than 300 language editions. Yet speakers of these languages don't get access to the same content, or the same amount of information. We wanted to go deeper in asking and answering: how good is Wikipedia's content coverage across its language editions? Are some languages more well-represented than others? Do certain language communities have access to more content than others? We answered some of these questions in detail in Martin and Mark's Wikipedia analysis.

We used 2018 data with geotags (a way to embed geographic references like coordinates within Wikipedia articles), and analyzed the number of articles and growth of content in different languages. We also based our analysis on "local" languages, defining these as languages that are either classified as an official language in Unicode CLDR (the code that supports languages on the internet), or that are in use by at least 30 % of the population in any country.

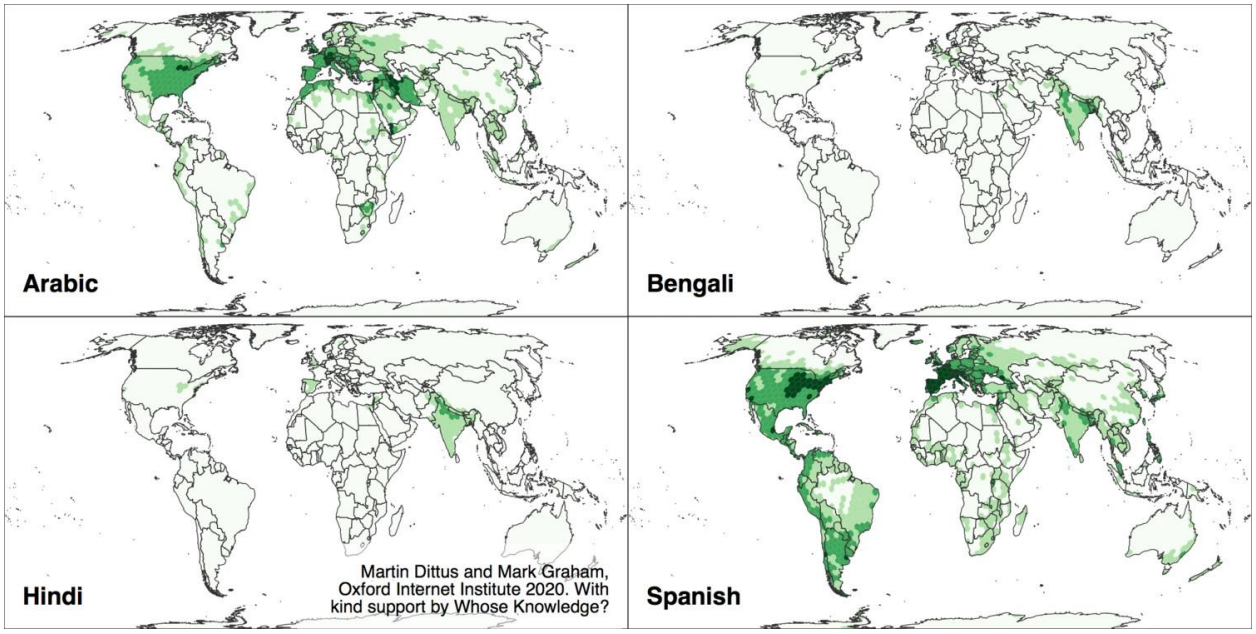
We then identified the most prevalent local language, i.e. spoken by the largest number of people in each country. We found 73 such languages, which are most prevalent in at least one country. English is most widely spoken, and is the most prevalent language in 34 countries. It is followed by Arabic and Spanish (18 countries), French (13 countries), Portuguese (seven countries), German (four countries), and Dutch (three countries). Chinese, Italian, Malay, Romanian, Greek, and Russian are the most prevalent languages in two countries, with the remaining 60 languages being most prevalent in a single country.

To compare this local language distribution with Wikipedia content in each country, we identified the Wikipedia language edition with the largest number of articles about that country. We found a bias towards English-language content. English is the dominant Wikipedia language in 98 countries, followed by French (nine countries), German (eight countries), Spanish (seven countries), Catalan and Russian (four countries), Italian and Serbian (three countries), and Dutch, Greek, Arabic, Serbo-Croatian, Swedish, and Romanian (two countries). The remaining 21 Wikipedia languages are most prevalent in a single country.

While the number of articles in each language Wikipedia is dynamic and keeps growing, it's clear that Wikipedia's language editions vary greatly in size and scale — both in terms of number of articles, and in terms of the size of their editor communities. English Wikipedia is the largest by far, with more than six million articles and almost 40 million registered contributors. The next-largest contributor communities are the Spanish, German and French Wikipedia editions, each with between four and six million contributors, and around two million articles. The remaining language editions are small in comparison: only around 20 language editions have more than one million articles, and only 70 have more than 100,000 articles. Most Wikipedia language editions have only a small fraction of the content in English Wikipedia.



The information density of English Wikipedia in early 2018. Darker shading indicates a greater number of geotagged articles.

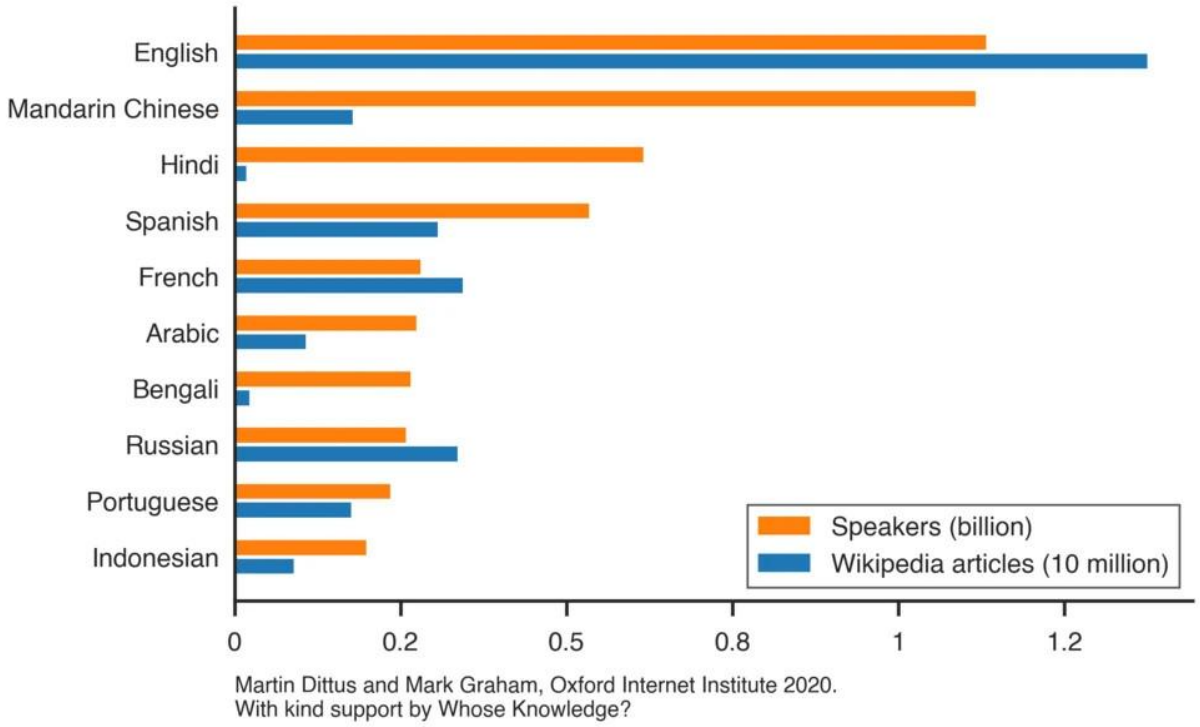


The information density of the Arabic, Bengali, Hindi and Spanish Wikipedias in early 2018. Darker shading indicates a greater number of geotagged articles.

What is fascinating is how content across language Wikipedias echoes the same distribution in the Google Maps we saw earlier.

This is also true when we look at the number of articles in different language Wikipedias in comparison with the number of speakers of those languages (as both first and second

languages). We find that in European languages like English, French, Spanish, Russian, and Portuguese, the number of articles in those Wikipedias are proportional to the number of speakers. But this is not true for other widely spoken languages: Mandarin Chinese, Hindi, Arabic, Bengali, and Indonesian (Bahasa Indonesia) are each spoken by hundreds of millions of people, yet their Wikipedia editions are much smaller, with a smaller number of articles compared to the editions in European languages. There are more articles in the French, Spanish or Portuguese Wikipedias than there are in the Mandarin, Hindi or Arabic versions, although these are in the top five spoken languages of the world, with more speakers than French and Portuguese.



Wikipedia content and number of speakers for the 10 most widely spoken languages in the world. (Population estimate: Ethnologue 2019, which includes second-language speakers.)

Martin and Mark's essay has much more analysis and data visualizations on Wikipedia in different languages, but what is clear from these numbers is what is echoed by the lived experiences of our contributors across the world.

The marginalizations and exclusions of languages that are not primarily European colonial languages are deep across the physical and virtual worlds, including in otherwise dominant and global languages like Arabic. In order to write Wikipedia in one's own language, using references that are based in that language context, we need reliable and extensive published sources, which (as we found earlier) is rare in most languages of the world. As a Wikipedian, Emna talks about how difficult it is to find resources and references in different languages

across Africa: "...the struggle for example for me as a Wikipedian to find references in our own languages, when I say our own languages, it's not only the Tunisian language, our dialect, or the Arabic language but also when we walk across Africa we find a huge gap in terms of resources and references."

Even in Europe, minority language speakers struggle with using or editing Wikipedias in their languages. While Claudia found that many of her respondents in Breton were aware of the existence of a Breton Wikipedia, with "19 % of them even contributing to it by editing existing articles or writing new ones (8 %)," she feels that most minority language speakers will switch to a dominant language because it's easier: "availability does not imply that services, interfaces, apps and Wikipedias are actually used. Some studies reveal that minority language speakers switch easily to their dominant language when using language-based digital technologies, either because the technology is inherently better, or because the range of services available is much wider."

Wikipedia and its constellation of free and open-source (where the code is openly available and collectively built) knowledge projects are one of the most hopeful and helpful spaces for multilingual knowledge online. For instance, its volunteer communities know and understand that there is no single form of English or Arabic or Chinese, but expressing this plurality of language context and content is not always easy. As we see from our analysis and experiences, Wikipedia too suffers from historical and ongoing structures of power and privilege that skew the ways and forms in which we create and share knowledge in different languages, and within families of languages.

What are some ways forward for us as individuals, organizations and communities who wish for a more multilingual internet? In these next and final sections, we draw upon the insights from all the numbers and stories we have shared with you so far, to offer an overview of what we've learned, and the contexts, understandings, and actions that may lead us to a truly multilingual internet.

What have we learned about the multilingual internet?

We have learned a lot about languages, the internet, and languages on the internet, as we've worked on this report. Here is a brief overview of the most important insights of our journey so far.

Learning: Language is a proxy for knowledge and an essential way of being in the world, not only a tool for communication. This is why multilinguality is so important, so we honor and affirm the full richness and textures of our many selves and our different worlds better.

Context: People know their worlds and express themselves in over 7000 languages. Our languages can be oral (spoken and signed), written, or transmitted through sounds.

Yet language support on the main tech platforms and apps are for a fraction of these 7000 languages, with only about 500 of these languages represented online in any form of information or knowledge. Some of the most widely spoken languages in the world do not have much language support or information online. The richest language support, the most extensive information on the internet (including in Google Maps and Wikipedia), and most websites, are in English.

Reflection: The internet is nowhere near as multilingual as we imagine or need it to be.

Analysis: Most people have to use their nearest European colonial language (English, Spanish, Portuguese, French…) or regionally dominant language (Mandarin Chinese, Arabic…) to access the internet. Historical and ongoing structures of power and privilege are intrinsic to the way languages are accessible (or not) online.

How can we do better?: Contexts and Actions for a Multilingual Internet

In the majority of cases, using a minority language requires a good amount of perseverance, will, and resilience, since the user experience in using minority languages is interspersed with flaws and difficulties."
Claudia Soria

The goal of a multilingual internet that is inclusive and representative of Indigenous peoples must consider and reckon with the social legacies and ongoing experience of colonial oppression. A multilingual internet cannot merely aspire to be representative, but, considering colonial history, must also seek to promote environments that enhance the survival and learning of Indigenous languages for and by Indigenous people
Jeffrey Ansloos and Ashley Caranto Morford

Mapuche youth and children grow side by side with technology and the Internet. The internet is a space where these people can approach the Mapuzugun... The stories of our people have to be written, and they need to be written or spoken in Mapuzugun... Our stories are not necessarily those of great heroes, as the colonial and postcolonial States celebrate. Our history is the story of each Mapuche who survived adversity and violence. The woman who had to migrate to the city to get a salary, the women and men who returned from the city, but there was no more room in their "lofs", and they had to go back to the cities with their roots cut off and with no place to return to. These experiences and the memories of each Mapuche person constitute our collective memory as a people.
Kimeltuwe project

In this section of our summary State of the Internet's Languages report, we bring together the different insights from our contributors to this report, as well as from our 2019 Decolonizing the Internet's Languages convening, to understand the different contexts, challenges and opportunities around multilinguality in the world and online. Asking ourselves and each other four key questions may give us the ways forward to imagine and design for a much more multilingual internet.

- **Whose power and resources?**
- **Whose values and knowledges?**
- **Whose technologies and standards?**
- **Whose designs and imaginations?**

Whose power and resources?

Contexts

Both our statistical analysis and people's lived experiences make it clear that the marginalization of languages in physical and virtual worlds is not based only on the number of people who speak the language.

Indigenous languages are spoken by citizens of over 6000 Indigenous Nations globally, who used to be the primary inhabitants of much of the world till colonization and genocide destroyed or reduced their peoples and languages. Dominant languages in some of the most linguistically diverse continents like Asia and Africa, spoken and signed by millions of people, are not represented well, or sometimes at all, online. If we consider the diaspora of people who speak languages like the many different varieties of Arabic, Chinese, Hindi, Bengali, Punjabi, Tamil, Urdu, Indonesian (Bahasa Indonesia), Malay, Swahili, Hausa and so on, across different countries and continents, it's clear that these are marginalized languages online, even as they may dominate over others within their own regions.

These forms of digital marginalization and exclusion are not accidental; they are caused by historical and ongoing structures and processes of power and privilege. They also mean that the resources that go into language infrastructure — from publishing and academia, governments and technology companies — are already skewed towards certain regions (Europe and North America) and certain languages (English and other western European languages). Even within Europe and North America, Indigenous, Black and other marginalized communities find it difficult to preserve their languages across generations.

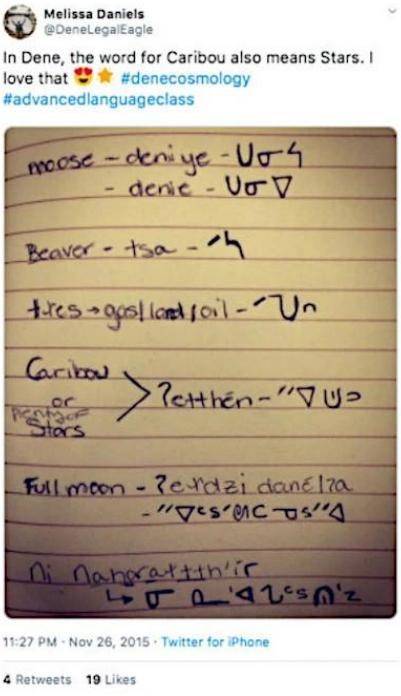
In particular, the forces of colonization and capitalism cause and interlock with other systems of discrimination and oppression like racism, patriarchy, homophobia, ableism, classism and casteism. This means that certain languages — primarily European colonial languages — are the most prominent online, whether or not they are spoken by the most number of people in the world. It also means that when there is information and knowledge in other more marginalized languages, the content in those languages is limited by who has access and power to create them, or to prevent others from producing alternative information. For example, the lack of feminist online content in Sinhala, or the lack of positive content for queer and disabled people in Bengali or Indonesian (Bahasa Indonesia).

Because language is at the heart of who we are, not being able to express ourselves in our own languages and in the fullness of our many selves is a form of violence. But the consequences of these marginalizations are violent in other ways. As Uda says, "The lack of digital content that is feminist, human rights friendly, and respectful, makes online spaces where communication primarily takes place in local languages, hostile to women, queer people, and minorities. There is an obvious lack of digital media that counters the mainstream narrative which remains

contaminated with negative stereotypes. This exacerbates hate speech and sexual and gender based violence online." These forms of violence extend to Indigenous, caste oppressed and minority religion communities, to disabled people, and other marginalized communities.

At the same time, these communities are using the internet to fight back against different forms of transgenerational violence against themselves and their languages. Jeffrey and Ashley analyzed about 3800 tweets with over 35 Indigenous language hashtags and 57 keywords, across 60 federally recognized Indigenous language groups in Canada. They found that, through hashtags on Twitter, Indigenous peoples across Canada and elsewhere in the world are actively connecting, participating, and collaborating in a revival and flourishing of Indigenous languages.

As they put it, "In a social context where Indigenous languages across Canada have been targeted for erasure under colonial policies of assimilation, Twitter hashtag networks have convened a unique and meaningful context for Indigenous people to share knowledge about Indigenous languages. Across the various networks included in our study, there are examples of language reclamation and reconnection for intergenerational survivors of colonial assimilationist policies."



Dene Melissa Daniels in providing consent to cite this tweet wished to recognize the language teacher who provided this teaching, Dene elder and educator Eileen Beaver.

Actions

- Recognize structures and processes of power and privilege in the different institutions and processes that support language infrastructures.
- Ensure reparative resources into marginalized languages and communities, including for language learning and language programming.
- Enable resources towards creating and expanding information and knowledge from and for communities that experience multiple, intersecting forms of oppression and violence, in the languages and forms of their choice.

Whose values and knowledges?

Contexts

The internet's history and technology are based on a worldview from Western epistemologies, or ways of knowing and doing. More specifically, the internet was and continues to be designed and governed primarily by white (and now some brown) privileged men. This means that the values that are most often at the core of the internet's architectures and infrastructures are the values of technological determinism — where technology is seen as being the primary (and beneficial) cause of any societal changes — and individualism, where the primary focus and driver is the individual, not the collective.

In addition, this worldview is rooted in the Age of Enlightenment, the 18th century move in the Global North towards a certain kind of rationality-based science and technology. What we forget is that mathematics and science flourished across the Global South well before the 18th century. The first writing and number systems, for instance, came out of Mesopotamia, present day Iran and Iraq. Even more critically, we forget that the resources for the Enlightenment, this "golden age" of science and technology in the Global North, was the Age of Empire in the Global South, with mass movements of colonization, slavery, genocide and extraction of resources from Asia, Africa, Latin America and the Caribbean and Pacific Islands. The foundations for the extractive nature of modern capitalism are rooted in these past histories of colonization, and continue to be part of tech capital.

Along with material resources, what was destroyed, ignored, or undermined in these processes were other forms of knowing, doing, and being — that is, non-Western epistemologies, like Indigenous knowledges, or the knowledges of those from less privileged parts of the world. The most devastating consequence of this for language, which is a significant proxy for knowledge, as we've said before, was the complete devaluation of non-Western European languages, and the active destruction or benign neglect of oral and non-written forms of language. The bias towards written content in a small set of privileged languages continues to skew us towards a certain kind of written "knowledge" in publishing and academia, which then impacts the

underlying documentation and data for natural language processing, or some of the automated language systems that are part of the internet's infrastructures, like Google Translate.

As Ana describes it, "Despite the fact that just under half of the world's languages are languages that do not have a writing system and maintain a long oral tradition, languages with a widely recognized and widespread alphabetic system dominate the web. The web reinforces a systematic exclusion, where only those languages that are written can be preserved for the future."

Through the loss of language, we're losing more of our futures than we realize. We're losing both the forms of expression in different languages and entire worldviews and knowledges inherent in these languages. At a time in which humanity is on the verge of planetary collapse, Indigenous communities and their knowledges are guarding our biodiversity and preserving life as we know it. Unsurprisingly, we know that the loss of languages is directly linked with the loss of biodiversity and the destruction of ecosystems around the world.

The internet could be an exciting infrastructure for the preservation and expansion of different forms of language and knowledge, because its rich media forms can mimic and represent embodied languages that are spoken, signed, and beyond text. But it's critical that this radical promise of the internet is not based once more on colonial-capitalist and patriarchal values. How do communities of people preserve and revitalize their languages and identities, and share their knowledges on their own terms? For instance, in many Indigenous communities, some knowledge is sacred and not to be shared openly.

One such community-led effort is Papa Reo, a speech recognition technology for te reo Māori (the Māori language) of Aotearoa/New Zealand. The Māori community created and maintains the tech and data for this initiative, and believes that this form of data sovereignty is critical in ensuring that the knowledge shared through the language is used for and by Māori, rather than by profit-seeking companies. Interestingly, while recognizing the value of open-source tech, the Papa Reo team also decided not to add their data to open-source databases, because the Māori community has been denied the resources and privileges of most open-source communities. On the other hand, Mukurtu is an example of an open-source platform that is built with Indigenous communities to manage their own language data.

Actions

- Create, collaborate, and share languages infrastructures for the internet that are for the public good, designed with collective and community values at their heart, and that center feminist and Indigenous concepts of sovereignty and embodiment.
- Continue to challenge and critique languages tech infrastructures that are oppressive in nature: capitalist, proprietary, humanly extractive and environmentally destructive.

- Recognize that free and open-source language technologies also need to pay attention to their own relative privilege, and honor how marginalized communities self-determine and define "open" and what they want to share with the world.

Whose technologies and standards?

Contexts

The tech industry is not fully responsible for the current lack of representation and support for the vast majority of the world's languages. However, technology from the Global North is responsible for continuing to maintain and increase language-based inequities and digital colonialism online.

The big tech companies — who are designing and creating most of the digital platforms, tools, hardwares and softwares we use — can ignore the need to create a truly multilingual internet, because they not consider or see the majority of our 7000 spoken languages as an essential part of the internet's infrastructure. After all, they know that they need to provide language support only when it makes business sense: either for European colonial languages, or languages in what they call "emerging markets". In fact, South and South East Asian dominant languages are starting to get better language support on the big tech platforms as they become the largest customer base for these companies.

At the same time, some of the most widespread language technologies we have on the internet are created and controlled by these companies, because they have the resources and capacity. Wikipedia is a notable exception, because it is open-source and supported by its volunteer communities from around the world. In general, proprietary and profit-driven motivations do not result in the tools and technologies necessary for deep and nuanced content in different marginalized languages, and from marginalized communities. Even worse, the language technologies that are being built currently by these companies, are large scale automated systems that rely on huge amounts of language data from any kind of source — even if that source might be violent and hate-filled speech or writing against marginalized groups.

As Jeffrey and Ashley describe, "Across most Indigenous language survivance and learning networks, racism has been identified as a major social challenge within the Twitter ecology. Specifically, there are various ways in which these networks are actively targeted with incendiary text and multimedia content, sometimes constituting hate speech, by a variety of user types, including actual people, and automated bots. In terms of bot users, these accounts seem to follow similar patterns for dissemination of misinformation, and often distribute nonsensical text reflecting aggregate analytics and content generation."

If a company doesn't care enough about the consequences of their lack of local language expertise and context, then it can lead to immeasurable harm and active violence. In Myanmar,

where Facebook (or Meta) essentially is the internet, activists warned the company about hate speech for years before it set up a Burmese language team. In 2015, Facebook had four Burmese speakers reviewing content, with 7.3 million active users in Myanmar. What did this lack of attention to language and context lead to? It meant that the United Nations found that Facebook was part of the genocide against Rohingya Muslims in Myanmar, and the company is at the heart of a case against the Myanmar government at the International Court of Justice.

Similarly, hate speech in India against Muslims, Dalits and other marginalized communities continues with very little active moderation, despite India being Facebook's largest market, and having some of the most spoken languages in the world. In fact, the company spends 84 % of its "global remit/language coverage" on misinformation in the United States, which has less than 10% of its users. The remaining 16 % is for the rest of the world.

Technology companies need to recognize that the building out of language technologies needs a breadth and depth of resources, sociopolitical context, and a commitment to a safe and welcoming multilingual digital experience.

Our contributors talk about the range of needs, challenges and opportunities in creating such experiences. In particular, the lack of infrastructure (from internet access to effective devices) and technologies for all languages makes the digital use of marginalized languages tiring, difficult, slow and impractical. We offer a few powerful examples.

Language tech is rarely designed for a marginalized language.

Donald talks about how rare it is for his community in Malawi to have internet access and devices for easy communication in their languages. He interviewed 20 Chindali speakers, 10 of whom were students and 10 were elders of the community. Of the 20 he interviewed, only five had smartphones or feature phones, and seven did not have any device. Only four out of the 20 — all students — had laptops. As for internet access, only the college students had access through their colleges or personal data plans.

Once online, it is rare for most people in the world to be able to access keyboards in their own languages. Most communities have to work around a keyboard designed primarily for European languages, and paste characters from their own languages onto the keyboard. This is difficult enough for a personal computer's keyboard, but impossible for a small phone keyboard. It is also even more difficult for languages that are primarily oral and do not have an agreed-upon writing system.

As Ana says of her language, "keyboards do not have the correct Zapotec symbols to represent the sounds and tones of our languages. Efforts to get indigenous languages written have been in years of discussion trying to reach a consensus for a standardized format such as the Latin, a format somewhat imposed and influenced by the West and in some way accepted and required

by a sector of speakers." This is even more difficult for languages like Arrernte, which combine voice and gesture, as Joel and Caddie told us through the Indigemoji project.

If you are from a community that already feels unsafe and insecure in the world, and internet access is not designed in your own language, it is unlikely that you will be able to comfortably produce and make visible relevant and critical content for and with your community. Paska, in their essay on LGBTQIA+ content in Indonesian (Bahasa Indonesia), suggests, "Many LGBTQIA+ individuals in Indonesia are still not familiar with the technical aspect of a website, nor have sufficient knowledge about how a search engine works."

Marginalized communities in privileged regions of the Global North also face these challenges of not having life-affirming and sometimes life-saving content in their own languages. Jeffrey and Ashley explain: "One of the key challenges... is the technical limitations of current translation technologies for Indigenous languages in the Canadian context. Indigenous languages such as Hul'qumi'num, Sḵwx̱wú7mesh (Squamish), Lewkungen, and Neheyawewin (Cree) are being technically misrecognized as German, Estonian, Finnish, Vietnamese, and French by Twitter's translation technology."

Overall, Uda confirms, "creating digital content in local languages remains a challenge due to the unavailability of tools, and the complexity of tools that are available. Creating content in a local language requires special tools and skills. This challenge contributes to the lack of progressive content in local languages."

Language tech is mostly designed in a top-down approach, prioritizing profit over equity and safety. Claudia describes the approach by most technology companies very clearly, when she says: "[the] provision of technology and media is poured top-down by big companies, with little or no involvement of speakers' communities. In this case, a patronizing approach can also be spotted: since very little is available, anything that is provided must be good and welcome by definition. Very often companies offer ready-made solutions without taking into accounts the real needs, desires, and expectations of minority language speakers. It is as if the assumption was that these speakers should be grateful for whatever product or opportunity is given to them, no matter if it is actually interesting or relevant to their lives. Notable exceptions to this behavior are exemplified by van Esch et al. (2019), who repeatedly stress the need for close collaboration with language speakers when planning the development of Natural Language Processing applications."

This approach rarely understands the contexts in which marginalized language communities live, and the differences of design needed to bring their languages online in a rich and nuanced way. When Emna interviewed Gamil from Sudan, he spoke in Sudanese Arabic to say, "Sudan is a country with a lot of tribes and a variety of traditions and customs. So the North speaks Arabic dialect (the same I am using) and that's because of colonialism of course. While for the East and West, the tribes speak a different local language, only them are able to speak and understand it.

For people from the North, it is very rare to find someone who understands it unless s/he lived there and interacted with them. The origin of our language is the Cushitic language. The Cushitic and Nubian civilizations are our reference. When the High Dam sank, we lost the Cushitic identity, we didn't find a dictionary to decrypt the language and translate it into other languages. The Nubian language, however, is known and translated. Nubian language exists even on Huawei mobile phones as one of the languages in the mobile settings. For the East and West languages, they are not written (or may be not, I'm not aware, I need to verify). For the North, we speak Arabic. We are Africans and Arabic speakers, not Arabs."

Emna's interviews prompted her to say, "The web needs all its users, those who write and those who do not. However, change is not something that is strictly the responsibility of users, companies that design and develop software must also take responsibility for the future design of the web. It seems that everyone is talking today about being inclusive and fighting racism, discrimination, and other forms of colonialism. However, to make a change on the web, corporations should discuss how they perpetuate exclusion; web designers, engineers of web technologies, owners of tech companies should also contribute in making their software accessible to all users."

One way in which to critically analyze these different forms of exclusion is to recognize that the very basis of digital technology — the code itself — is mostly in English. There are very few programming languages based on other languages, which means that technologists themselves need to know English well enough to code in it. The programming language Qalb, based on Arabic syntax and calligraphy, is one of the few examples that tries to break this trend. In general, though, the tech industry needs to understand how language privilege is encoded in every technology and most technologists.

As Jeffrey and Ashley put it, "Our study makes clear that the promotion of Indigenous languages on the internet must be rooted within a critical analysis of the internet's technology itself, as well as the social processes that surround its utilization."

Actions

- Recognize that multilingual technology and content, designed and created by language speakers themselves, is a fundamental human right that needs to be prioritized by technology companies and standards organizations, and supported by global institutions like UNESCO.
- Build a model for community-led, global governance of language infrastructures that works in partnerships of trust and respect with technology companies and other institutions.
- Center community consent and ethics in creating language data and tech, ensuring that language communities have power and safety over what and how they share, especially for those marginalized further by different interlocking systems of oppression and discrimination.

Whose designs and imaginations?

From both the numbers and stories we've shared, we know that meaningful and effective language infrastructures are only possible when we center the needs, designs, and imaginations of the language community itself. Marginalized languages will flourish and expand only when the minoritized majority of the world are part of building the tech.

Our contributors suggest a range of ways in which to ensure this, including technology companies hiring technologists and other experts from marginalized language communities, as well as responsibly resourcing communities in this work. They recommend a plurality of structures and processes based on language contexts, rather than a single formulaic approach. As Claudia puts it, "Minority language speakers do not need ready-made solutions: their precise needs and requirements must be listened to and accommodated into products that are tailored to those needs. The sociolinguistic contexts of the various minority languages can differ greatly, and so must the solutions that are provided."

Our contributors also describe their own communities' responsibilities and approaches to preserving and expanding their languages, using the technologies available to them. Ishan says, "As a queer, disabled individual, I believe that we must make the best use of our available socio-cultural-economic resources to make our experiences, aspirations and demands heard and seen. We, the internet users, must take the responsibility of making the internet inclusive and accessible."

Ana explains how platforms that have better oral and visual infrastructures are used more by her communities than those reliant on text. "Today, in the case of Zapotec languages, orality is used more than written language in cyberspace. Social platforms such as YouTube, Facebook, WhatsApp and Instagram appear available and friendly resources for oral languages. These platforms allow users to upload visual content enriching the message, in the way users want it

to be delivered, without falling into the rigor of writing. It is precisely these platforms that have the most users globally and are being used by indigenous communities. In the case of the Sierra Zapotec communities, the majority of users connect to the internet through Facebook, where they broadcast traditional festivals, dances, music, narrations of important events and announcements for migrant and local communities. Before Covid-19, Facebook was already playing an important role in mourning for migrant families, as funerals and rituals were transmitted through the platform. This same platform is used by some Zapotec communities to retransmit radio programming, making an important bridge between a widely used analogue communication medium in rural areas such as the radio, with a universal and ubiquitous space such as the Internet."

For those that can use written forms of language, hashtags have become an exciting way for communities to connect digitally to learn and spread their own languages, and inspire other communities too. As Jeffrey and Ashley describe: "In a social context where Indigenous languages across Canada have been targeted for erasure under colonial policies of assimilation, Twitter hashtag networks have convened a unique and meaningful context for Indigenous people to share knowledge about Indigenous languages... For example, a hashtag network of Gwichin language learners inspired learners of Anishnaabemowin (Ojibway) to create their own hashtag network. And similarly, a Neheyawewin (Cree) language network on Twitter inspired learners of Hul'qumi'num to start their own Twitter-based Word of the Day."

The broad and creative use of these proprietary platforms by marginalized communities is a significant reason for technology companies to work with them, rather than against them.

At the same time, Claudia warns that language activists from marginalized communities need to be better informed and coordinated so that they do not lose energy and resources in the process. "Though commendable, these [activists'] initiatives tend to suffer from lack of coordination, little planning, and even lesser discoverability. This leads to a very serious problem for communities where resources are not unlimited: reduplication of efforts. In both cases, the main problem lies in the limited knowledge of what is already available and of what is needed. In order to decolonize language technology for minority languages, it is important to get a clearer picture of the extent to which minority languages are used over digital media, with what frequency, and for what purposes. Equally important is to know about the obstacles that minority language speakers face when (if) trying to use these languages: do they experience technical difficulties? Are they blocked by some kind of self-induced paranoia? As writing in a minority language is a kind of exposure to the outside world, do people refrain from it for fear of being mocked or stigmatized? Similarly, little is known about the desire of minority languages speakers regarding digital opportunities: what do they want or expect to be made available?"

This report has been one attempt to understand these challenges, and the ways forward. Doing the same things over and over again for different languages will not work. For instance,

building an app in English and assuming it will work reasonably similarly for Indonesian (Bahasa Indonesia) is deeply problematic. Making the internet better is about changing the dynamics of power between people, not just fixing technical issues. And multilinguality on the internet is a set of complex socio-technical and political issues. We need to put the needs of language communities first, not internet technologies — doing it this way round will actually make language tech more effective and useful.

Most critically, we know that it is the designs and imaginations of what we call the "minoritized majority" of the world that will change our language infrastructures for the better. "[Indigemoji] comes at a critical time of rapid technology uptake and new connectivity in Central Australia. It invites local people to imagine what they could do with these new platforms. How are they not just another colonizing force? And how can we embed our languages and culture in them, to make them our own?"

Actions

- Center language tech in the contexts, needs, designs, and imaginations of locally based but globally connected language communities, rather than tech that assumes one-language-fits-all.
- Be creative about using the fullest range of the internet's technologies to explore the fullest range of embodied languages (oral, visual, gestural, text...) so that different forms of knowledge can be easily and accessibly expressed and shared.
- Learn from our Indigenous Nations to design language tech by honoring collective and community memories at the same time as planning for what lies ahead. Let us walk backwards into our futures.

Finally, what can you do?

If someone doesn't speak English as well as you do, it doesn't mean they're stupid. It means that they speak better in one of the other 7000 languages of the world.

All of us with many different skills and experiences need to work together to create and expand a truly multilingual internet. We also need to ensure that the information and knowledges we share in these many languages do not cause harm, but instead lead to the collective good of our world. We need what we call 'solidarity in action'.

If you're in technology:

- Recognize how the policies of your company are contributing (or not) to the multilinguality of the internet, and the deepening of shared human knowledge.
- Center (marginalized) language internationalization and localization work in your strategies, rather than seeing them as peripheral. Do so with a community-partnered approach, rather than a top-down, context-free approach.
- Accept the well-researched criticism of large language models and automated language technologies, and the extensive harms they can cause without thoughtful human oversight.
- Build in careful human processes of context, curation, and moderation of all language work, using smaller community-governed datasets.
- Work respectfully with communities, especially those most marginalized and likely to be most harmed by any lack of care and attention.
- Resource those who give you time and expertise from these language communities.

If you're in a technology standards organization:

- Recognize how context-rich language standards need to be.
- Build better relationships and processes with marginalized language communities so that more standards can be community-partnered, if not community-led.
- Invite with intention more members from marginalized language communities to be part of governance, and give them the resources necessary to participate fully.

If you're in government:

- Recognize that content in the languages of your citizens needs to be accessible for all, not just a privileged minority.
- Support the expansion of content in your languages from and for those who are marginalized or discriminated against, within your regions.

- Support the preservation and digitization of marginalized languages in your region, rather than only the dominant languages.

If you're in free and open-source tech and open knowledge:

- Recognize that free and open-source tech and knowledge also has its own power imbalances and limitations, even though it is meant for collective good.
- Be respectful of the boundaries that marginalized communities set in their sharing of knowledges, because of the ways their knowledges have been historically exploited and commodified in the past.
- Work with marginalized language communities to create the technologies and knowledges they need, rather than those you think they need.

If you're in a GLAM (galleries, libraries, archives, museums and memory) institution:

- Recognize that language is at the heart of the knowledges and cultures you are curating, preserving, and displaying.
- Work with marginalized language communities to ensure that their histories and languages are affirmed, acknowledged and amplified as they want, through the ways you mark provenance (or the ownership and location of materials). This includes the rights of marginalized communities to choose not to have certain knowledges and materials shared publicly. This is critical because many GLAM institutions, especially in the Global North, are rooted in complex histories of colonization and capitalism.
- Ensure that language materials that are stored in your collections are freely and easily accessible to marginalized communities and their allies, so that we can build a collective languages infrastructure together.

If you're in education:

- Recognize how biased our education is towards text-based sources, and certain languages.
- Expand your ways of teaching and learning to include multiple languages and the different forms of language and knowledge they embody.
- Read, listen, and cite work in translation where possible and encourage others to do so too.

If you're in publishing:

- Recognize how skewed towards European colonial languages most publishing currently is in the world.
- Expand the number of languages you publish in, and digitize in all these languages.
- Publish more multilingual books and material.
- Experiment with multimodal forms of publishing, so different forms of oral, visual, and textual language can be simultaneously shared more easily.
- Honor and recognize your translators.

If you're in philanthropy:

- Recognize that language is at the core of human expertise, experience and knowledge of all kinds, no matter what issue you fund.
- Resource multiple language interpretations in all the different global and regional events and convenings you support.
- Support the production, preservation and digitization of materials in the languages of the communities you serve, and ensure your own materials are in the languages of the communities you serve.

If you're in a marginalized language community:

- Recognize that you are not alone.
- Know that it is your community's right to decide what knowledge you would like to share with the world and how.
- Work with the elders, scholars and younger generations in your community, as well as friends from other communities, to collect and share this knowledge.
- If you'd like to be connected to others doing similar work, reach out to us!

If you simply love languages and are wondering what to do:

- Recognize that language is at the heart of who we are and what we do, and is central to different knowledges and cultures — including your own!
- Have conversations with your family, friends and communities to notice how and why English and a few other languages dominate internet access and content, and how to change that together.
- Actively look for, read, listen to and share the contributions of marginalized language communities (including this report!)
- If you'd like to receive updates from our initiative, follow us on social media!

Gratitude

We hold in love, respect, and solidarity the many marginalized communities around the world (Indigenous and beyond) who see language at the core of their identities and ways of being. Their efforts to preserve, vitalize and expand these languages and forms of expressions in meaningful ways, inspire us to imagine more multilingual and plural internet(s) in which we can be the fullest and richest of our many selves. We are also deeply appreciative of all the community and institutional scholars and technologists who love languages as we do, and who work so hard every day to make the internet as multilingual as our physical worlds.

To our many contributors, translators, and communities around the world (especially those who joined our Decolonizing the Internet's Languages conversation in 2019): thank you for all that you do and are in the world, and for being patient with us as we tried to survive the last two years! Thank you especially to our illustrator for her creative visualizations of the essays, and to our animator, who gifted us his animation of these illustrations.

So much gratitude to all our friends and community who reviewed our work from different perspectives, and in different languages. All the mistakes are ours, but your support and solidarity helped make this work-in-progress so much better. And finally, to each other and our families of blood and choice: we couldn't have made it through the last few years (especially 2019-21) without holding each other close, even if only virtually. Love and trust is the best language of all.

Definitions

There are many different ways to define the different aspects of language, and the histories we have been discussing. Not all of these definitions agree with each other! We have used certain terms and phrases in particular ways throughout this report. These are our definitions of these key words and phrases.

- **Dominant languages**: Languages that are either the languages spoken by the majority of the population in a certain area, or that dominate through specific forms of historical power and validation, through legal, political, or cultural forces. For example, Hindi is a dominant language in South Asia, in comparison with many other languages, especially considering that Hindi itself is a family of languages or what some call "dialects". Similarly, Mandarin Chinese is a dominant language in China, through government policy, in comparison with other forms of Chinese as well as other Indigenous languages in the region. Some dominant languages are also "official" or "national" languages in a region or country.

- **European colonial languages**: Languages from Western Europe that spread across Africa, Asia, the Americas, the Caribbean and the Pacific Islands through the processes of colonization by Western European companies and governments, from the 16th century onwards. These include English, Spanish, French, Portuguese, Dutch, and German. It's important to note that these languages were also "colonizer" languages for the Indigenous peoples of North America, not only Latin America (Central and South America).

- **Global South and Global North**: The term "Global South" refers to the regions of Africa, Asia, Latin America and the Caribbean and Pacific Islands that were colonized by Western European countries. It is not meant to be a term of geography. Rather, it is meant to reflect the historical and ongoing socio-economic and political conditions that characterize these countries and regions, and distinguish it from the countries in the privileged countries of Europe and North America, or the "Global North". It was created and amplified by scholars and activists from the Global South to move beyond what they felt were the pejorative and undesirable nature of terms like "least developed" and "developing" nations, and "Third World". Because colonization led to the genocide or decimation of many Indigenous Nations in the Global North, and because some individuals and communities in the Global South benefited and participated in the colonization of their own peoples, we sometimes say that there is a Global South in the Global North and a Global North in the Global South. These structures and processes affect the status of languages in these regions as well (see the term "minoritized majority").

- **Indigenous languages**: Languages spoken by the Indigenous Nations of a particular region or place are Indigenous languages. Indigenous peoples are seen as the "first peoples" or first inhabitants of places across the world that were later colonized and settled by a different cultural group. The majority of the over 7000 languages in the world are spoken by Indigenous communities.

- **Languages and dialects**: We consider any structured system of expression between humans, whether by voice, sound, sign, gesture, or writing, to be language. Some linguists define a "dialect" to describe what sound like different varieties of the same language, but that may be "mutually intelligible" — understood by all speakers of these different varieties who can then talk to each other. However, because the difference between how "language" and "dialect" is defined most commonly is a political (rather than linguistic) choice — based on historical processes of power and privilege — we have rarely used the term dialect in this report. We prefer to use the term "language family", showing that there are many languages that may have similar histories but that also have different characteristics, like the family of languages that is Arabic, Chinese, or Hindi.

- **Local languages**: In this report, we have defined local languages as those languages spoken by the largest number of people in a country or region.
- **Marginalized languages**: In this report, marginalized languages are those languages that are not prominent on the internet in terms of language support or content, i.e. information and knowledge in that language. These languages are marginalized by historical and ongoing structures and processes of power and privilege, including colonization and capitalism, rather than by the population or the number of speakers. Some marginalized languages are already endangered in the world (like many Indigenous languages). But some marginalized languages are spoken by a significant majority of peoples in their region or the world, and yet under-represented online (like Punjabi and Tamil, or Hausa and Zulu, as just a few examples from many dominant languages in Asia and Africa).
- **Minority and majority languages**: Minority languages are those spoken by a minority (in terms of numbers) of the population, in any described territory or region, while a majority language is spoken by the majority (in numbers) of that population.
- **Minoritized majority of the world**: Historical and ongoing structures of power and privilege result in the discrimination and oppression of many different communities and peoples, across the world. These forms of power and privilege are often interlocking and intersecting, so some communities are disadvantaged or oppressed in multiple ways: for instance, by gender, race, sexuality, class, caste, religion, region, ability, and of course, by language. Whether online or in the physical world, these communities make up the majority of the world by population or number, but they are often not in positions of power, and therefore they are treated like a minority. In other words, they are the “minoritized majority” of the world.

More about how to cite and use this report ↪

More about our resources and inspirations ↪