ÉTAT DES LIEUX DES LANGUES D'INTERNET

Rapport de Synthèse · 2022

ÉTAT DES LIEUX DES LANGUES D'INTERNET

Rapport de synthèse

Pourquoi un tel rapport ? Qui sommes-nous ?

Les dictionnaires et les grammaires définissent les langues comme une manière structurée d'exprimer des informations, le plus souvent entre humains. Pourtant, les langues sont bien plus que cela : elles sont l'héritage fondateur que nous offrons à autrui, dont nous héritons le plus souvent de nos ancêtres et, si nous avons de la chance, que nous léguerons à celles et ceux qui nous succèdent. Lorsque nous pensons, lorsque nous parlons, lorsque nous écoutons, lorsque nous imaginons, nous utilisons une langue pour nous et pour les autres. La langue se trouve au cœur de notre identité, de notre rapport au monde. Elle nous aide à raconter des histoires et à partager ce que nous savons de nous-mêmes et des autres. Quelle langue parlez-vous ? Dans quelle langue rêvez-vous ? Pensez-vous dans une langue différente de celle que vous parlez au travail ? La musique que vous aimez est-elle dans une langue que vous ne comprenez pas toujours ?

Chaque langue est un système d'existence, d'action et de communication dans le monde, mais surtout, de connaissance et d'imaginaire. Chacune de nos langues est un système de savoir en elle-même : dans notre langue, nous donnons du sens au monde et l'expliquons aux autres de manière fondamentale. **Nos langues peuvent être orales (parlées et signées), écrites ou transmises par des sons** faits avec un sifflet ou un tambour ! Quelle que soit sa forme, **la langue est une manière d'accéder au savoir**. En d'autres termes, la langue est la manière la plus évidente pour exprimer ce que nous pensons, croyons et savons.

Maintenant, réfléchissez aux langues que vous parlez, dans lesquelles vous pensez, rêvez ou écrivez. Dans combien de ces langues êtes-vous en mesure de partager et communiquer véritablement dans les espaces numériques ? Quelle est votre expérience de l'utilisation d'une ou plusieurs langues en ligne ? Le matériel que vous utilisez est-il équipé des caractères de votre langue ? Devez-vous modifier vos claviers pour les utiliser dans votre langue ? Lorsque vous cherchez des informations à l'aide d'un moteur de recherche, les résultats sont-ils retournés dans la langue de votre choix ? Avez-vous dû apprendre une autre langue que la vôtre pour accéder et contribuer à Internet ? Si vous répondez « non » à une ou plusieurs de ces questions, alors vous faites partie de la minorité des personnes privilégiées dans le monde capables d'utiliser facilement Internet dans leur propre langue. Et il y a de grandes chances que votre langue soit… le français.

Actuellement, Internet et ses différents espaces numériques représentent l'une des infrastructures les plus essentielles pour diffuser des savoirs, communiquer et agir. Pourtant, parmi les 7000 langues du monde (y compris les langues parlées et signées), **de combien pouvons-nous faire véritablement l'expérience en ligne ? Quels seraient l'aspect, l'ambiance et les sonorités d'un Internet vraiment multilingue ?**

Ce rapport est une tentative de réponse à cette question. Nous sommes un collectif de trois organisations : Whose Knowledge?, Oxford Internet Institute et le Centre for Internet and Society (Inde). Nous nous sommes rassemblé·es pour proposer différents points de vue, expériences et analyses des langues sur Internet. En partenariat avec des tiers intéressés par ces problématiques, nous espérons créer un Internet, des technologies et des pratiques numériques plus multilingues.

Ce rapport a trois objectifs :

- **Décrire l'état actuel des langues sur Internet** : nous essayons de comprendre quelles langues sont actuellement représentées sur Internet et comment. Pour cela, nous utilisons des données quantitatives (en étudiant des chiffres provenant de différents outils, plateformes et espaces numériques), ainsi que des données qualitatives (en écoutant des histoires et expériences personnelles en lien avec les langues sur Internet).
- **Sensibiliser aux défis et opportunités d'un Internet plus multilingue** : créer et gérer les technologies, le contenu et les communautés pour toutes les langues du monde représente un défi de taille, ainsi que des possibilités et opportunités prometteuses. Ce rapport exposera certains de ces défis et possibilités.
- **Proposer un programme d'action** : avec ces éléments d'information et de sensibilisation, nous proposons des manières selon lesquelles nous, et beaucoup d'autres parties prenantes travaillant sur ces problématiques dans le monde, souhaitons nous organiser et agir pour garantir un Internet plus multilingue.

Portée et limites de ce rapport

Ce rapport est un travail et un processus en cours (d'amélioration continue).

De nombreuses personnes, communautés et institutions travaillent sur différents aspects des langues depuis très longtemps et, plus récemment, sur les différents aspects des langues sur Internet. Leurs travaux nous inspirent, mais ce rapport n'a pas vocation à tou·tes les représenter. Nous ne connaissons pas non plus tou·tes les acteur·ices travaillant sur les langues et Internet, bien que nous ayons tenté d'inclure la plupart de celles et ceux que nous connaissons et qui nous inspirent, d'une manière ou d'une autre, en les faisant figurer dans nos sections Ressources et Gratitude.

Nous sommes limité·es par les données que nous avons pu rassembler et nous exposons certaines de ces contraintes dans la section Chiffres. Nous accueillons les commentaires et les suggestions d'amélioration ou de mise à jour des informations présentées ici. Nous serions ravi·es d'avoir des retours de la part des personnes qui travaillent déjà sur ces problématiques et qui voudraient faire partie des prochaines mises à jour de ce rapport.

Nous avons fait tout notre possible pour rédiger ce rapport dans un style accessible. Nous souhaitons que des générations et des communautés variées nous rejoignent dans notre travail et nous ne voulons pas que la langue « académique » ou le jargon soient un obstacle à la lecture et la réflexion. Nous souhaitons également que ce rapport soit traduit dans le plus de langues possible (traducteurs et traductrices : contactez-nous !). Bien que ce rapport ait été rédigé d'abord en anglais, nous ne voulons pas que la maîtrise de cette langue soit un prérequis à la réflexion ou à l'action.

Nous espérons que ce rapport servira de « point de départ » à des recherches, discussions et actions futures sur ces problématiques, tout en se fondant sur les nombreux efforts déjà effectués.

Qui sommes-nous et pourquoi nous sommes-nous associé·es pour ce rapport ?

Trois organisations se sont rapprochées pour effectuer les recherches de ce rapport : le Centre for Internet and Society, l'Oxford Internet Institute et Whose Knowledge?. Nous nous intéressons toutes les trois aux implications d'Internet et des technologies numériques du point de vue de la recherche, de la politique et de la sensibilisation.

Au cours des années passées, nous avons travaillé indépendamment pour comprendre les inégalités et injustices liées au savoir sur Internet : qui contribue au contenu en ligne et comment ? Nous avons vite compris qu'il n'existait que peu de données sur le savoir en différentes langues sur Internet. Nous avons alors voulu en savoir plus : quelle proportion des langues du monde se trouve sur Internet à l'heure actuelle ? Dans quelle mesure Internet est-il multilingue ? Notre exploration s'est limitée aux quelques domaines dans lesquels nous avons pu trouver des informations publiques, ouvertes et utiles, mais nous espérons que cette contribution supplémentaire soutiendra toutes les personnes qui, comme nous, travaillent pour un Internet multilingue.

Remarque sur la COVID-19 et ses conséquences sur ce rapport : nous avons commencé à travailler sur ce rapport en 2019, avant la COVID-19, mais la plupart des travaux d'analyse, d'entretien et de rédaction se sont produits pendant la pandémie qui a changé nos vies à un niveau individuel et collectif dans le monde entier. Toutes les personnes ayant contribué à ce rapport ont été

affectées par cette crise et nous avons mis plus de temps que prévu pour le publier. Mais la COVID-19 nous a également aidé·es à nous rappeler à quel point nous sommes interconnecté·es, combien il est essentiel d'être capable de transmettre des idées complexes dans différentes langues et l'importance de disposer d'infrastructures (numériques) résilientes et accessibles qui soient vraiment multilingues.

Comment lire ce rapport ?

Ce rapport est conçu pour le « numérique d'abord », c'est-à-dire que le meilleur moyen de le lire, de l'écouter et d'en tirer des connaissances est de le consulter via ce site Web. La version numérique est la plus adaptée, car le rapport est composé de différents niveaux et couches combinant des chiffres et des histoires. Nous informer sur l'état des langues en ligne depuis une perspective statistique nous donne un aperçu de ces problématiques et des différentes situations vécues par les personnes. Toutefois, ce sont les expériences individuelles des langues sur Internet dans le monde, dans différents contextes, qui nous aident à approfondir notre compréhension de la facilité ou la difficulté ressentie par les personnes utilisant Internet dans leurs langues. Avec des histoires et des chiffres, nous pouvons commencer à entrevoir des opportunités, des contextes et des défis sous-jacents.

C'est pourquoi ce rapport est composé de trois parties principales :

- La synthèse résumant la création et le contenu de l'état des lieux des langues (ce que vous lisez à cet instant !)
- Des chiffres qui analysent plusieurs problématiques linguistiques critiques sur certains appareils, plateformes et applications numériques que nous utilisons tous les jours. Ce sont nos ami·es de l'Oxford Internet Institute qui ont mené ce travail et vous trouverez des analyses et visuels intéressants de leurs données dans cette partie. Veuillez noter que cette analyse se limite aux données auxquelles nous avions accès, soit des jeux de données et supports ouverts et disponibles au public. D'autres contraintes méthodologiques sont exposées plus en détail dans ces essais, mais la principale difficulté est de trouver une manière unique et cohérente d'identifier les langues. Il est également complexe d'estimer le nombre de personnes utilisant une langue spécifique, notamment en raison du caractère dynamique et évolutif de l'utilisation des langues.
- Des histoires qui nous apportent une compréhension approfondie de l'expérience des différentes personnes et communautés du monde sur Internet et souvent, de leurs difficultés à trouver les informations dont elles ont besoin dans leurs propres langues. Nous avons sollicité ces histoires sous forme écrite et parlée. Vous trouverez donc des essais rédigés, ainsi que des entretiens audio et vidéo. Nos ami·es du Centre for Internet and Society ont mené ce travail en combinant ce riche patchwork d'expériences des langues du monde entier. Nous avons inclus des contributions concernant les langues

autochtones d'Afrique, des Amériques et d'Australie comme le chindali, le cri, l'ojibwé, le mapudungun, le zapotèque et l'arrernte ; des langues minoritaires comme le breton, le basque, le sarde et le carélien en Europe ; ainsi que des langues dominantes aux niveaux régional ou mondial comme le bengali, l'indonésien (bahasa Indonesia) et le cinghalais en Asie et différentes formes d'arabe en Afrique du Nord.

Nos contributeur·ices ont écrit ou parlé dans leurs propres langues, ainsi qu'en anglais. Notre synthèse est également écrite et parlée dans différentes langues. Nous espérons que vous apprécierez votre lecture ou votre écoute dans plusieurs langues !

Nous nous sommes aussi efforcé·es de donner vie à ces contributions sous forme visuelle à l'aide d'illustrations et d'animations créatives qui synthétisent les aspects sociaux et techniques des langues. Comme pour tous les autres éléments de notre rapport, ces visuels ont été développés en collaboration avec notre illustratrice et nos contributeur·ices.

Dans quelle mesure Internet est-il multilingue ?

Internet n'est pas encore aussi multilingue que la vie réelle (et malheureusement, il n'est pas près de le devenir). Nous essayons de comprendre pourquoi en analysant des chiffres et des expériences vécues par des personnes du monde entier. Voici un bref résumé de la richesse et de l'envergure des travaux menés par nos contributeur·ices. N'hésitez pas à consulter leurs essais pour plus de détails et d'inspiration.

Nous analysons d'abord les contextes dans lesquels les personnes utilisent Internet dans le monde et en différentes langues. Nous étudions de quelles manières les informations et connaissances sont réparties, ou non, selon les langues et les régions du monde. Puis, nous examinons plus en détail les plateformes et applications majeures que nous utilisons pour créer du contenu, communiquer et partager des informations en ligne, et combien de langues sont prises en charge par chacune d'elles. Nous nous intéressons à Google Maps et Wikipédia, des espaces de contenus multilingues utilisés dans la vie quotidienne, et observons leur fonctionnement en plusieurs langues.

Tout au long de notre analyse, nous partageons les histoires et expériences des personnes qui accèdent et contribuent aux savoirs dans leurs propres langues sur Internet. Comme nous l'avons appris, la plupart de nos contributeur·ices sont obligé·es d'utiliser une autre langue que leur première langue choisie afin d'accéder et de contribuer aux problématiques qui leur tiennent à cœur.

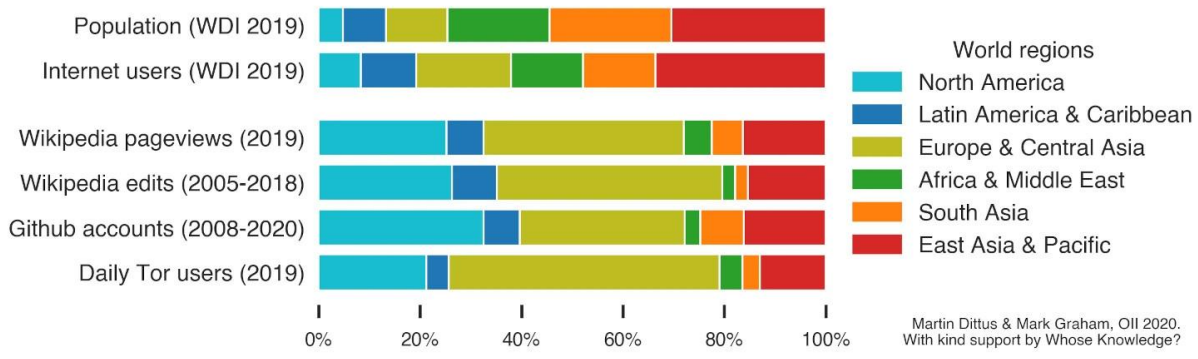
Le contexte des langues : inégalités géographiques et numériques du savoir

«Les langues de tradition orale ne trouvent pas leur place dans le Web d'aujourd'hui.»
Ana Alonso

«Nous avons l'impression que ces plateformes, en général, perpétuent l'idée colonialiste qu'il existe des langues ayant une plus grande valeur et capacité à communiquer, ce qui nuit à l'image des langues minoritaires comme le mapudungun.»
Kimeltuwe project

Nous savons que plus de 60 % du monde est désormais connecté au numérique, pour la plupart des gens par téléphone ou appareil mobile. Parmi toutes les personnes connectées, les trois quarts se situent dans les pays du Sud : Asie, Afrique, Amérique latine, les îles des Caraïbes et du Pacifique. Pourtant, notre accès à Internet est-il véritable et équitable ? Sommes-nous en mesure de créer et produire des savoirs publics en ligne dans une proportion égale à notre consommation ?

L'étude de Martin et Mark sur le rapport entre population mondiale et nombre d'utilisateur·rices d'Internet montre que certains groupes peuvent accéder à Internet de façon plus significative que d'autres, y compris dans des espaces numériques très connus. Par exemple, alors qu'une majorité d'entre nous se trouve dans les pays du Sud, nous ne sommes pas en mesure d'accéder à Internet dans un rôle de créateur·ices et producteur·ices de savoir, mais uniquement en tant que consommateur·ices. La plupart des modifications sur Wikipédia, la majorité des comptes sur Github (un répertoire de code) et la majorité des utilisateurs de Tor (un navigateur sécurisé) sont originaires d'Europe et d'Amérique du Nord.

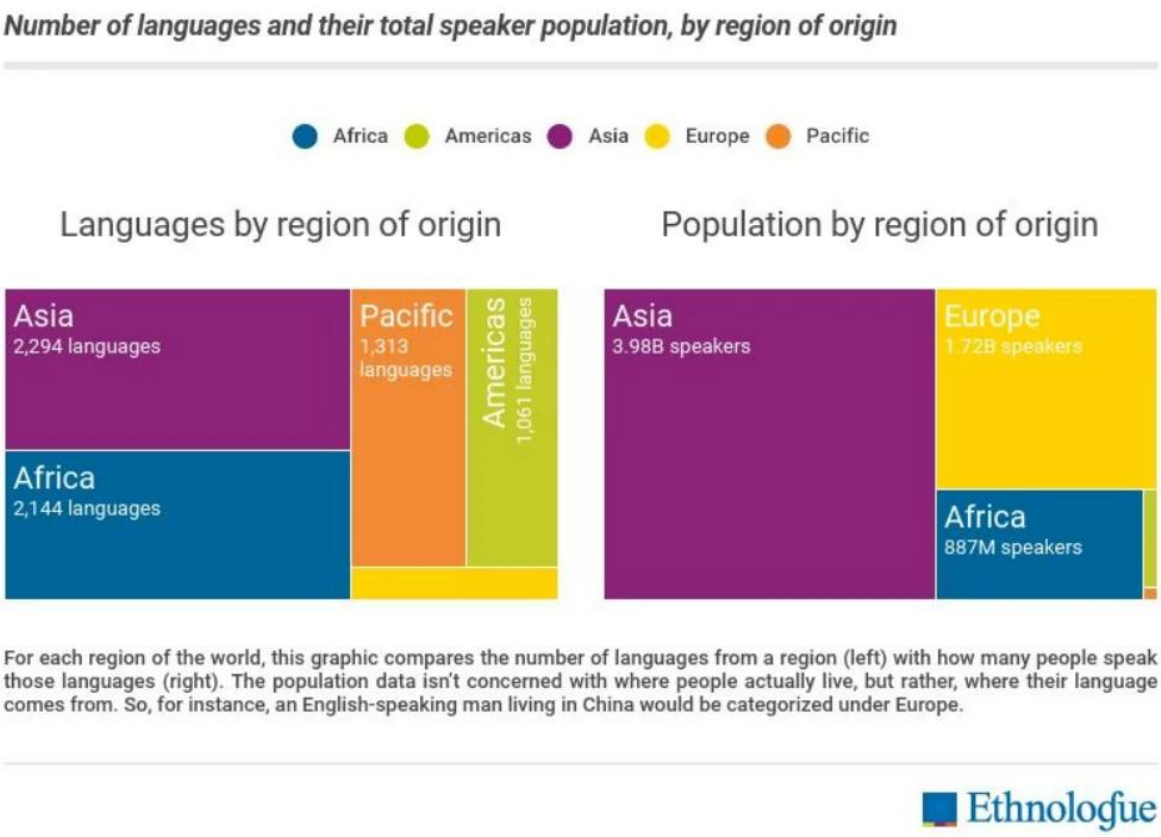


Mesures de la participation au numérique par région du monde. (Données : Banque mondiale 2019, Wikimedia Foundation 2019, Wikipédia 2018, Github 2020, Tor 2019)

Que signifie cette inégalité d'accès pour les langues ? Sommes-nous tous et toutes en mesure d'accéder à Internet dans nos propres langues ? Sommes-nous en mesure de créer du contenu et des informations dans nos propres langues ?

Comme le montrent d'autres projections, plus de 75 % des personnes accèdent à Internet dans seulement 10 langues, dont la plupart ont un passé colonial européen (anglais, français, allemand, portugais, espagnol…) ou sont dominantes dans des régions spécifiques où d'autres langues ont des difficultés à rester actuelles (chinois, arabe, russe…). En 2020, on estimait que 25,9 % de l'ensemble des internautes utilisaient l'anglais, tandis que 19,4 % accédaient à Internet en chinois. La Chine est le pays comptant le plus d'utilisateur·ices d'Internet au monde. Il convient de ne pas oublier que le « chinois » n'est pas une langue unique, mais une famille de nombreuses langues différentes.

Paradoxalement, la plupart des langues actuelles d'Internet sont originaires d'Europe, alors que c'est le continent qui compte le moins de langues au monde. On dénombre plus de 7000 langues dans le monde, dont plus de 4000 sont parlées en Asie et en Afrique (avec plus de 2000 langues pour chaque continent), tandis que les îles du Pacifique et les Amériques comptent plus de 1000 langues pour chaque région. La Papouasie–Nouvelle-Guinée et l'Indonésie sont les pays comptant le plus de langues, avec 800 langues en Papouasie–Nouvelle-Guinée et plus de 700 en Indonésie.



Le nombre de langues et la population totale des locuteur·ices par région dans le monde. Source : Ethnologue

De nombreuses langues de l'Asie du Sud (hindi, bengali, ourdou...) se trouvent parmi les 10 langues du monde ayant le plus de locuteur·rices] dont c'est la première langue (ou natifs). Pourtant, elles ne permettent pas aux habitant·es de cette région d'accéder à Internet. Et bien sûr, l'histoire d'Ishan, dont la première langue est le bengali, nous apprend que même si vous pouvez accéder à des connaissances numériques dans la langue de votre choix, le type d'informations que vous recherchez risque de ne pas exister. Dans le cas d'Ishan : du contenu sur le handicap et les droits liés à la sexualité. En Asie du Sud-Est, une région qui concentre l'un des plus importants foyers d'utilisation d'Internet au monde et l'une des plus grandes diversités linguistiques, la situation est semblable. Paska se heurte aux mêmes problèmes de disponibilité du contenu sur les droits liés à la sexualité en indonésien (bahasa Indonesia) qu'Ishan en bengali.

Nous savons également que, sur plus de 7000 langues du monde, environ 4000 disposent de systèmes ou de caractères d'écriture. Toutefois, la plupart de ces écritures n'ont pas été développées par les locuteur·rices de ces langues, mais par les nombreux mécanismes

colonisateurs à l'œuvre dans le monde entier. L'existence d'une forme d'écriture ne signifie pas que celle-ci est largement comprise ou utilisée. La plupart des langues du monde sont transmises sous forme parlée ou signée et non par l'écrit. Même dans le groupe des langues disposant de formes d'écriture, l'édition favorise les langues coloniales européennes, et dans une moindre mesure, les langues dominantes régionales. En 2010, Google estimait qu'il existait un total d'environ 130 millions de livres publiés, dont une importante proportion était écrite en environ 480 langues. La plupart des revues académiques réputées en sciences ou en sciences sociales sont publiées en anglais. Le livre le plus traduit au monde est la Bible (en plus de 3000 langues). Le document le plus traduit au monde est la Déclaration universelle des droits humains des Nations Unies (en plus de 500 langues).

Pourquoi est-ce important ? Car les technologies linguistiques numériques reposent sur le traitement automatisé des publications dans chaque langue afin d'améliorer la prise en charge des langues et contenus. Ainsi, lorsque la publication de textes dans le monde surreprésente certaines langues et omet les langues non écrites, les inégalités linguistiques que nous constatons sont aggravées. Et bien sûr, les langues non basées sur des textes, qui utilisent des signes, sons, gestes et mouvements, sont totalement absentes du secteur de l'édition, donc souvent absentes des technologies linguistiques numériques.

Par exemple, Ana nous explique : « le Web n'est pas conçu pour répondre aux utilisateur·ices parlant des langues de tradition orale uniquement ». Dans ce contexte de domination des langues écrites sur Internet, il est difficile de trouver du contenu issu des traditions linguistiques orales ou visuelles. Nous ne pouvons pas facilement rechercher des gestes, des signes et des sifflements, par exemple. Dans un entretien, Joel et Caddie nous parlent du premier ensemble d'émojis autochtones d'Australie conçu sur le territoire Arrernte à Mparntwe/Alice Springs. Ils nous expliquent comment le geste corporel est souvent associé à la parole pour produire du sens en arrernte. Emna nous dit la même chose sur la Tunisie et les différentes langues parlées par son peuple : « lorsque qu'il s'agit de préserver une langue, nous ne devons pas nous limiter à l'écrit, nous devons également en préserver les formes orales, gestes, signes, sifflements, etc. qu'il est impossible de capturer entièrement à l'écrit ».

Les technologies numériques nous offrent la possibilité de représenter la pluralité des formes linguistiques basées sur le texte, le son, la gestuelle et bien plus. Elles peuvent également nous aider à préserver et ramener à la vie des langues menacées d'extinction : plus de 40 % de toutes les langues. Tous les mois, deux langues autochtones et leurs savoirs meurent et sont perdus à jamais.

Pourquoi ces différents contextes linguistiques ne sont-ils pas mieux représentés en ligne ?

Dans son essai, Claudia nous propose trois axes d'étude pour comprendre les relations entre langues et technologies : disponibilité, utilisabilité et mode de développement des technologies.

Comme nous l'observons dans ce rapport, les langues dites « majoritaires » (étant pour la plupart des langues coloniales européennes ou des langues régionales dominantes) sont disponibles sur toute une gamme de média, services, interfaces et applications, tandis que les autres langues sont beaucoup moins disponibles, notamment dans les infrastructures comme les claviers, la traduction automatique ou la reconnaissance vocale. Les entreprises technologiques dépensent beaucoup de temps et de ressources sur l'utilisabilité de ces langues majoritaires, car c'est là qu'elles identifient le plus de bénéfices. Pour finir, Claudia conclut que la plupart des technologies linguistiques sont développées par des procédures descendantes avec peu de collaboration des communautés linguistiques, ou que les rares efforts de travail avec des communautés sont mal planifiés et manquent de coordination.

Ces difficultés et défis liés au contexte représentent également des pistes d'action pour créer un Internet plus multilingue et nous reviendrons sur ces possibilités plus tard.

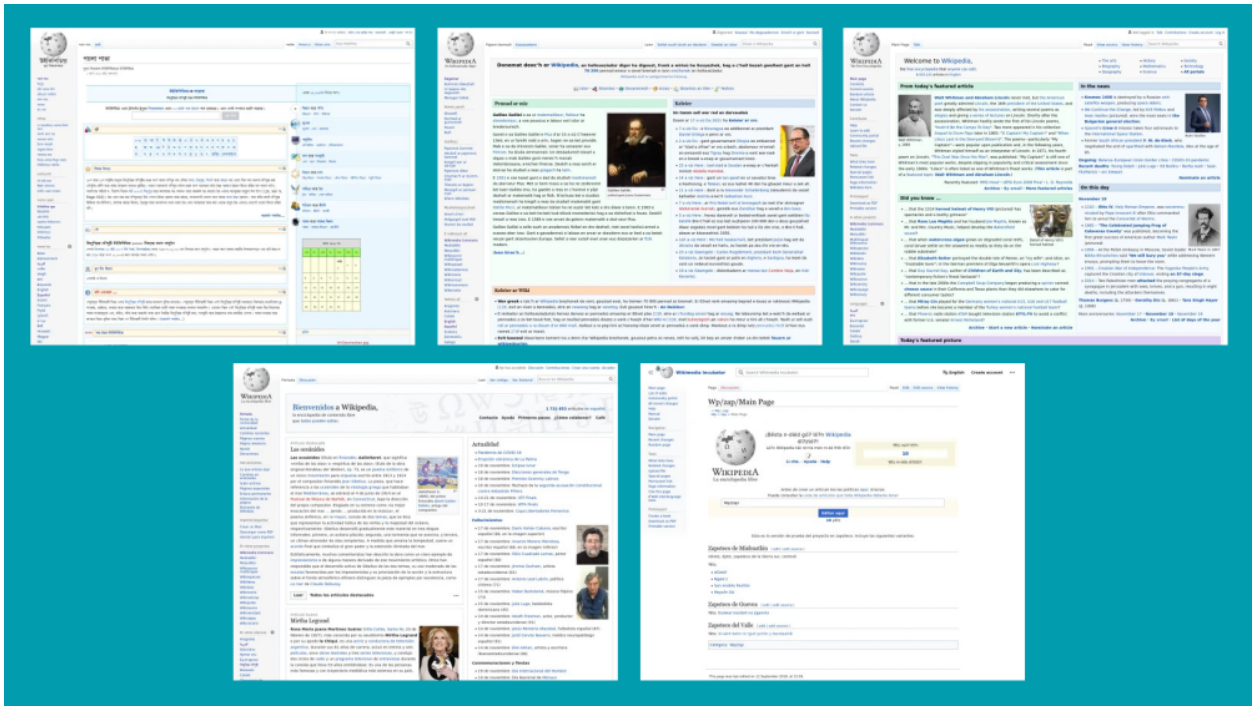
Prise en charge linguistique : plateformes et applications de messagerie

«Lorsque vous écrivez le mot « bonjour », avant d'avoir fini de taper toutes les lettres, votre téléphone ou ordinateur vous suggère le mot. Lorsque j'écris le même mot en chindali « mwalamusha », je dois taper le mot entier, ce qui prend plus de temps et il sera souligné parce que l'ordinateur ou le téléphone ne le reconnaît pas.»
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

«Si les applications populaires et les interfaces des logiciels essentiels ne sont pas disponibles en breton rapidement, cette langue, qui ne peut pas rivaliser avec les applications en français, va forcément perdre son attractivité pour les plus jeunes générations.»
Claudia Soria

Nous avons creusé la question pour comprendre dans quelle mesure Internet manque de multilinguisme par rapport au monde dans lequel nous vivons. Nous avons analysé le type de prise en charge linguistique, c'est-à-dire les interfaces en différentes langues, que les plateformes et applications numériques majeures nous fournissent pour communiquer, créer et partager du contenu dans nos langues.

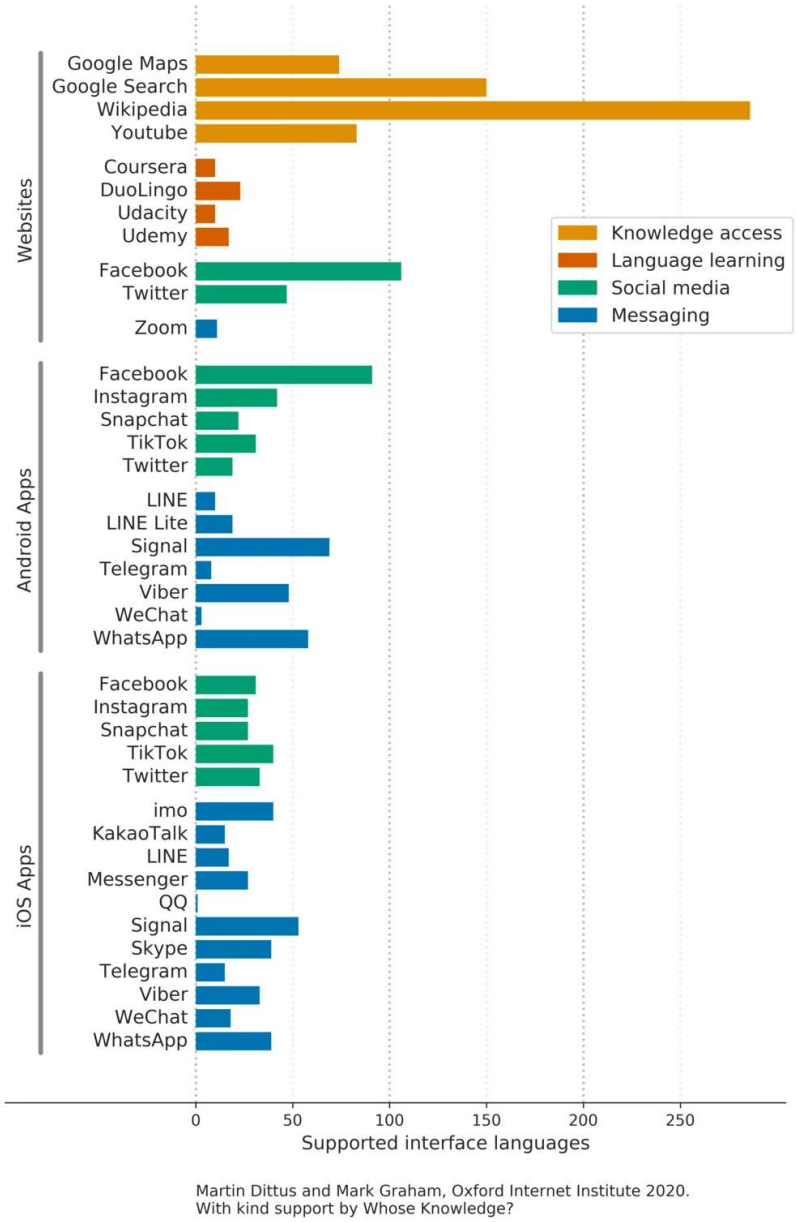


Interface de Wikipédia en plusieurs langues.

Martin et Mark ont analysé la prise en charge linguistique sur 11 sites Web, 12 applications Android et 16 applications iOS. Ils ont choisi des plateformes largement utilisées, spécialisées dans le recueil et le partage des savoirs, notamment celles qui visent une présence et une audience mondiales. Eux aussi ont dû utiliser les données disponibles publiquement sur ces plateformes et applications.

Ces plateformes sont groupées en quatre grandes catégories (pouvant se superposer) :

- **Accès au savoir** (plateformes de savoirs et informations, y compris les moteurs de recherche) : Google Maps, Google Search, Wikipédia, YouTube.
- **Apprentissage linguistique** (plateformes d'apprentissage des langues autoguidé) : DuoLingo et les plateformes d'études Coursera, Udacity, Udemy.
- **Réseaux sociaux** (plateformes de réseaux sociaux publics) : Facebook, Instagram, Snapchat, TikTok, Twitter.
- **Messagerie** (messagerie privée et groupée) : 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.

Nous avons découvert que la prise en charge des langues écrites est fortement inégale en fonction des plateformes numériques. Les plateformes Web majeures comme Wikipédia, Google Search et Facebook offrent actuellement la prise en charge linguistique la plus large. Wikipédia (une organisation à but non lucratif dont les articles sont rédigés par des bénévoles du monde entier) est de loin la plateforme la plus traduite. Wikipédia prend en charge plus de 400 langues avec une interface utilisateur de base et environ 300 langues disposent d'au moins 100 articles. Google Search prend en charge 150 langues, tandis que Facebook prend en charge entre 70 et 100 langues. Signal est en tête des applications de messagerie avec presque 70 langues prises en charge sur Android et 50 sur iOS. Cependant, la plupart des plateformes

concentrent leur prise en charge linguistique sur un petit nombre de langues très répandues, ce qui laisse une majorité de langues non prises en charge. L'application de messagerie QQ, par exemple, ne prend en charge que le chinois.

Les quelques langues souvent prises en charge par la plupart des plateformes étudiées incluent les langues européennes telles que l'anglais, l'espagnol, le portugais et le français, ainsi que certaines langues asiatiques, comme le chinois mandarin, l'indonésien, le japonais et le coréen. Des langues majeures comme l'arabe et le malais sont moins souvent prises en charge et d'autres langues parlées par des dizaines, voire des centaines de millions de locuteur·rices ne sont pas bien représenté·es en matière de prise en charge d'interface.

Que signifie cette absence de prise en charge linguistique pour la plupart des personnes dans le monde ? En 2021, la population mondiale est estimée à environ 7,9 milliards d'humains, dont plus de la moitié vit en Asie (presque 4,7 milliards) et en Afrique (presque 1,4 milliard). Pourtant, la majorité de la population mondiale n'est pas servie par les langues d'Internet :

- *Les personnes parlant des langues africaines* : la vaste majorité des langues africaines n'est prise en charge par aucune des interfaces des plateformes étudiées. En conséquence, plus de 90 % des Africain·es doivent utiliser une seconde langue afin d'accéder à une plateforme. Pour nombre d'entre eux, cela signifie utiliser une langue coloniale européenne ou une langue plus dominante dans leur région.
- *Les personnes parlant des langues d'Asie du Sud* : presque la moitié des plateformes étudiées ne prennent en charge aucune langue de cette région sur leur interface. Même les langues majeures d'Asie du Sud, comme l'hindi et le bengali, parlées par des centaines de millions de personnes, sont moins largement prises en charge que d'autres langues.
- *Les personnes parlant des langues d'Asie du Sud-Est* : la prise en charge des langues d'Asie du Sud-Est est tout aussi inégale. Tandis que l'indonésien, le vietnamien et le thaï ont tendance à être très bien pris en charge par les plateformes étudiées, la plupart des autres langues d'Asie du Sud-Est ne le sont pas.

Les conclusions de Martin et Mark sont renforcées par les réalités quotidiennes des personnes vivant dans ces régions du monde. Par exemple, au Malawi, Donald a constaté que lorsqu'il demandait aux locuteur·ices du chindali (langue bantoue en danger d'extinction) comment ils et elles communiquaient par téléphone, toutes les réponses obtenues expliquaient combien c'était lent et laborieux dans cette langue. En effet, la plupart de leurs téléphones, conçus pour une prise en charge linguistique en anglais, français ou arabe, ne reconnaissent pas le chindali. Ces difficultés technologiques s'additionnent aux contraintes économiques et sociales qui limitent la capacité des locuteur·ices chindalis à acquérir un smartphone ou un abonnement Internet. Même pour celles et ceux qui utilisent la langue officielle du Malawi, le chichewa, le

manque de prise en charge linguistique représente une difficulté : « Pourquoi acheter un téléphone coûteux ou perdre mon temps à aller sur Internet, si tout est en anglais, une langue que je ne comprends pas ? »

En effet, en 2018, l'absence de prise en charge linguistique pour la plupart des langues africaines s'est fait cruellement ressentir lorsque Twitter a pour la première fois reconnu le swahili, une langue parlée (comme première ou seconde langue) par plus de 50 à 150 millions de personnes en Afrique de l'Est et dans d'autres régions. Avant cette date, le swahili et la plupart des langues africaines étaient considérés comme de l'indonésien par la plateforme. La reconnaissance des mots swahilis et la prise en charge de la traduction n'ont pas été initiées par l'entreprise : elles résultent d'une campagne menée par les locuteurs du swahili utilisant Twitter.

La situation n'est pas tellement meilleure pour les langues autochtones d'Amérique latine. Dans l'entretien sur le projet Kimeltuwe portant sur le mapudungun, parlé par les peuples mapuches au Chili et en Argentine, on apprend : « il serait génial de pouvoir publier en mapudungun sur des plateformes comme YouTube ou Facebook. Pas seulement que l'interface soit traduite, mais de pouvoir étiqueter, dans les menus disponibles, la langue comme étant du mapudungun. Par exemple, lorsque je charge une vidéo sur YouTube ou Facebook, je ne peux pas ajouter une transcription en mapudungun puisque la langue n'apparaît pas dans la liste prédéterminée. Si vous voulez charger une transcription en mapudungun, vous devez indiquer qu'elle est en espagnol ou en anglais. »

Martin et Mark n'ont pas analysé la prise en charge linguistique sur des appareils spécifiques, comme les téléphones portables, mais nous savons que les claviers numériques sont l'un des espaces critiques dans lesquels les linguistes et les spécialistes des technologies ont fait le plus de progrès. Par exemple, Gboard, le clavier de smartphone de Google pour le système d'exploitation Android, prend en charge plus de 900 langues grâce à un travail important réalisé avec des communautés linguistiques et des universitaires. Toutefois, un clavier de smartphone avec ces capacités n'est accessible que si l'on peut se permettre un appareil assez haut de gamme.

En parallèle, l'expérience d'Uda avec le cinghalais, une langue parlée par plus de 20 millions de personnes au Sri Lanka comme première ou seconde langue, montre toute la difficulté à créer du contenu dans une langue dont l'écriture n'est pas facilement comprise par certains des spécialistes des technologies travaillant sur la prise en charge linguistique, surtout quand les caractères diffèrent beaucoup de l'alphabet latin des langues d'Europe occidentale. Elle explique : « le principal problème avec le cinghalais Unicode est lié à l'ordre dans lequel les différents caractères doivent être tapés afin de produire une lettre. Cet ordre exige une consonne suivie d'une voyelle. C'est une logique qui suit les règles des langues européennes basées sur un alphabet latin. Cependant, en cinghalais, la voyelle précède parfois la consonne.»

Unicode est le standard technologique de codage du texte exprimé dans un système d'écriture ou alphabet. La version 13 comprend 143 859 caractères pour plus de 30 systèmes d'écriture en usage aujourd'hui, puisque plusieurs langues partagent le même (par exemple, l'alphabet latin couvre la plupart des langues européennes, les caractères han servent pour le japonais, le chinois et le coréen, et l'écriture devanagari est utilisée par plusieurs langues d'Asie du Sud). Elle contient aussi les caractères d'écriture ancienne pour des langues mortes. L'Unicode Consortium (une ONG californienne) décide également des émojis, les symboles que nous utilisons quotidiennement sur différentes interfaces.

L'étude de Martin et Mark sur la prise en charge linguistique et les expériences liées aux langues d'autres contributeur·ices du monde entier contient beaucoup plus d'informations que cette brève description des prises en charge techniques hétérogènes et limitées pour la plupart des langues sur les plateformes et applications à l'heure actuelle. N'hésitez pas à la lire.

Contenu linguistique : accessibilité et production

«Le contenu féministe est particulièrement inaccessible dans les langues locales. La Women's Development Foundation (Fondation pour le développement des femmes) est un groupe de femmes rurales travaillant sur les problématiques des droits des femmes depuis 1983. Mais c'est seulement en 2019 que nous avons commencé à partager du contenu féministe en cinghalais concernant des problématiques sociopolitiques et économiques en ligne.»
Uda Deshapriya

«Malheureusement, il était, et il reste toujours difficile de trouver du contenu queer éducatif et positif en indonésien sur Internet… Si nous cherchons les termes « LGBT » ou « homoseksualitas » (homosexualité) sur Google, le moteur de recherche le plus exhaustif et le plus utilisé, nous trouverons beaucoup de résultats contenant les mots « penyimpangan » (déviance), « dosa » (péché) et « penyakit » (maladie).»
Paska Darmawan

«Les informations sur l'intersection de la sexualité queer et du handicap (ou même son absence) disponibles en bengali sur Internet sont largement influencées par l'homophobie et le validisme qu'elles renforcent à leur tour.»
Ishan Chakraborty

Nous avons voulu comprendre le contenu sur Internet en analysant quelle version du monde et quels savoirs nous sont présentés lorsque nous sommes en ligne. Après tout, plus de 63 % de tous les sites Internet ont l'anglais comme langue principale pour leur contenu.

Dans leurs essais et entretiens, nos contributeur·ices parlent des différentes constellations d'obstacles historiques, sociopolitiques, économiques et technologiques qui empêchent un accès significatif à Internet dans leurs langues. Plus révélateur encore, tous et toutes évoquent les difficultés éprouvées à trouver du contenu pertinent sur Internet et à créer du contenu important pour eux dans ces langues. En d'autres termes, il ne suffit pas d'être en mesure d'accéder à des informations et à des savoirs : souvent, ces contenus sont créés pour nous dans d'autres langues par des personnes qui ne comprennent pas toujours nos contextes et expériences, ou pire, qui y sont hostiles. Nous devons être en mesure de produire des savoirs importants pour nous-mêmes et nos communautés, ou à minima, d'être en mesure de prendre en charge la production et l'expansion de ce contenu dans toutes nos langues différentes.

C'est particulièrement vrai pour les personnes qui rencontrent des difficultés d'accessibilité et celles qui subissent plusieurs formes de marginalisation et d'exclusion.

Comme Joel nous l'a décrit dans son entretien à propos du projet Indigemoji, tout a commencé par un tweet de frustration depuis sa voiture. Un jour, il s'est arrêté sur le bord de la route et il a commencé à associer des mots arrernte avec des émojis pour en décrire la signification. Pendant des dizaines d'années après l'apparition des émojis sur Internet, les Premières Nations ou les peuples autochtones ont présenté des demandes d'émojis pour exprimer leurs langues orales ou visuelles, comme l'arrernte, en vain. Comme nous l'avons décrit plus tôt, c'est l'Unicode Consortium qui étudie les demandes publiques pour de nouveaux émojis. Des pétitions telles qu'un drapeau aborigène australien ont été refusées. Pour Joel, Caddie et bien d'autres, le projet Indigemoji est devenu un effort multigénérationnel pour combattre physiquement et virtuellement de multiples formes de marginalisation et créer leur propre contenu de manière cohérente avec leurs identités et langues autochtones.

Un tweet associant une liste d'émojis avec des mots arrernte. Source : Indigemoji

Il est important de se souvenir que si les langues autochtones sont des langues « minoritaires » dans le monde d'aujourd'hui, c'est en raison du génocide de masse dû à la colonisation. Les Nations autochtones ont été détruites ou réduites à des minorités, alors qu'elles représentaient la population principale d'une région ou d'un territoire particulier. Ces processus de colonisation affectent également les langues dominantes parlées par des millions de personnes dans le monde.

Ishan est un universitaire malvoyant et queer pour qui aller sur Internet est une prouesse. Il a également des difficultés à trouver des informations pertinentes en bengali sur les sujets du handicap, de la sexualité queer et encore plus sur l'intersection entre ces problématiques. Cela mène à ce qu'il appelle la « marginalité au sein de la marginalité » : « d'un côté les attitudes homophobes et validistes de la société et de l'autre, l'homophobie et/ou le validisme internalisés par les individus (queer et/ou handicapé·es). Ces situations se complètent et perpétuent le mécanisme de la marginalisation. L'emplacement sociétal d'un individu handicapé queer peut être décrit comme à la « marge de la marge » ».

En d'autres termes, des dynamiques de critique de l'accès et de l'information, même dans une langue dominante telle que le bengali, parlée par environ 300 millions de personnes dans le monde, sont absentes d'Internet.

Martin et Mark ont décidé d'aller plus loin en analysant la portée et le type de contenu en plusieurs langues, sur deux plateformes d'information et de savoirs différentes : Google Maps et Wikipédia.

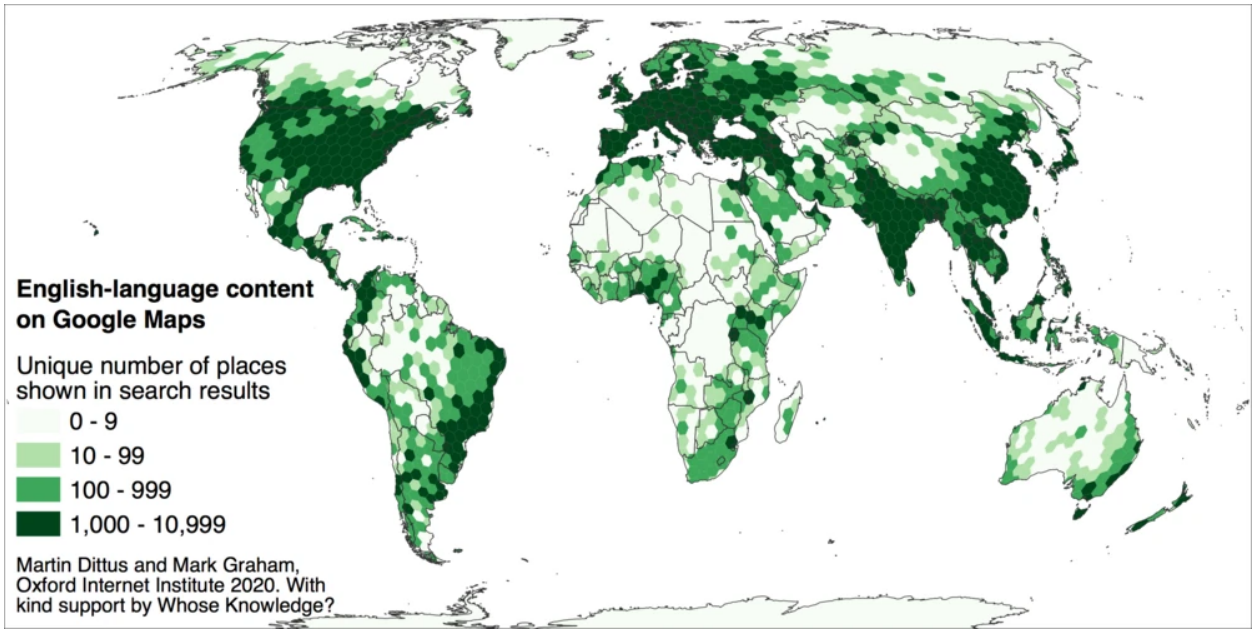
Google Maps

Pouvons-nous accéder à Google Maps dans toutes nos langues ? La langue que nous utilisons change-t-elle la version du monde que nous voyons sur Google Maps ?

Pour répondre à ces questions, Martin et Mark ont collecté des données à propos de la couverture mondiale du contenu de Google Maps dans les 10 langues les plus courantes : anglais, chinois mandarin, hindi, espagnol, français, arabe, bengali, russe, portugais, et indonésien (bahasa Indonesia). Ils ont compilé des dizaines de millions de résultats de recherches individuelles dans ces langues et ont identifié et cartographié environ trois millions d'endroits uniques (établissements et autres lieux).

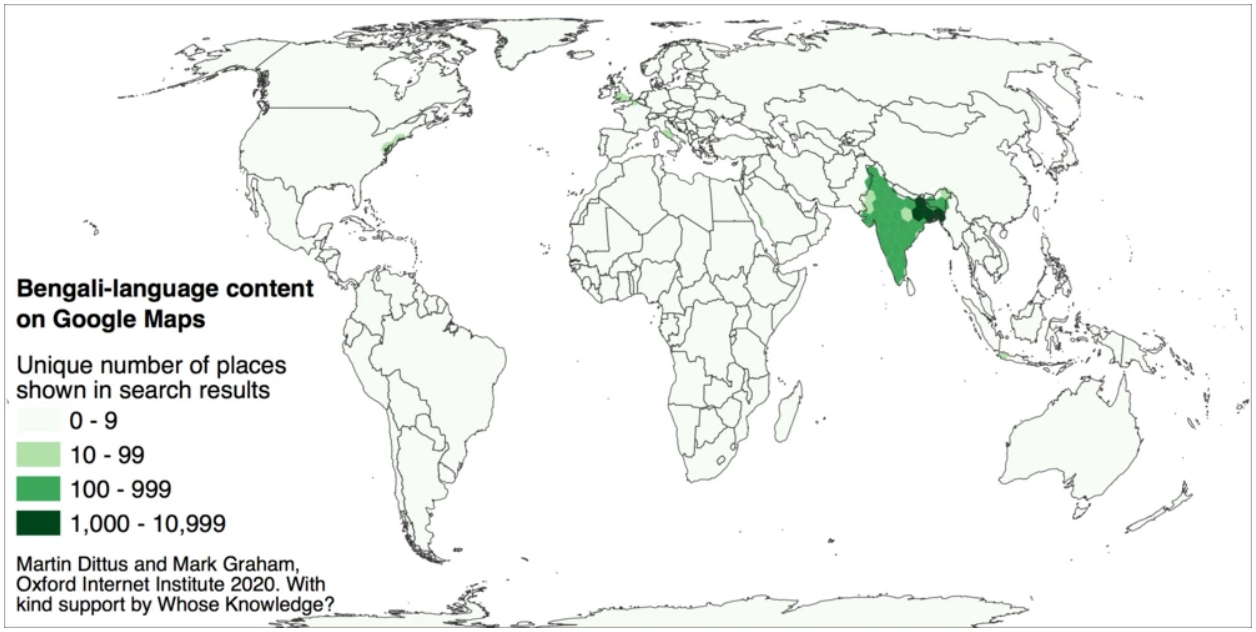
Sans surprise, c'est en accédant à Google Maps en anglais que les cartes présentent le plus de contenu. Les cartes en langue anglaise de Google couvrent le monde entier, néanmoins elles sont beaucoup plus denses (c'est-à-dire qu'elles contiennent plus d'informations) dans les pays du Nord, tout particulièrement en Europe et en Amérique du Nord. Elles couvrent également

plutôt bien l'Asie du Sud et des portions de l'Asie du Sud-Est, ainsi que de grandes zones d'Amérique latine. En comparaison, toutefois, de nombreuses parties d'Afrique contiennent peu de contenu.



La densité d'informations de Google Maps pour les anglophones. Les zones plus sombres indiquent les endroits où les résultats de recherche incluent un plus grand nombre de lieux.

Nous avons pu constater qu'à l'inverse des cartes en anglais relativement bien fournies, celles en bengali (première langue d'Ishan) se cantonnent principalement à l'Asie du Sud, notamment à l'Inde et au Bangladesh. Google Maps ne dispose de peu, voire aucun contenu pour les locuteur·ices du bengali dans la majorité du reste du monde. Afin de découvrir du contenu supplémentaire et de naviguer vers des endroits autres que l'Inde et le Bangladesh, les locuteur·ices du bengali doivent utiliser une langue secondaire, telle que l'anglais. C'est également vrai pour Google Maps en hindi (la troisième langue la plus parlée au monde, après l'anglais et le chinois mandarin).



La densité d'informations de Google Maps pour les locuteur·ices du bengali. Les zones plus sombres indiquent les endroits où les résultats de recherche incluent un plus grand nombre de lieux.

Vous trouverez plus d'informations sur Google Maps en différentes langues dans l'essai détaillé de Martin et Mark.

Wikipédia

Comme l'a démontré l'enquête sur les plateformes de Martin et Mark, Wikipédia se trouve à l'avant-garde de la prise en charge linguistique sur Internet, grâce à une interface traduite par ses utilisateur·rices dans plus de langues que n'importe quelle autre plateforme commerciale étudiée, y compris Google et Facebook.

En termes de contenu concret (les informations et savoirs des articles de Wikipédia), le site compte des éditions en plus de 300 langues. Pourtant les locuteur·ices de ces langues n'ont pas accès au même contenu, ni à la même quantité d'information. Nous avons voulu aller plus loin dans le questionnement : le contenu sur Wikipédia est-il de qualité égale d'une langue à l'autre ? Certaines langues sont-elles mieux représentées que d'autres ? Certaines communautés linguistiques ont-elles accès à plus de contenu que d'autres ? Nous avons répondu à plusieurs de ces questions en détail dans l'analyse de Wikipédia par Martin et Mark.

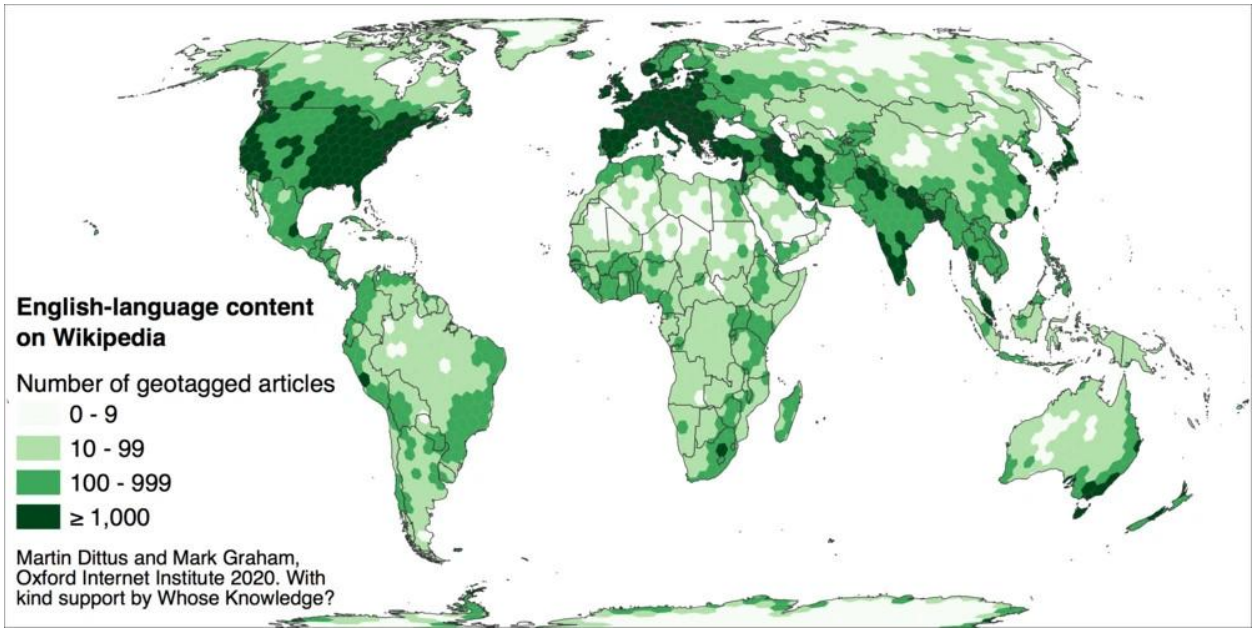
Nous avons utilisé des données de 2018 avec des balises géographiques (une manière d'insérer des références géographiques, comme des coordonnées, dans les articles de Wikipédia) et analysé le nombre d'articles et la croissance du contenu en différentes langues. Nous avons également basé notre analyse sur les langues « locales », c'est-à-dire celles classées comme

langue officielle dans Unicode CLDR (le code qui prend en charge les langues sur Internet), ou qui sont utilisées par au moins 30 % de la population d'un pays.

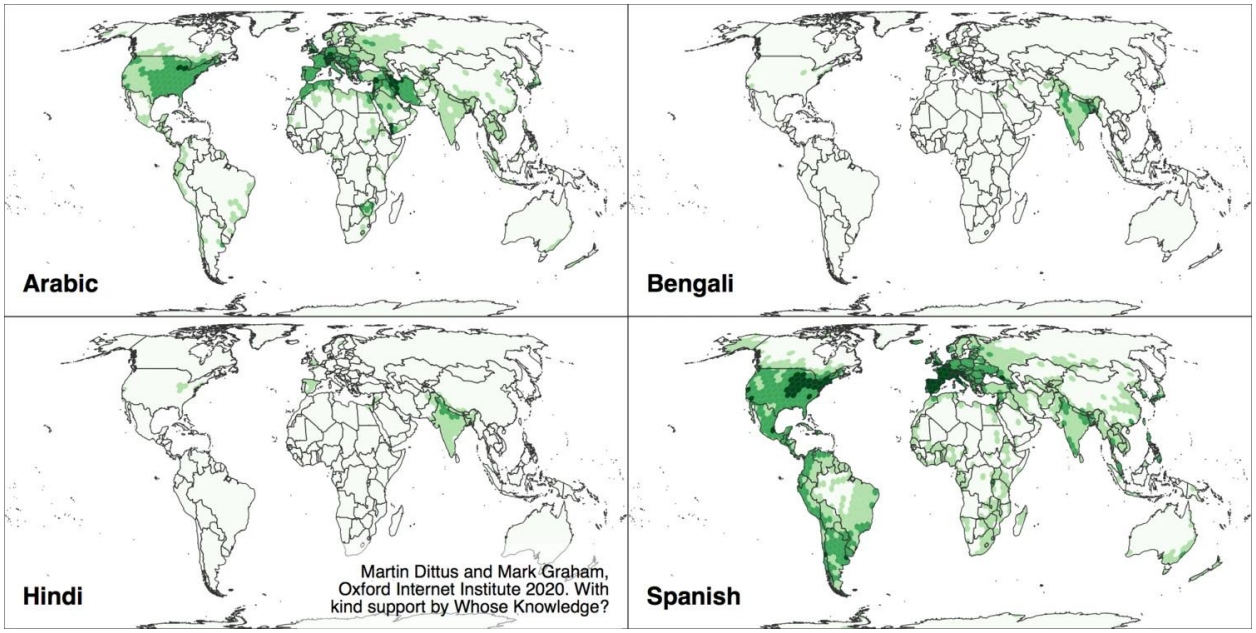
Nous avons ensuite identifié les langues locales les plus prévalentes, c'est-à-dire parlées par le plus grand nombre de personnes dans chaque pays. Nous avons trouvé 73 langues prévalentes dans au moins un pays. L'anglais est la langue la plus couramment parlée et prédomine dans 34 pays. Elle est suivie par l'arabe et l'espagnol (18 pays), le français (13 pays), le portugais (7 pays), l'allemand (4 pays) et le néerlandais (3 pays). Le chinois, l'italien, le malais, le roumain, le grec et le russe sont les langues plus prévalentes dans 2 pays et les 60 langues restantes prédominent dans un seul pays.

Pour comparer la distribution de ces langues locales avec le contenu Wikipédia dans chaque pays, nous avons identifié l'édition de Wikipédia comptant le plus d'articles à propos de ce pays. Nous avons constaté un biais envers le contenu en anglais. L'anglais est la langue dominante sur Wikipédia dans 98 pays, suivie par le français (9 pays), l'allemand (8 pays), l'espagnol (7 pays), le catalan et le russe (4 pays), l'italien et le serbe (3 pays) et le néerlandais, le grec, l'arabe, le serbo-croate, le suédois et le roumain (2 pays). Les 21 langues de Wikipédia restantes sont dominantes dans un seul pays.

Alors que le nombre d'articles dans chaque langue de Wikipédia est dynamique et croît constamment, la variation considérable en taille et échelle (nombre d'articles et de contributeur·ices) entre les éditions linguistiques du site est flagrante. Wikipédia en anglais est de loin la plus grande édition avec plus de 6 millions d'articles et presque 40 millions de contributeur·ices enregistré·es. Les éditions en espagnol, allemand et français comptent chacune entre 4 et 6 millions de contributeur·ices et environ 2 millions d'articles. Les éditions linguistiques restantes sont petites en comparaison : environ 20 langues comptent plus d'un million d'articles et seulement 70 ont plus de 100 000 articles. La plupart des éditions linguistiques de Wikipédia ne contiennent qu'une petite portion du contenu de Wikipédia en anglais.



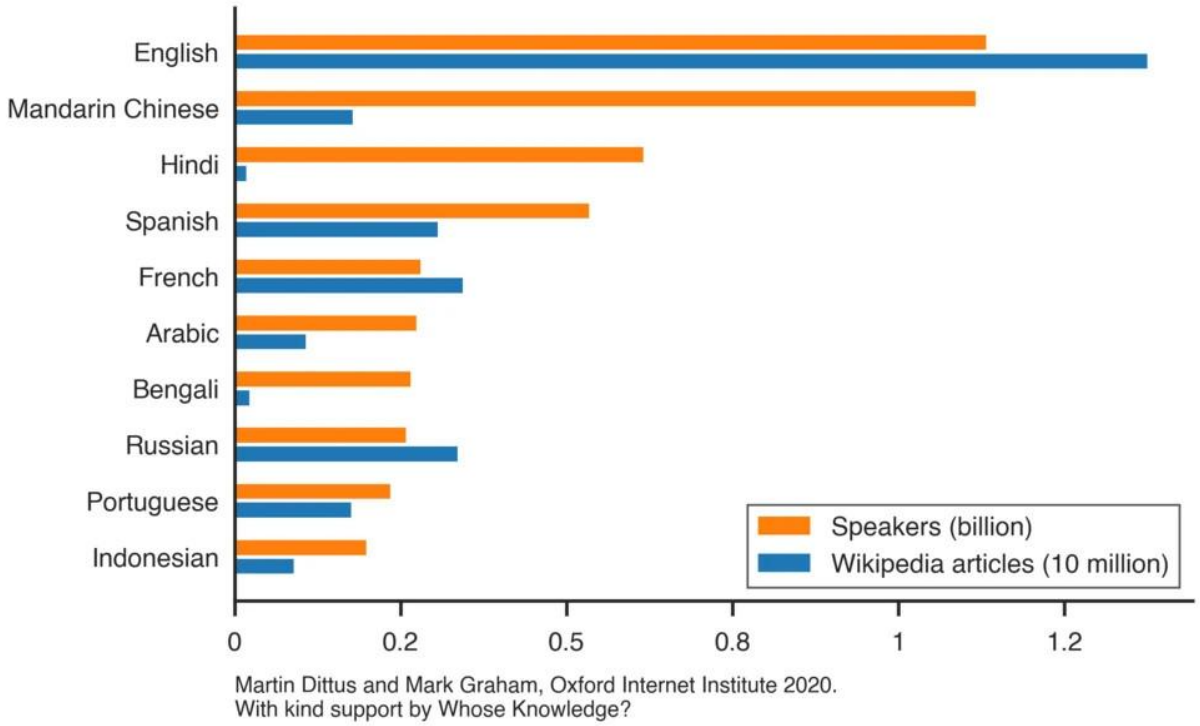
La densité d'informations de Wikipédia en anglais début 2018. Les zones plus sombres indiquent le plus grand nombre d'articles avec balise géographique.



La densité d'informations de Wikipédia en arabe, bengali, hindi et espagnol début 2018. Les zones plus sombres indiquent le plus grand nombre d'articles avec balise géographique.

Ce qui est fascinant, c'est de voir combien le contenu multilingue de Wikipédia rappelle la répartition de Google Maps étudiée plus haut.

Lorsque l'on compare le nombre d'articles dans les différentes langues de Wikipédia au nombre de locuteur·ices de ces langues (comme première ou seconde langue), nous constatons que pour les langues européennes comme l'anglais, le français, l'espagnol, le russe et le portugais, le nombre d'articles de Wikipédia est proportionnel au nombre de locuteur·ices. Mais cette règle ne s'applique pas à d'autres langues très répandues : le chinois mandarin, l'hindi, l'arabe, le bengali et l'indonésien (bahasa Indonesia) sont parlés par des centaines de millions de personnes. Pourtant les éditions de Wikipédia dans ces langues sont beaucoup plus modestes et comptent moins d'articles que les éditions en langues européennes. Il existe plus d'articles sur Wikipédia en français, espagnol ou portugais qu'en mandarin, hindi ou arabe, alors que ces dernières figurent parmi les cinq langues les plus parlées au monde et comptent plus de locuteur·ices que le français et le portugais.



Contenu de Wikipédia et nombre de locuteur·ices des 10 langues les plus courantes au monde. (Estimation de la population : Ethnologue 2019, incluant les locuteur·ices d'une seconde langue.)

L'essai de Martin et Mark contient beaucoup plus d'analyses et de présentations visuelles des données en différentes langues. Ces chiffres confirment les expériences vécues par nos contributeur·ices du monde entier.

Les marginalisations et exclusions des langues qui ne sont pas principalement des langues coloniales européennes sont profondes dans les mondes réel et virtuel, y compris dans le cas de langues mondiales et dominantes comme l'arabe. Afin de rédiger des articles Wikipédia dans

notre propre langue, à l'aide de références basées dans ce contexte linguistique, nous avons besoin de sources publiées fiables et vastes ce qui (comme nous l'avons constaté plus haut) est rare dans la plupart des langues du monde. Emna, contributrice Wikipédia, explique les problèmes qu'elle rencontre à trouver des ressources et des références en différentes langues d'Afrique : « ...les difficultés par exemple pour moi, en tant que contributrice Wikipédia, ne concernent pas que la langue tunisienne, notre dialecte, ou la langue arabe, c'est également dans l'Afrique toute entière que l'on constate un immense manque de ressources et références».

Même en Europe, les locuteur·ices de langues minoritaires ont du mal à utiliser ou modifier Wikipédia dans leur propre langue. Claudia a constaté que de nombreux et nombreuses répondant·es en breton connaissaient l'existence de Wikipédia en breton, et que « 19 % d'entre eux contribuaient en modifiant les articles existants et 8 % en écrivaient de nouveaux ». Pourtant, elle ressent que la plupart des locuteur·ices de langues minoritaires utilisent une langue dominante par facilité : « la disponibilité de services, d'interfaces, d'applications et d'éditions Wikipédia n'implique pas qu'ils soient réellement utilisés. Certaines études révèlent que les locuteur·ices de langues minoritaires passent facilement à leur langue dominante lorsqu'ils utilisent des technologies numériques basées sur le langage, soit parce que les technologies sont bien meilleures, soit parce que la gamme de services disponibles est beaucoup plus large. »

Wikipédia et sa constellation de projets de transmission des savoirs gratuits et open source (où le code est disponible ouvertement et conçu de manière collective) représentent l'un des espaces les plus utiles et prometteurs pour le savoir multilingue en ligne. Par exemple, ses communautés de bénévoles savent et comprennent qu'il n'existe pas une forme unique de l'anglais, de l'arabe ou du chinois, mais exprimer la pluralité des contextes et contenus linguistiques n'est pas toujours facile. Comme nous le voyons dans nos analyses et expériences, Wikipédia aussi souffre des structures de pouvoir et de privilèges héritées du passé et toujours présentes qui biaisent les manières et les formes de création et de partage des savoirs en différentes langues et au sein des familles linguistiques.

Quels sont les moyens d'avancer pour les individus, les organisations et les communautés qui souhaitent un Internet plus multilingue ? Dans les sections suivantes et finale, nous tirons les conclusions de tous les chiffres et histoires que nous avons partagés avec vous jusqu'à maintenant afin de proposer une vue d'ensemble de ce que nous avons appris et des contextes, considérations et actions qui peuvent nous mener vers un Internet vraiment multilingue.

Qu'avons-nous appris sur l'Internet multilingue ?

Nous en avons appris beaucoup sur les langues, sur Internet et sur les langues sur Internet en travaillant sur ce rapport. Voici un aperçu des éléments les plus importants de notre parcours jusqu'ici.

Apprentissage : les langues ne sont pas juste un outil de communication, elles constituent une manière d'accéder au savoir et d'exister dans le monde. C'est pourquoi le multilinguisme est si important : il nous permet de respecter et affirmer toute la richesse et la texture de nos multiples identités et de nos différents mondes.

Contexte : les personnes connaissent leurs mondes et s'expriment dans plus de 7000 langues, qui peuvent être orales (parlées et signées), écrites ou transmises par des sons.

Pourtant, la prise en charge linguistique sur les principales plateformes et applications technologiques ne couvre qu'une fraction de ces 7000 langues, dont seulement environ 500 sont présentes en ligne pour l'information ou les savoirs. Certaines des langues les plus parlées au monde sont rarement prises en charge et incluent peu d'informations en ligne. La prise en charge linguistique la plus riche, les informations les plus complètes sur Internet (y compris sur Google Maps et Wikipédia), et la plupart des sites Web sont disponibles en anglais.

Réflexion : Internet est loin d'être aussi multilingue que nous l'imaginons et le souhaitons.

Analyse : la plupart des personnes doivent utiliser la langue coloniale européenne la plus proche (anglais, espagnol, portugais, français…) ou la langue dominante régionalement (chinois, arabe…) afin d'accéder à Internet. Les structures passées et présentes du pouvoir et des privilèges jouent un rôle déterminant dans la manière dont les langues sont accessibles (ou non) en ligne.

Comment mieux faire ? : contextes et modes d'action pour un Internet multilingue

«La plupart du temps, utiliser une langue minoritaire demande beaucoup de persévérance, de volonté et de résilience, car l'expérience utilisateur est semée d'embûches.»
Claudia Soria

«Pour construire un Internet multilingue, inclusif et représentatif des peuples autochtones, il est essentiel de tenir compte de l'héritage social et de la réalité actuelle de l'oppression coloniale. Un Internet multilingue ne peut pas avoir la représentativité pour unique objectif. Étant donnée l'histoire coloniale, il faut s'efforcer de promouvoir activement des environnements qui renforcent la survie et l'apprentissage des langues autochtones par et pour les peuples autochtones.»
Jeffrey Ansloos and Ashley Caranto Morford

«Les enfants et jeunes mapuches grandissent avec le numérique et Internet, un espace où ces personnes peuvent aller à la rencontre du mapudungun… Les histoires de notre peuple doivent être écrites, et elles doivent être écrites ou parlées en mapudungun… Nos histoires ne sont pas nécessairement celles de personnalités héroïques comme celles mises en avant par les États coloniaux et postcoloniaux. Notre histoire raconte chaque Mapuche qui a survécu à l'adversité et à la violence : les femmes qui ont dû migrer vers la ville pour gagner un salaire, les femmes et les hommes qui ont quitté les villes, mais ont dû y retourner, privés de leurs racines et de leurs foyers d'origine, car leurs "lofs" n'avaient plus de place pour les accueillir. Ce sont les expériences et les souvenirs de chaque Mapuche qui constituent la mémoire collective de notre peuple.»
Kimeltuwe project

Dans cette section de notre synthèse de l'état des lieux des langues d'Internet, nous recueillons les différentes perspectives des contributeur·rices à ce rapport, ainsi que celles de notre rassemblement « décoloniser les langues d'Internet » de 2019, afin de comprendre les différents contextes, enjeux et opportunités en lien avec le multilinguisme dans le monde et en ligne. Quatre questions peuvent nous aider à identifier des pistes pour imaginer et concevoir un Internet beaucoup plus multilingue.

- **À qui appartiennent les ressources et le pouvoir ?**
- **À qui appartiennent les valeurs et les savoirs ?**
- **À qui appartiennent les technologies et les normes ?**
- **À qui appartiennent les créations et les imaginaires ?**

À qui appartiennent les ressources et le pouvoir ?

Contextes

Notre analyse statistique et les expériences vécues par les personnes elles-mêmes montrent que la marginalisation des langues dans les mondes réel et virtuel n'est pas seulement liée au nombre de locuteur·ices.

Les langues autochtones sont parlées par les membres de plus de 6000 nations autochtones dans le monde. Les populations autochtones vivaient dans une grande partie du monde avant que la colonisation et les génocides détruisent et affaiblissent leurs peuples et leurs langues. Les langues dominantes dans des continents comptant une très vaste diversité linguistique, comme l'Asie ou l'Afrique, sont parlées et signées par des millions de personnes, mais ne sont pas bien représentées en ligne, voire totalement absentes. Par exemple, les diasporas présentes dans beaucoup de pays et de continents parlent de nombreuses formes d'arabe, de chinois, d'hindi, de bengali, de pendjabi, de tamoul, d'ourdou, d'indonésien (bahasa Indonesia), de malais, de swahili, de haoussa, etc. Or, même si ces langues occupent une place dominante dans leurs régions d'origine, elles sont incontestablement marginalisées sur Internet.

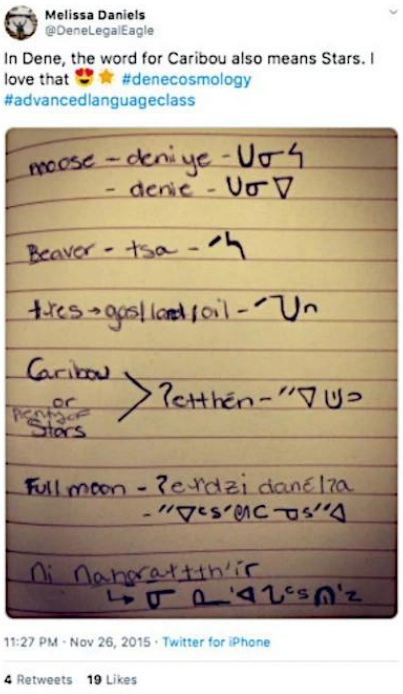
Ces formes de marginalisation et d'exclusion numériques ne sont pas fortuites ; elles sont la conséquence de structures et de dynamiques passées et présentes de pouvoir et de privilèges. Cela veut aussi dire que la quantité de ressources allouées à l'infrastructure linguistique (dans les domaines de l'édition, de la recherche, du secteur public ou des entreprises technologiques) est, dès le départ, biaisée en faveur de certaines régions (Europe, Amérique du Nord) et de certaines langues (l'anglais et d'autres langues d'Europe occidentale). Même en Europe et en Amérique du Nord, les communautés autochtones, noires et toutes les autres communautés marginalisées rencontrent des difficultés à préserver leurs langues de génération en génération.

Les dynamiques de la colonisation et du capitalisme provoquent et renforcent d'autres systèmes de discrimination et d'oppression tels que le racisme, le patriarcat, l'homophobie, le validisme, le classisme et le castisme. Cela signifie que certaines langues (majoritairement des langues coloniales européennes) sont les plus répandues sur Internet, quel que soit leur nombre de locuteur·rices dans le monde. Une autre conséquence : lorsque des informations et du savoir existent dans des langues plus marginalisées, ces contenus sont limités par les personnes qui y ont réellement accès et qui sont en mesure d'en créer ou d'empêcher autrui de produire des informations contraires. On peut souligner, par exemple, l'absence de contenu féministe en cinghalais ou de contenu positif pour les personnes queer et handicapées en bengali et en indonésien (bahasa Indonesia).

Les langues étant au cœur de notre identité, être privé·e de la possibilité d'exprimer toutes les nuances de son individualité dans ses propres langues est une forme de violence. Ces marginalisations sont également violentes d'autres manières. Comme l'explique Uda, « le manque de contenus numériques féministes, pro-droits humains et respectueux rend les espaces où la communication passe principalement par les langues locales hostiles aux femmes, aux personnes queer et aux minorités. Le manque criant de contenus numériques allant à l'encontre du discours dominant, qui est encore pollué par des stéréotypes négatifs, exacerbe les propos haineux et les violences sexistes et sexuelles en ligne ». Ces formes de violence s'appliquent aussi bien aux communautés autochtones, aux personnes subissant des oppressions de caste et aux minorités religieuses, qu'aux personnes handicapées et aux autres communautés marginalisées.

En parallèle, ces communautés utilisent Internet pour riposter contre les différentes formes de violences transgénérationnelles qui s'exercent contre elles et leurs langues. Jeffrey et Ashley ont analysé environ 3800 tweets contenant plus de 35 hashtags et 57 mots-clés en langues autochtones, couvrant 60 groupes de langues autochtones reconnues au niveau fédéral canadien. Leur étude montre que, grâce aux hashtags sur Twitter, les peuples autochtones du Canada et d'ailleurs entrent en contact, participent et collaborent activement à la renaissance et à la prospérité des langues autochtones.

Ils expliquent que « dans un contexte social où les langues autochtones du Canada ont été activement éliminées par les politiques coloniales d'assimilation, les réseaux de hashtags sur Twitter ont créé un environnement unique et précieux où les peuple autochtones peuvent partager du savoir sur leurs langues. Dans les différents réseaux inclus dans notre étude, on retrouve des exemples de réappropriation des langues et de restauration des liens parmi les survivant·es intergénérationnel·les des politiques coloniales assimilationnistes ».



Quand Melissa Daniels, du peuple Déné, a donné son consentement pour citer ce tweet, elle a souhaité rendre hommage à la professeure de langue qui lui a transmis cet enseignement : Eileen Beaver, aînée déné et éducatrice.

Actions

- Prendre conscience des structures et dynamiques de pouvoir et de privilèges dans les différentes institutions et processus qui sous-tendent les infrastructures linguistiques.
- Garantir que les communautés et les langues marginalisées ont accès à des ressources réparatrices, notamment pour l'apprentissage et la programmation en différentes langues.
- Développer des ressources pour créer et diffuser le savoir des communautés à l'intersection de multiples formes d'oppression et de violence, et ce dans les langues et les formats de leur choix.

À qui appartiennent les valeurs et les savoirs ?

Contextes

L'histoire et les technologies d'Internet reposent sur une vision du monde provenant d'épistémologies (manières de savoir et de faire) occidentales. Plus précisément, Internet a été et continue d'être conçu et gouverné majoritairement par des hommes blancs (et maintenant quelques hommes racisés) privilégiés. Ainsi, les valeurs qui se retrouvent le plus souvent au

cœur des architectures et des infrastructures d'Internet sont celles du déterminisme technologique (où la technologie est perçue comme le vecteur principal et positif de n'importe quel changement social) et l'individualisme, qui se focalise non pas sur le collectif, mais sur l'individu comme principal moteur.

D'autre part, cette vision du monde a pour origine le siècle des Lumières, la transition faite au XVIIIe siècle par les pays du Nord vers une forme spécifique de sciences et technologies fondées sur la rationalité. On a souvent tendance à oublier que les mathématiques et la science se sont développées dans les pays du Sud bien avant le XVIIIe siècle. Par exemple, le premier système d'écriture et de numérotation est originaire de Mésopotamie, un territoire qui correspond à l'Iran et l'Irak actuels. Plus important encore, on passe sous silence que les ressources qui ont permis aux Lumières d'être « l'âge d'or » de la science et des technologies dans les pays du Nord provenaient de l'impérialisme exercé sur les pays du Sud, des mouvements massifs de colonisation, d'esclavage, de génocides et d'extraction de ressources en Asie, Afrique, Amérique latine et dans les îles des Caraïbes et du Pacifique. Les bases de la nature extractiviste du capitalisme moderne se retrouvent dans l'histoire de la colonisation, et continuent à faire partie du capital technologique.

En plus des ressources matérielles, ce sont les épistémologies (manières de savoir, de faire et d'être) non occidentales qui ont été détruites, ignorées ou affaiblies par ces processus, comme les savoirs autochtones ou ceux de communautés d'autres zones moins privilégiées. Les langues étant, comme dit précédemment, une manière d'accéder au savoir, elles en ont subi les effets les plus dévastateurs : la dévalorisation totale des langues non occidentales a provoqué une négligence, si ce n'est une destruction active des formes de langue orales et non écrites. Ce biais qui favorise le contenu écrit d'un petit nombre de langues privilégiées alimente le déséquilibre en faveur d'un certain type de « savoir » écrit dans l'édition et la recherche. Ce déséquilibre se répercute ensuite sur la documentation et les données utilisées pour le traitement naturel du langage, ou sur certains systèmes linguistiques automatisés qui constituent l'infrastructure d'Internet, comme Google Traduction.

Comme le décrit Ana, « bien que près de la moitié des langues du monde n'aient pas de système d'écriture et maintiennent une longue tradition orale, les langues qui possèdent un système alphabétique largement reconnu dominent le Web. Le Web renforce une exclusion systématique, où seules les langues écrites peuvent être préservées pour la postérité ».

La disparition des langues nous fait perdre une plus grande partie de notre avenir que nous le pensons. Nous perdons à la fois les manières de nous exprimer dans différentes langues, les visions du monde et les savoirs uniques inhérents à ces langues. À une époque où l'humanité est au bord de l'effondrement planétaire, les communautés autochtones et leurs savoirs conservent notre biodiversité et protègent la vie telle que nous la connaissons. Or, sans

surprise, il existe un lien direct entre la disparition des langues, l'appauvrissement de la biodiversité et la destruction des écosystèmes sur Terre.

Internet pourrait être une infrastructure prometteuse pour la préservation et l'expansion de différents types de langues et de savoirs, car la richesse de ses supports multimédias permet d'imiter et de représenter des langues parlées, signées et allant au-delà de l'écrit. Toutefois, il est crucial que cette promesse radicale d'Internet ne soit pas une fois de plus fondée sur des valeurs coloniales, capitalistes et patriarcales. Comment les communautés peuvent-elles préserver et revitaliser leurs langues, leurs identités et partager leurs savoirs comme elles l'entendent ? Par exemple, dans beaucoup de communautés autochtones, certains savoirs sont sacrés et ne peuvent pas être partagés ouvertement.

Parmi les efforts menés par les communautés, on peut citer Papa Reo, une technologie de reconnaissance vocale pour te reo Māori (la langue māori) d'Aotearoa/de la Nouvelle-Zélande. La communauté māori a créé et maintient la technologie et les données de cette initiative, convaincue que cette forme de souveraineté des données est essentielle pour s'assurer que le savoir transmis par la langue est utilisé par et pour des Māori, plutôt que par des entreprises à la recherche de profit. Il est intéressant de souligner que l'équipe de Papa Reo, tout en reconnaissant l'intérêt des technologies open source, a décidé de ne pas participer à des bases de données ouvertes, car la communauté māori s'est vu refuser les ressources et les privilèges de la plupart des communautés open source. Mukurtu adopte une approche différente avec une plateforme open source créée par des communautés autochtones pour gérer leurs propres données linguistiques.

Actions

- Créer, collaborer et partager des infrastructures linguistiques sur Internet au service de l'intérêt général, avec pour valeurs clés celles du collectif, de la communauté, des concepts féministes et autochtones de souveraineté et incarnation.
- Continuer à remettre en question et à critiquer les infrastructures numériques linguistiques qui sont par nature oppressives, parce que capitalistes, propriétaires, exploitant les humains et détruisant l'environnement.
- Prendre conscience que les technologies linguistiques libres et open source doivent également rester attentives à leur propre privilège relatif et respecter le droit des communautés marginalisées à déterminer et définir par elles-mêmes leur « ouverture » et ce qu'elles veulent partager avec le reste du monde.

À qui appartiennent les technologies et les normes ?

Contextes

Le secteur des industries technologiques n'est pas totalement responsable du manque actuel de représentation et de soutien vis-à-vis de la grande majorité des langues du monde. Cependant, les technologies développées par les pays du Nord sont responsables du maintien et de l'augmentation des inégalités liées à la langue et au colonialisme numérique en ligne.

Les grandes entreprises technologiques, qui conçoivent et développent la plupart des plateformes, outils, matériels et logiciels que nous utilisons, peuvent ignorer le besoin de créer un Internet véritablement multilingue, parce qu'elles ne considèrent pas la majorité de nos 7000 langues parlées comme un élément essentiel de l'infrastructure d'Internet. Après tout, elles savent qu'elles peuvent se contenter de fournir une prise en charge linguistique uniquement pour servir leurs intérêts commerciaux : soit pour les langues coloniales européennes, soit pour les langues de ce qu'elles appellent les « marchés émergents ». Ainsi, les langues dominantes des pays d'Asie du Sud et du Sud-Est, qui deviennent petit à petit la clientèle la plus importante de ces entreprises, commencent à bénéficier d'une meilleure prise en charge linguistique de leur part.

En parallèle, certaines des technologies linguistiques les plus répandues sur Internet sont créées et contrôlées par ces entreprises qui disposent des ressources et des compétences nécessaires. Wikipédia est une exception remarquable parce qu'elle est open source et soutenue par ses communautés de bénévoles du monde entier. En général, les projets axés sur la recherche de bénéfices et le développement d'outils propriétaires ne mènent pas à des outils et technologies adaptés à du contenu riche, nuancé, dans différentes langues marginalisées et créés par les communautés concernées. Pire encore, les technologies linguistiques actuellement développées par ces entreprises sont des systèmes automatisés à grande échelle qui s'appuient sur des quantités énormes de données provenant de toutes sortes de sources, même si elles contiennent des propos violents, haineux ou dirigés contre les groupes marginalisés.

Comme le décrivent Jeffrey et Ashley : « dans la plupart des réseaux de survivance et d'apprentissage des langues autochtones, le racisme représente un défi social majeur de l'écosystème de Twitter. En particulier, ces réseaux sont activement visés de différentes manières par des propos et des contenus multimédia incendiaires, voire haineux, et par divers types d'utilisateur·rices, dont des vraies personnes et des robots automatisés. Les comptes gérés par des robots semblent suivre les mêmes modèles que la dissémination de la mésinformation, et propagent généralement des contenus incompréhensibles issus d'analyses agrégées et de génération automatique de contenu. »

Si une entreprise qui manque d'expertise dans les langues et contextes locaux ne prend pas ces enjeux assez au sérieux, cela peut mener à des dommages incommensurables et à des actes de violence. Au Myanmar, où Facebook (aujourd'hui Meta) constitue pratiquement l'intégralité d'Internet, des militant·es ont alerté l'entreprise sur les contenus haineux pendant des années avant qu'une équipe en langue birmane ne soit enfin créée. En 2015, Facebook avait quatre modérateur·rices birmanophones pour 7,3 millions d'utilisateur·rices actifs au Myanmar. Les conséquences de ce manque de prise en compte de la langue et du contexte ? Les Nations Unies ont conclu que Facebook avait joué un rôle dans le génocide contre les musulman·es Rohingya du Myanmar, et l'entreprise se trouve au cœur d'une affaire judiciaire contre le gouvernement du Myanmar devant la Cour internationale de Justice.

De la même manière, en Inde, les propos haineux contre les musulman·es, les dalits et d'autres communautés marginalisées perdurent en raison de la très faible modération active, alors que ce pays représente le marché le plus important de Facebook et qu'on y trouve certaines des langues les plus parlées au monde. Ainsi, l'entreprise consacre 84 % de ses ressources dédiées à la prise en compte des langues à la mésinformation aux États-Unis, pays où se trouvent moins de 10 % de ses utilisateur·rices. Les 16 % de ressources restantes sont alloués au reste du monde.

Les entreprises du numérique doivent prendre conscience que le développement de technologies linguistiques nécessite d'y investir des ressources importantes, de prendre en compte le contexte sociopolitique et de s'engager pour offrir une expérience numérique multilingue sécurisée et accueillante.

Nos contributeur·rices évoquent la diversité des besoins, défis et opportunités lorsqu'on souhaite créer une telle expérience. Le manque d'infrastructures (de l'accès à Internet à l'efficacité des appareils) et de technologies adaptées à toutes les langues rend l'utilisation de langues marginalisées en ligne pénible, difficile, lente et peu pratique. Nous présentons ici quelques exemples frappants.

Les technologies linguistiques sont rarement conçues pour une langue marginalisée.

Donald raconte à quel point il est difficile pour sa communauté au Malawi d'avoir accès à Internet et à des appareils pour communiquer facilement dans leurs langues. Il s'est entretenu avec 20 locuteur·rices du chindali : 10 étudiant·es et 10 aîné·es de la communauté. Parmi ces 20 personnes, seulement 5 avaient des smartphones ou des téléphones mobiles, et 7 ne possédaient aucun appareil. Seulement 4 personnes (tou·tes des étudiant·es) possédaient un ordinateur portable. En matière d'accès à Internet, seul·es les étudiant·es disposaient de l'abonnement de leur université ou d'un abonnement personnel.

Une fois sur Internet, la plupart des gens ne disposent pas de claviers dans leurs propres langues. La majorité des communautés doit bricoler avec un clavier conçu principalement pour les langues européennes, sur lequel elles collent les caractères de leurs propres langues. L'opération, déjà ardue pour le clavier d'un ordinateur, est impossible pour le petit clavier d'un téléphone. La difficulté est encore plus grande dans les langues principalement orales qui n'ont pas de système d'écriture communément accepté.

Ainsi, Ana, qui parle une langue zapotèque, explique que « les claviers n'ont pas les symboles qui permettent de retranscrire correctement les sons et les tons de nos langues. Depuis des années, on s'efforce de proposer une forme écrite des langues autochtones, en essayant d'atteindre un consensus autour d'un format standard tel que le l'alphabet latin, un format plus ou moins imposé et orienté par les pays occidentaux, mais d'une certaine manière accepté et requis par une partie des locuteur·rices ». Comme l'illustrent Joel et Caddie à travers le projet Indigemoji, cette démarche est encore plus difficile pour les langues comme l'arrernte, qui allient le geste à la parole.

Si vous faites partie d'une communauté qui se sent déjà en danger dans le monde physique, et que vous ne pouvez pas accéder à Internet dans votre propre langue, il y a peu de chances que vous vous sentiez à l'aise pour produire et rendre visible du contenu pertinent et critique pour et avec votre communauté. Paska souligne que « de nombreuses personnes LGBTQIA+ en Indonésie ne maîtrisent pas bien les aspects techniques d'un site Web et manquent de connaissances sur le fonctionnement d'un moteur de recherche ».

Les communautés marginalisées de régions privilégiées des pays du Nord sont également confrontées à la difficulté de ne pas disposer de contenus importants voire essentiels dans leurs propres langues. Jeffrey et Ashley expliquent : « l'un des obstacles principaux… réside dans les nombreuses limites des technologies de traduction actuelles pour traduire les langues autochtones dans le contexte canadien. Les technologies de traduction de Twitter prennent les langues autochtones comme le hul'qumi'num, le sḵwx̱wú7mesh (le squamish), le lewkungen et le neheyawewin (le cri) pour de l'allemand, de l'estonien, du finlandais, du vietnamien et du français ».

Globalement, confirme Uda, « il reste difficile de créer du contenu numérique dans des langues locales en raison de la rareté des outils et de la difficulté d'utilisation de ceux qui existent. La création de contenus dans une langue locale nécessite des outils et des compétences spécifiques. Ces obstacles contribuent au manque de contenus progressifs dans les langues locales ».

Les technologies linguistiques sont le plus souvent conçues de manière descendante, en faisant passer le profit avant l'équité et la sécurité. Claudia dresse un portrait très clair de l'approche de la plupart des entreprises du numérique : « la tech et les médias sont distribués

de manière descendante par les grandes entreprises, en impliquant à peine, voire pas du tout, les communautés de locuteur·rices. À cela s'ajoute une tendance paternaliste : puisqu'il existe très peu de choses, tout ce qui est fourni doit forcément être apprécié et accepté. Très souvent, les entreprises mettent en place des solutions toutes faites sans prendre en compte les réels besoins, désirs et attentes des personnes parlant des langues minoritaires. C'est comme si l'on considérait que ces personnes doivent être reconnaissantes de tous les produits ou opportunités qui leur sont donnés, quels que soient leur intérêt ou leur pertinence pour elles. Il existe des exceptions notables, telles que van Esch et al. (2019) qui insistent fortement sur le besoin de concevoir les applications de traitement du langage naturel en collaboration étroite avec des locuteur·rices de la langue ».

L'approche descendante prend rarement en compte les contextes de vie des communautés locutrices de langues marginalisées, ou les modifications de conception nécessaires pour faire exister leurs langues en ligne de manière riche et nuancée. Quand Emna a interviewé le Soudanais Gamil, ce dernier a répondu en arabe soudanais : « le Soudan est un pays avec beaucoup de tribus et de nombreuses traditions et coutumes. Donc le Nord du pays parle le dialecte arabe (celui que j'utilise actuellement) et c'est bien sûr à cause de la colonisation. En revanche, à l'Est et à l'Ouest, les tribus parlent une langue locale différente qu'elles sont les seules à pouvoir parler et comprendre. C'est très rare de trouver une personne du Nord qui la comprenne à moins qu'elle n'ait habité dans ces régions et interagi avec ces communautés. Notre langue est issue de la famille des langues couchitiques. Nos références sont les civilisations couchitiques et nubiennes. Quand le Haut Barrage s'est rempli, nous avons perdu notre identité couchitique et nous n'avons pas trouvé de dictionnaire pour décrypter la langue et la traduire. La langue nubienne, en revanche, est connue et traduite. C'est même l'une des langues disponibles dans les paramètres de base des téléphones Huawei. Les langues de l'Est et de l'Ouest ne sont pas écrites (ou peut-être qu'elles le sont, je ne sais pas, je dois vérifier). Au Nord, on parle arabe. Nous sommes des Africains qui parlons arabe, pas des Arabes ».

S'appuyant sur ses entretiens, Emna affirme : « le Web a besoin de tous·tes ses utilisateur·rices, ceux qui écrivent et ceux qui n'écrivent pas. Cependant, le changement n'est pas uniquement de la responsabilité des utilisateur·rices. Les entreprises qui conçoivent et développent des logiciels doivent également assumer leur part de responsabilité dans la conception de l'Internet du futur. En ce moment, on a l'impression que tout le monde parle d'inclusivité, de lutte contre le racisme, les discriminations et les autres formes de colonialisme. Cependant, pour transformer le Web, le secteur privé doit s'interroger sur la manière dont il perpétue l'exclusion. Les designers Web, ingénieur·es de technologies numériques et dirigeant·es d'entreprises de la tech doivent elles et eux aussi contribuer à rendre leurs logiciels accessibles à tous·tes les utilisateur·rices ».

Une des manières d'analyser ces différentes formes d'exclusion de manière critique est de reconnaître que le fondement même du numérique, le code, est majoritairement en anglais. Il

existe très peu de langages de programmation qui s'appuient sur d'autres langues, ce qui veut dire que les développeur·euses sont obligé·es d'acquérir une connaissance de l'anglais assez poussée pour pouvoir coder. Le langage de programmation Qalb, qui s'appuie sur la syntaxe et la calligraphie arabes, est l'un des rares contre-exemples à cette tendance. Toutefois, en général, le secteur du numérique doit comprendre comment le privilège linguistique est encodé dans chaque technologie et dans la plupart des spécialistes des technologies.

Pour reprendre les propos de Jeffrey et Ashley : « notre étude montre que la promotion des langues autochtones sur Internet doit s'appuyer sur une analyse critique des technologies mêmes d'Internet et des processus sociaux à l'œuvre dans leur utilisation ».

Actions

- Prendre conscience que la possibilité pour les locuteur·rices d'une langue de concevoir et de créer des technologies et contenus numériques multilingues est un droit humain fondamental qui doit figurer parmi les priorités des entreprises du numérique et des organismes de standards, et être soutenu par des institutions mondiales telles que l'UNESCO.
- Construire un modèle de gouvernance internationale des infrastructures linguistiques piloté par les communautés et qui collabore dans la confiance et le respect avec les entreprises du numérique et les autres institutions.
- Placer l'éthique et le consentement des communautés au centre de la création de données et d'outils liés au langage, dans le but d'assurer la maîtrise et la sécurité des contenus et méthodes de partage, en particulier pour les communautés qui subissent une marginalisation encore plus importante du fait de l'imbrication de différents systèmes d'oppression et de discrimination.

À qui appartiennent les créations et les imaginaires ?

Les chiffres et les histoires que nous avons partagés montrent que des infrastructures linguistiques utiles et efficaces sont possibles uniquement à condition que les besoins, conceptions et imaginaires de la communauté de locuteur·rices elle-même soient placés au centre des réflexions. Les langues marginalisées ne prospèreront et ne se diffuseront que quand la majorité minorisée du monde sera incluse dans le développement des technologies.

Nos contributeur·rices suggèrent différentes manières d'y arriver : par exemple, par l'embauche dans les entreprises du numérique de spécialistes des technologies et d'autres spécialistes venant de communautés de langues marginalisées, et le recours responsable aux communautés dans ces travaux. Elles et ils recommandent de mettre en place des structures et des processus adaptés aux contextes linguistiques, plutôt que d'adopter une seule approche prescriptive. Comme le dit Claudia : « les personnes qui parlent des langues minoritaires n'ont pas besoin de

solutions toutes faites : il faut écouter leurs besoins et contraintes précises et les intégrer dans des produits adaptés à ces besoins. Les contextes sociolinguistiques des langues minoritaires peuvent être très variés ; les solutions proposées doivent l'être également. »

Nos contributeur·rices soulignent également le rôle de leurs propres communautés dans la préservation et la diffusion de leurs langues en utilisant les outils à leur disposition. Ishan explique : « en tant que personne queer et handicapée, je pense que nous devons utiliser les ressources sociales, culturelles et économiques déjà à notre disposition pour faire entendre nos expériences, ambitions et exigences. Nous, utilisateur·rices d'Internet, devons prendre la responsabilité de rendre Internet inclusif et accessible. »

Ana explique que ses communautés utilisent plus les plateformes qui ont de meilleures infrastructures orales et visuelles que celles qui reposent sur du texte. « Aujourd'hui, dans les langues zapotèques, l'oral est plus utilisé en ligne que l'écrit. Des réseaux sociaux tels que YouTube, Facebook, WhatsApp et Instagram sont des outils accessibles et faciles à prendre en main pour les langues orales. Ces plateformes permettent aux utilisateur·ices de mettre en ligne du contenu visuel qui enrichit leur message de la manière dont elles et ils le souhaitent, sans se plier aux exigences de l'écrit. C'est pour cela que ces plateformes ont le plus d'utilisateur·rices dans le monde et sont utilisées par les communautés autochtones. Dans les communautés zapotèques de la Sierra, la plupart des personnes vont sur Internet via Facebook, où elles diffusent des festivals traditionnels, des danses, des musiques, des récits d'événements importants et des annonces pour les communautés locales et émigrées. Avant la COVID-19, Facebook jouait déjà un rôle important dans le processus de deuil pour les familles émigrées, grâce à la diffusion des enterrements et des rituels sur la plateforme. Facebook est également utilisé par certaines communautés zapotèques pour retransmettre des émissions de radio. Cela crée un pont précieux entre la radio, qui est un moyen de communication analogique répandu dans les zones rurales, et l'espace universel et omniprésent qu'est Internet. »

Pour les personnes qui peuvent utiliser des formes de langage écrites, les hashtags sont une belle opportunité d'entrer en contact avec sa communauté sur Internet afin d'apprendre, de diffuser leurs langues et d'inspirer d'autres communautés à faire de même. Comme le décrivent Jeffrey et Ashley : « dans un contexte social où les langues autochtones du Canada ont été activement éliminées par les politiques coloniales d'assimilation, les réseaux de hashtags Twitter ont créé pour ces peuples un environnement unique et précieux de partage de savoirs sur leurs langues… Par exemple, un réseau de hashtags créé par les personnes apprenant la langue gwich'in a poussé celles apprenant l'anishnaabemowin (l'ojibwé) à faire de même. De la même manière, un réseau linguistique autour de la langue neheyawewin (le cri) sur Twitter a donné envie aux personnes apprenant le hul'qumi'num de lancer leur propre « mot du jour » sur le réseau social. »

Les usages variés et créatifs que font les communautés marginalisées de ces plateformes propriétaires sont une raison importante pour que les entreprises du numérique travaillent avec elles et non contre elles.

D'un autre côté, Claudia souligne que les militant·es de communautés de langues marginalisées doivent être mieux informé·es et coordonné·es afin de ne pas gâcher leur énergie et leurs ressources. « Bien qu'admirables, ces initiatives [militantes] pâtissent souvent d'un manque de coordination, de planification et d'une faible visibilité. Cela entraîne un problème grave pour les communautés aux ressources limitées : le dédoublement des initiatives. Dans les deux cas, toute la difficulté réside dans la connaissance limitée de ce qui est déjà disponible et des besoins auxquels répondre. Afin de décoloniser les technologies linguistiques pour les langues minoritaires, il est primordial d'avoir une image plus claire de l'étendue de l'utilisation des langues minoritaires en ligne, de sa fréquence et de ses objectifs. Il est tout aussi important de connaître les obstacles que les locuteur·rices de langues minoritaires rencontrent lorsqu'ils et elles essayent d'utiliser ces langues : quelles sont les difficultés techniques ? Sont-ils et elles bloqué·es par une forme d'autoparanoïa ? Écrire dans une langue minoritaire implique de s'exposer aux yeux du monde : est-ce que des personnes se censurent par peur d'être moquées ou stigmatisées ? De même, on en sait peu sur ce que les personnes parlant des langues minoritaires attendent des opportunités offertes par le numérique : qu'aimeraient-elles avoir à leur disposition ? »

Ce rapport représente une tentative de compréhension de ces enjeux et propose des pistes de solutions. Répéter les mêmes choses constamment pour différentes langues ne fonctionnera pas. Par exemple, il est problématique de développer une application en anglais et de supposer qu'elle fonctionnera à peu près aussi bien en indonésien (bahasa Indonesia). Améliorer Internet implique de transformer les dynamiques de pouvoir entre les personnes, pas simplement de résoudre des problèmes techniques. Le multilinguisme sur Internet est un ensemble de problématiques sociales, technologiques et politiques complexes. Nous devons prioriser les besoins des communautés linguistiques, et non les technologies d'Internet, afin de rendre les technologies linguistiques plus efficaces et utiles.

Par-dessus tout, ce sont les créations et imaginaires de ce que nous appelons la « majorité minorisée » du monde qui vont permettre d'améliorer nos infrastructures linguistiques.

« [Indigemoji] émerge dans une période cruciale d'accélération de l'adoption du numérique et d'une meilleure connectivité en Australie centrale. Le projet invite la population locale à imaginer ce qu'elle pourrait faire avec ces nouvelles plateformes. Comment éviter que celles-ci ne soient qu'une arme colonisatrice de plus ? Et comment peut-on y intégrer nos langues et cultures, pour nous les approprier ? »

Actions

- Promouvoir des technologies linguistiques centrées sur les contextes, besoins, créations et imaginaires des communautés linguistiques locales et connectées à l'échelle internationale, plutôt qu'un numérique qui a recours à une approche universelle.
- Utiliser tous les outils d'Internet de manière créative pour explorer l'éventail des manières dont les langues sont matérialisées (oralement, visuellement, par des gestes, par du texte...).
  pour que différentes formes de savoir puissent être exprimées et partagées facilement en étant accessibles.
- Se tourner vers les nations autochtones pour apprendre à concevoir des technologies linguistiques qui respectent les souvenirs collectifs et communautaires tout en préparant le futur. Marchons vers l'avenir à reculons.

Enfin, que pouvez-vous faire ?

Si une personne ne parle pas aussi bien français que vous, cela ne veut pas dire qu'elle est stupide. Cela signifie qu'elle maîtrise mieux une des 7000 autres langues du monde.

Nous devons tous·tes, avec nos nombreuses compétences et expériences, travailler main dans la main pour créer et développer un Internet réellement multilingue. Nous devons aussi faire en sorte que les informations et les savoirs que nous partageons dans ces langues ne fassent de mal à personne, et qu'au contraire ils contribuent au bien commun. Nous avons besoin de ce que nous appelons « la solidarité en action ».

Si vous travaillez dans le secteur du numérique :

- Prenez conscience que les politiques de votre entreprise peuvent contribuer (ou non) au multilinguisme d'Internet et à l'approfondissement du savoir humain commun.
- Mettez l'internationalisation et la localisation de vos outils dans des langues (marginalisées) au centre de vos stratégies, au lieu de les traiter comme des sujets secondaires. Faites-le en partenariat avec les communautés, plutôt qu'avec une approche descendante et hors contexte.
- Acceptez les critiques étayées par la recherche des grands modèles de langues et technologies linguistiques automatisées ainsi que les dommages importants qu'ils peuvent causer s'ils ne sont pas attentivement supervisés par des humains.
- Intégrez dans tout travail linguistique des processus humains et attentifs de contextualisation, modération et d'édition, en utilisant de plus petits ensembles de données gouvernés par les communautés.
- Travaillez dans le respect des communautés, surtout les plus marginalisées et celles risquant le plus de souffrir si on ne leur prête pas assez attention.

- Citez les personnes des communautés de locuteur·rices qui vous font profiter de leur temps et de leur expertise.

Si vous êtes un organisme de standards technologiques :

- Prenez conscience que les standards linguistiques doivent prendre en compte le contexte.
- Construisez des liens plus forts et de meilleurs processus avec les communautés de langues marginalisées afin que plus de standards puissent être établis en partenariat avec des communautés, voire pilotés par elles.
- Invitez activement plus de membres de communautés de langues marginalisées dans la gouvernance des standards, et donnez-leur les ressources nécessaires pour participer pleinement.

Si vous travaillez pour le secteur public :

- Prenez conscience que le contenu dans les langues de vos citoyen·nes doit être accessible à tout le monde, et pas seulement à une minorité privilégiée.
- Soutenez l'expansion de contenus dans les langues par et pour les personnes marginalisées ou victimes de discriminations dans vos régions.
- Soutenez la préservation et la numérisation des langues marginalisées dans vos régions, et pas seulement celles des langues dominantes.

Si vous travaillez dans le milieu des technologies libres et open source et de la connaissance ouverte :

- Prenez conscience que les technologies libres et open source ne sont pas exemptes d'inégalités de pouvoir et limitations, même si elles entendent agir pour le bien commun.
- Respectez les limites imposées par les communautés marginalisées dans leur manière de partager leurs savoirs, en réaction à l'exploitation et la marchandisation du passé.
- Travaillez avec les communautés de langues marginalisées pour créer les technologies et les savoirs dont elles ont besoin, plutôt que ceux dont vous croyez qu'elles ont besoin.

Si vous êtes dans le milieu « GLAM » (galeries, bibliothèques, archives, musées et institutions de mémoire) :

- Prenez conscience que les langues sont au cœur des savoirs et des cultures que vous préservez, exposez et mettez en valeur.

- Travaillez avec les communautés de langues marginalisées pour que leurs histoires et leurs langues soient entendues, reconnues et amplifiées comme elles le souhaitent, par la signalisation de la provenance (ou la propriété et la localisation de ressources). Respectez notamment les droits des communautés marginalisées à ne pas partager publiquement certains savoirs et ressources. Ce point est crucial, car beaucoup d'institutions GLAM, en particulier dans les pays du Nord, sont issues d'histoires complexes de colonisation et de capitalisme.
- Assurez-vous que les ressources linguistiques qui sont dans vos collections sont librement et facilement accessibles aux communautés marginalisées et à leurs allié·es, afin que nous puissions construire ensemble une infrastructure linguistique collective.

Si vous travaillez dans le milieu de l'éducation :

- Prenez conscience du biais de notre éducation en faveur des sources textuelles et de certaines langues.
- Diversifiez vos techniques d'enseignement et d'apprentissage en y incluant plusieurs langues, ainsi que les différentes formes de langage et de savoir qu'elles incarnent.
- Lisez, écoutez et citez des œuvres traduites quand vous en avez la possibilité, et encouragez les autres à faire de même.

Si vous travaillez dans le milieu de l'édition :

- Prenez conscience que la plus grande partie monde de l'édition favorise actuellement les langues coloniales européennes.
- Élargissez le nombre de langues dans lesquelles vous publiez des œuvres, et numérisez les contenus dans toutes ces langues.
- Publiez plus de livres et de contenus multilingues.
- Expérimentez des formes multimodales de publication, afin que différentes formes de langage oral, visuel et textuel puissent être partagées simultanément plus facilement.
- Respectez et reconnaissez le travail de vos traducteur·rices.

Si vous travaillez dans le milieu de la philanthropie :

- Prenez conscience que les langues sont au cœur de l'expertise, de l'expérience et des savoirs humains, quelle que soit la cause que vous financez.
- Prévoyez une interprétation multilingue dans tous les événements et rassemblements internationaux et régionaux que vous organisez et soutenez.
- Soutenez la production, la préservation et la numérisation de contenus dans les langues des communautés que vous servez, et assurez-vous que vos propres contenus sont disponibles dans les langues de ces communautés.

Si vous êtes dans une communauté linguistique marginalisée :

- Prenez conscience que vous n'êtes pas seul·e.
- Sachez que votre communauté a le droit de décider quels savoirs partager avec le reste du monde, et de quelle manière.
- Travaillez avec les aîné·es, les chercheur·euses, les jeunes générations de votre communauté et des ami·es d'autres communautés, pour collecter et partager ces savoirs.
- Si vous souhaitez entrer en contact avec d'autres personnes faisant un travail similaire au vôtre, contactez-nous !

Si vous adorez tout simplement les langues et vous demandez quoi faire :

- Prenez conscience que les langues sont au cœur de ce que nous sommes, de ce que nous faisons, et sont indispensables pour de nombreux savoirs et cultures, dont les vôtres !
- Discutez avec votre famille, vos ami·es et vos communautés pour vous rendre compte que l'anglais et quelques autres langues dominent l'accès à Internet et les contenus en ligne, comprendre pourquoi et trouver des moyens d'y remédier ensemble.
- Faites la démarche de chercher, lire, écouter et partager les contributions des communautés linguistiques marginalisées (y compris ce rapport !)
- Si vous souhaitez recevoir des nouvelles de notre initiative, suivez-nous sur les réseaux sociaux !

Gratitude

Nous témoignons tout notre amour, notre respect et notre solidarité envers les nombreuses communautés marginalisées de par le monde (autochtones et autres) qui considèrent que les langues sont au cœur de leurs identités et de leurs manières d'être. Leurs efforts pour préserver, faire vivre et diffuser ces langues et ces formes d'expression de manière concrète nous poussent à imaginer un (des) Internet(s) plus multilingues et pluriels, où exprimer toutes nos facettes de la manière la plus riche et authentique qui soit. Nous sommes également très reconnaissantes envers les chercheur·euses dans les institutions et les communautés et spécialistes des technologies qui aiment les langues autant que nous, et travaillent d'arrache-pied tous les jours pour rendre Internet aussi multilingue que nos environnements physiques.

À nos nombreux·ses contributeur·rices, traducteur·rices, et communautés à travers le monde (en particulier celles qui ont participé à notre conversation « Décoloniser les langues d'Internet » en 2019) : merci pour tout ce que vous faites et êtes dans le monde. Merci pour votre patience pendant que nous tentions de survivre à ces deux dernières années ! Un merci tout particulier à notre illustratrice qui a créé des visuels originaux pour les essais et à notre animateur qui nous a fait le cadeau de l'animation pour ces illustrations.

Une gratitude sans bornes envers tous·tes nos ami·es et notre communauté qui ont révisé notre travail sous différents points de vue et dans différentes langues. Nous sommes responsables de toutes les erreurs restantes, mais votre soutien et votre solidarité nous ont énormément aidé·es à améliorer ce travail en cours. Et enfin, aux membres de notre équipe et à nos familles de sang et de cœur : nous n'aurions pas pu survivre à ces dernières années (en particulier 2019, 2020 et 2021) sans votre présence, même quand elle était seulement virtuelle. L'amour et la confiance sont la meilleure langue qui existe.

Définitions

Il existe de nombreuses manières de définir les différents aspects des langues et des histoires dont nous avons parlé. Ces définitions ne s'accordent pas toutes entre elles ! Dans ce rapport, nous avons utilisé certains termes et expressions de manière spécifique. Voici les définitions que nous donnons à ces mots et expressions clés.

- **Langues dominantes** : langues parlées par la majorité de la population d'une zone donnée, ou qui dominent par des formes spécifiques de pouvoir et de validation historiques, par des dynamiques juridiques, politiques ou culturelles. Par exemple, l'hindi, une famille de langues ou de « dialectes » selon certaines personnes, est une langue dominante en Asie du Sud par rapport à beaucoup d'autres langues. De la même manière, le chinois mandarin est une langue dominante en Chine, de par la politique gouvernementale, par rapport à d'autres formes de chinois et d'autres langues autochtones de cette région. Certaines langues dominantes sont aussi les langues « officielles » ou « nationales » d'une région ou d'un pays.

- **Langues coloniales européennes** : langues d'Europe occidentale ayant été diffusées en Afrique, Asie, dans les Amériques, dans les îles des Caraïbes et du Pacifique en raison de la colonisation menée par les entreprises et gouvernements d'Europe occidentale, à partir du XVIe siècle. Elles désignent l'anglais, l'espagnol, le français, le portugais, le néerlandais et l'allemand. Il est important de noter que ces langues ont été « colonisatrices » pour les peuples autochtones d'Amérique du Nord, et pas seulement d'Amérique latine (Amérique centrale et Amérique du Sud).

- **Les pays du Sud et les pays du Nord** : le terme « pays du Sud » désigne les régions d'Afrique, d'Asie, d'Amérique latine et des îles des Caraïbes et du Pacifique colonisées par les pays d'Europe occidentale. Ce n'est pas un terme géographique : il vise à mettre en lumière les conditions socio-économiques et politiques, historiques et actuelles, qui caractérisent ces pays et régions, et à les distinguer des pays privilégiés d'Europe et d'Amérique du Nord, les « pays du Nord ». Ce terme a été créé et diffusé par des

chercheur·euses et des militant·es des pays du Sud pour dépasser les expressions « les moins développés », « en voie de développement » ou « Tiers-Monde » qu'elles et ils considéraient comme péjoratifs et importuns. La colonisation a entraîné le génocide ou la quasi-destruction de beaucoup de nations autochtones dans les pays du Nord. Certaines personnes et communautés des pays du Sud ont participé à la colonisation de leurs propres peuples et en ont tiré avantage. Ainsi, on dit parfois qu'il y a des pays du Sud dans les pays du Nord et des pays du Nord dans les pays du Sud. Ces structures et ces processus affectent également le statut des langues de ces régions (voir le terme « majorité minorisée »).

- **Langues autochtones** : langues parlées par les nations autochtones d'une région ou d'un endroit donné. Les peuples autochtones sont considérés comme les « premiers peuples » ou premier·ères habitant·es de régions du monde qui ont plus tard été colonisées et occupées par un autre groupe culturel. Sur plus de 7000 langues existant dans le monde, une majorité est parlée par les communautés autochtones.

- **Langues et dialectes** : nous considérons que tout système structuré d'expression entre les êtres humains, que ce soit par la voix, des sons, des signes, des gestes ou l'écriture, est une langue. Certain·es linguistes définissent un « dialecte » comme des variétés différentes d'une même langue qui sont « mutuellement intelligibles », c'est-à-dire comprises par tous·tes les locuteur·rices de ces différentes variétés qui peuvent de ce fait communiquer. Cependant, la plupart du temps, la différence entre langue et dialecte dépend non pas de choix linguistiques, mais de choix politiques, issus de processus historiques de pouvoir et de privilèges. Pour cette raison, nous avons rarement utilisé le terme « dialecte » dans ce rapport. Nous préférons le terme « famille de langues » qui montre que beaucoup de langues peuvent avoir des histoires similaires, mais des caractéristiques différentes. Parmi elles, on peut citer l'arabe, le chinois ou l'hindi.

- **Langues locales** : dans ce rapport, nous avons défini les langues locales comme les langues parlées par le plus grand nombre de personnes dans un pays ou une région.

- **Langues marginalisées** : dans ce rapport, les langues marginalisées sont celles qui n'occupent pas une place importante sur Internet en termes de prise en charge linguistique ou de quantité de contenu, c'est-à-dire d'informations et de savoir dans ces langues. Ces langues sont marginalisées par des structures et dynamiques passées et présentes de pouvoir et de privilèges, dont la colonisation et le capitalisme, et non en raison de la taille de leur population ou de leur nombre de locuteur·rices. Certaines langues marginalisées, comme beaucoup de langues autochtones, sont déjà en voie de disparition dans le monde. D'autres sont parlées par une majorité importante de

personnes dans leur région du monde, et restent pourtant sous-représentées en ligne (par exemple, le pendjabi et le tamoul en Asie ou le haoussa et le zoulou en Afrique).

- **Langues minoritaires et majoritaires** : les langues minoritaires sont celles parlées par une minorité de la population, dans une région ou un territoire donné. Une langue majoritaire est parlée par la majorité de la population.

- **Majorité minorisée du monde** : les structures passées et présentes de pouvoir et de privilèges entraînent la discrimination et l'oppression de beaucoup de communautés et de peuples dans le monde. Souvent, ces formes de pouvoir et de privilèges se renforcent mutuellement, ce qui désavantage ou opprime certaines communautés de multiples manières : par exemple, selon le genre, la couleur de peau, la sexualité, la classe sociale, la caste, la religion, la zone géographique, le handicap et, bien sûr, la langue. Que ce soit dans le monde réel ou en ligne, ces communautés sont majoritaires dans le monde en termes de population ou de nombre, mais elles sont rarement dans des positions de pouvoir, et sont ainsi traitées comme des minorités. Autrement dit, c'est la « majorité minorisée » du monde.

En savoir plus sur comment citer et utiliser ce rapport ↪

En savoir plus sur nos ressources et nos sources d'inspiration ↪