La méthode à l'origine de la production inédite d'indicateurs de présence des langues dans l'Internet

Daniel Pimienta (*), Álvaro Blanco (*), Gilvan Müller de Oliveira (**)

(*) Observatoire de la diversité linguistique et culturelle dans l'Internet
https://obdilci.org/

(**) Chaire UNESCO des politiques linguistiques pour le multilinguisme

Traduction de l'article publié en anglais *The method behind the unprecedented production of indicators of the presence of languages in the Internet.* Frontiers Research Metrics & Analytics - Section Research Methods, Volume 8 – 2023. doi: 10.3389/frma.2023.1149347.

RESUMÉ

Des indicateurs fiables et actualisés de la présence des langues dans l'Internet sont nécessaires pour piloter efficacement les politiques linguistiques, pour prévoir le marché du commerce électronique ou pour soutenir de nouvelles recherches dans le domaine du support numérique des langues. Cet article présente une description complète des éléments méthodologiques impliqués dans la production d'un ensemble sans précédent d'indicateurs de la présence dans l'Internet des 329 langues comptant plus d'un million de locuteurs L1. Un accent particulier est mis sur le traitement de l'ensemble des biais impliqués dans le processus, provenant soit de la méthode, soit des différentes sources utilisées dans le processus de modélisation. Les biais liés à d'autres sources fournissant des données similaires sont également discutés, et en particulier, il est montré comment le manque de considération du haut niveau de multilinguisme du Web conduit à une énorme surestimation de la présence de l'anglais. La liste détaillée des sources est présentée dans les différentes annexes. Pour la première fois dans l'histoire de l'Internet, la production d'indicateurs sur la présence virtuelle d'un large ensemble de langues pourrait permettre des avancées dans les domaines de l'économie des langues, de la cyber-géographie des langues et des politiques linguistiques pour le multilinguisme.

MOTS CLÉS : Langues, Web, Internet, Indicateurs, Méthodologie, Biais, Webmetrics

REMERCIEMENTS

Le travail menant à la version 3 (et la version 1) du modèle décrit a été réalisé grâce au financement de l'Organisation de la Francophonie. Le fondement de la méthode repose en partie sur les idées exposées par Daniel Prado en 2012, notamment l'idée, appliquée au Web, d'utiliser des statistiques par pays, croisées avec des données démolinguistiques, pour obtenir des données par langue. Merci au gouvernement brésilien et à l'Institut international de la langue portugaise qui ont permis la version 2 du modèle, étape indispensable vers la version 3.

Table des matières

TABLEAUX ET FIGURES

1. INTRODUCTION

La mesure de l'espace de représentation des langues dans l'Internet ne passionne pas encore les foules, pourtant les enjeux, sur le plan linguistique, culturel, socio-économique et géopolitique, sont loin d'être neutres.

Concernant la situation des langues dans le monde, parmi les quelque 7000 langues encore existantes, environ 40% sont en danger[1] et l'intensité de leur présence dans l'Internet pourrait être un indicateur prédictif significatif. Afin de définir des politiques publiques efficaces pour les langues, mesurer la situation actuelle et son évolution est un préalable, notamment en ce qui concerne la capacité d'évaluer l'impact de ces politiques.

Aux premiers stades de l'Internet, certains chercheurs se sont penchés sur un nouveau domaine appelé cyber-géographie, qui est l'étude de la nature spatiale des réseaux de communication informatique.[2] L'acquisition d'indicateurs de la présence d'un plus grand nombre de langues dans l'Internet nous permet de proposer le concept de **cyber-géographie des langues** comme une notion connexe (Pimienta, Oliveira, 2022).

Bien que l'Internet ne soit pas un territoire homogène du point de vue de son fonctionnement et de sa gouvernance (O'Hara, Hall,2018), on peut le traiter comme un cyber-territoire réticulaire multilingue, analysant la distribution et l'interaction entre les langues dans un espace général. Dans une seconde perspective cependant, chaque langue est un territoire qui guide la densification des relations, notamment politiques et économiques. Chaque territoire linguistique est en même temps un marché, avec des capacités de production et de consommation spécifiques.

Cette vision territoriale permet d'inclure dans la discussion un autre concept pertinent : la **géopolitique des langues** et le multilinguisme (Oliveira, Pimienta, 2023). La géopolitique est constituée principalement de trois facteurs : le *territoire*, qui implique la localisation ; la *population*, dans ce cas, les locuteurs connectés de chaque langue ; et *l'effet de levier*, qui est ici l'équipement numérique de chaque langue, sa masse de contenus et ses politiques de promotion, c'est-à-dire sa capacité à recevoir des investissements (Silex, 2021). De ce point de vue, les cyber-territoires linguistiques sont des marchés en litige politique et économique (Bauböck, 2015).

Plusieurs économistes ont analysé la valeur économique des langues sous différentes perspectives (Grin, Vaillancourt, 1997 ; Gazzola, 2015). Mais malgré les instruments disponibles montrant que les langues sont fondamentales pour toutes les catégories de l'économie de services décrites par l'OMC[3], responsables d'une part croissante du PIB des pays

[1]Selon Ethnologue (https://www.ethnologue.com) le nombre exact de langues vivantes est de 7 168 alors que d'autres sources calculent qu'environ 30 000 langues ont existé (https://www.uh.edu/engines/epi2723.htm).
[2]https://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/about.html
[3]L'Organisation mondiale du commerce (OMC) propose quatre modes d'échange de services :a) du territoire d'un Membre vers le territoire de tout autre Membre (Mode 1 - Commerce transfrontières); b) sur le territoire d'un Membre au consommateur de services de tout autre Membre (Mode 2 - Consommation à l'étranger) ; c) par un fournisseur de services d'un Membre, par le biais d'une présence commerciale, sur le territoire de tout autre Membre (Mode 3 - Présence commerciale); et 4) par un fournisseur de services d'un Membre, grâce à la présence de personnes physiques d'un Membre sur le territoire de tout autre Membre (Mode 4 - Présence de personnes physiques).

au capitalisme avancé, les gouvernements et les investisseurs ont tardé à développer des perspectives plus contemporaines sur la gestion des langues.

En 2020, le e-commerce représentait à lui seul 20 % du total des ventes au détail mondiales[4], et les plateformes doivent communiquer dans la langue de leurs clients pour maintenir leur compétitivité sur le marché (voir diverses sources en annexe 8). Celui qui parvient à pénétrer les différents marchés linguistiques augmentera ses profits, ce qui amène les grandes entreprises à investir dans des stratégies multilingues (Oliveira, 2010). Un processus de *marchandisation* des langues est en cours et les données sur la présence des langues dans l'Internet sont essentielles pour la prise de décision dans ce domaine (Heller, 2010).

Depuis 2011, les décideurs politiques et les chercheurs en linguistique devaient s'appuyer exclusivement sur deux sources disponibles, toutes deux issues du domaine du marketing d'entreprise, pour évaluer l'impact de leurs politiques ou étayer leurs théories.

- ✓ W3Techs propose les pourcentages de contenus Web par langue[5], pour les 40 premières langues, avec une mise à jour quotidienne, et conserve également l'historique des pourcentages[6].

- ✓ InternetWorldStats rapporte les pourcentages de locuteurs connectés à l'Internet pour les 10 premières langues[7], avec une mise à jour annuelle.

L'analyse de la méthode de W3Techs révèle des biais sévères qui résultent de la non prise en compte de l'important multilinguisme qui prévaut sur le Web (voir 4.2 Les biais de W3Techs). Les calculs de InternetWorldStats reposent sur la combinaison du pourcentage de personnes connectées par pays, un chiffre fiable qui est publié chaque année et disponible auprès de l'Union des télécommunications internationales (UIT)[8], l'organisme des Nations Unies qui publie des statistiques sur les télécommunications, et des données démolinguistiques pour les locuteurs L1 (première langue) et L2 (deuxième langue) par pays. Les sources existantes sur les données démolinguistiques font état de grandes différences, notamment au niveau des chiffres L2 ; parmi eux, Ethnologue est généralement considéré comme la source la plus fiable ; cependant, cette source est propriétaire et non gratuite[9].

Depuis mars 2022, l'Observatoire de la Diversité Linguistique et Culturelle dans l'Internet (ci-après l'Observatoire) propose ces deux indicateurs, ainsi que des indicateurs complémentaires significatifs, pour les 329 langues comprenant une population de locuteurs L1 dépassant le million (voir résultats dans Pimienta 2022), avec des plans pour des mises à jour annuelles[10]. C'est l'aboutissement d'un long processus de dépuration des biais d'une méthode définie en 2017[11] et qui donne finalement des résultats avec un seuil de fiabilité acceptable.

L'Observatoire n'est pas un nouveau venu dans ce domaine : il a mené une série de mesures pionnières des contenus Web en langues anglaise, allemande et latine (français, italien,

[4] https://www.digitalcommerce360.com/article/global-ecommerce-sales/
[5] https://w3techs.com/technologies/overview/content_language
[6] https://w3techs.com/technologies/history_overview/content_language/ms/y
[7] https://www.internetworldstats.com/stats7.htm
[8] https://itu.int
[9] https://www.ethnologue.com/data-consulting
[10] https://obdilci.org/lc2022
[11] La méthode est décrite dans https://obdilci.org/lc2017/Alernative%20Languages%20Internet.docx .

portugais, espagnol et roumain), entre 1997 et 2007 (Pimienta, Prado et Blanco 2009). La méthode a tiré parti du nombre total d'occurrences de mots ou d'expressions dans les pages Web, qui a été rapporté par des moteurs de recherche explorant un pourcentage important de l'espace Web. L'Observatoire a été contraint de renoncer, après 2007, lorsque les moteurs de recherche ont cessé de fournir des chiffres fiables et que la proportion de pages Web indexées a été considérablement réduite.

La nouvelle méthode, développée en 2017, qui a permis de concevoir un ensemble d'indicateurs pour les 139 langues comptant plus de 5 millions de locuteurs L1, a inauguré une nouvelle approche, définie en 2012 et appliquée pour des langues uniques, principalement le français (Pimienta, 2014) et l'espagnol (Pimienta, Prado, 2016). Cette approche s'est concentrée sur la gestion d'un ensemble, aussi large que possible, de sources dispersées de données sur les langues ou les pays, ayant un certain type de rapport à l'Internet. Cette relation peut être directe (par exemple, répartition par pays des abonnés à un réseau social spécifique ou langues prises en charge dans les services de traduction en ligne) ou indirecte (par exemple, classement dans le domaine du commerce électronique ou nombre moyen de mobiles par personne dans chaque pays)[12]. La rareté des données relatives aux langues utilisées dans l'Internet a été compensée par l'utilisation de chiffres relatifs aux pays, plus nombreux, et ceux-ci ont été transformés en chiffres par langue par pondération avec les données démolinguistiques. Les données collectées ont été organisées en différentes catégories : *contenus, trafic, usages, index*[13]et *interfaces*[14]. En 2017, en apportant une cohérence mathématique et en utilisant des techniques statistiques pour extrapoler les données manquantes, la méthode a été généralisée pour de nombreuses langues, au-delà du français ou de l'espagnol. Un modèle a été conçu pour transformer l'ensemble des sources en indicateurs significatifs pour les 139 langues avec plus de 5 millions de locuteurs L1.

Par la suite et depuis 2017, les travaux ont été essentiellement consacrés à la lutte contre les différents **biais** propres à la méthode ou aux sources de données. En 2021, cela a abouti à une version 2, avec la même structure, mais avec certains biais importants contrôlés, notamment, par l'utilisation de la base de données Ethnologue Global 24 (mars 2021) pour les données démolinguistiques. Par la suite, la couverture linguistique a été étendue aux 329 langues avec plus d'un million de locuteurs L1. La poursuite de la lutte contre les biais s'est poursuivie et a conduit, en mars 2022, à une redéfinition définitive de la démarche et à la confiance dans l'atteinte d'un niveau raisonnable de contrôle des biais, avec la capacité de produire des chiffres fiables, dans un intervalle de confiance de plus ou moins 20%, une estimation empirique qui n'est soutenue par aucun calcul statistique.

Pourquoi est-il si important d'identifier les biais et, dans la mesure du possible, d'essayer de les atténuer ou, si ce n'est pas possible, d'évaluer l'impact sur les résultats obtenus de ces biais insurmontables ? Dans toute activité de recherche, la méthode scientifique exige une utilisation prudente des données et des statistiques car des biais peuvent survenir et, s'ils sont lourds, peuvent totalement discréditer les résultats obtenus. Bien qu'il s'agisse d'une démarche connue en matière de santé, où un grand nombre d'études statistiques sont menées, soit pour évaluer l'effet d'un traitement, soit pour mesurer la prévalence d'une maladie particulière dans une

[12]https://en.wikipedia.org/wiki/List_of_Wikipedias

[13]L'indice fait référence aux classements dans différents paramètres associés aux progrès de la société de l'information.

[14]La présence des langues comme option d'interface dans une liste d'applications incluant la traduction en ligne, comme une approximation d'une métrique jusqu'ici inexistante pour le niveau de support technologique des langues.

population spécifique, l'échantillonnage doit être soigneusement sélectionné et la méthode doit s'appuyer sur des bases solides (par exemple avec des procédures en double aveugle, dans lesquelles ni les participants ni les chercheurs ne savent quel est le traitement reçu par le participant). Cette préoccupation concernant les biais doit s'appliquer de la même manière à tous les domaines de la recherche.

Le domaine de la mesure des langues dans l'Internet se situe à l'intersection de deux domaines où les biais sont notables : la démolinguistique (démographie linguistique) et le Web. Dans les deux domaines, il n'y a pas de consensus fort sur les données et de grandes différences peuvent apparaître, selon les sources, sur des chiffres tels que le nombre de locuteurs de telle langue résidant dans tel pays ou le nombre total de pages Web.

Les biais peuvent se produire de différentes manières, propres aux sources de données utilisées, inhérentes à la méthode utilisée, à la sélection effectuée pour un échantillonnage, à l'hypothèse de calcul ou à l'hypothèse soutenant certaines simplifications nécessaires. S'il est de la responsabilité première du producteur de données de traiter systématiquement les biais éventuels et de documenter ceux qui subsistent, il est également de la responsabilité du chercheur utilisant ces données d'identifier les sources et de vérifier leur crédibilité, de trouver le descriptif de la méthode et analyser les biais possibles, tout cela avant de tirer des conclusions basées sur ces données. Un raisonnement juste sur des données fausses ne produira guère de conclusions fiables ! La facilité offerte aujourd'hui par les moteurs de recherche pour identifier dans la Toile des sources publiques de données n'élimine pas la nécessité de vérifier ces sources !

La méthode standard théorique pour mesurer l'espace des langues dans la Toile est de parcourir toutes les pages Web de l'Internet, d'appliquer à chacune d'elles un algorithme de reconnaissance de la langue et de compter la ou les langues de chaque page, en faisant attention qu'une seule page pourrait contenir plus d'une langue. Enfin, en divisant le nombre obtenu pour chaque langue par le nombre total de pages explorées, le pourcentage est obtenu. Avant ce processus, les éventuels biais de l'algorithme de reconnaissance du langage doivent évidemment être analysés.

D'après Netcraft[15], il existe aujourd'hui plus de 1,2 milliard de sites web, dont 200 millions sont actifs. Une source[16] évalue le nombre total de pages web à environ 50 milliards, dont moins de 10% seraient indexées par les moteurs de recherche. Dans ce contexte, cibler les sites web plutôt que les pages web est une simplification utilisée par la plupart des études, ce qui implique de nouveaux risques de biais à prendre en compte, encore plus si la reconnaissance de la langue s'applique exclusivement sur la page d'accueil de chaque site, qui bien souvent ont des éléments en anglais même pour des sites Web non anglophones. Cependant, explorer l'univers complet complète des sites Web existants est une option que même les puissants moteurs de recherche ne sont pas en mesure d'assumer ; une autre simplification est alors nécessaire, pratiquement pour sélectionner un échantillonnage réduit qui serait, espérons-le, représentatif de l'ensemble du Web. C'est là un autre risque de biais qu'un survol rapide de l'historique des tentatives mettra en évidence.

Avant 2007, le nombre d'initiatives pour tenter de mesurer le pourcentage de présence des langues sur le Web était limité ; ci-après une exploration rapide est conduite, pour une analyse plus approfondie consulter (Pimienta, Prado, Blanco 2009).

[15] https://news.netcraft.com/archives/category/web-server-survey
[16] https://www.worldwidewebsize.com

Trois des premières tentatives, dans la période 1995-1999, ont utilisées l'approche standard (équipe Babel, une initiative conjointe d'Alis Technologies et de l'Internet Society[17]) et deux autres. (Grefenstette, Noche, 2000) et l'Observatoire[18] ont utilisé des approches différentes. (Grefenstette, Noche, 2000) ont utilisé une technique pour estimer la taille d'un corpus spécifique à une langue à partir de la fréquence des mots courants dans ce corpus et l'ont appliquée à la Toile. L'Observatoire a comparé le nombre d'occurrences d'un vocabulaire équivalent dans les différentes langues étudiées (données fournies par les moteurs de recherche).

L'équipe Babel a défini son échantillonnage Web pour être analysé par une technique de randomisation des numéros IP qui a finalement consisté en un peu plus de 3000 sites Web sur la page d'accueil desquels la reconnaissance de la langue a été appliquée. Il y avait de nombreuses causes de biais, mais le problème majeur est qu'un seul échantillonnage, et donc une seule mesure, a été réalisé. En termes statistiques, l'absence d'une série de mesures invalide les résultats car un échantillonnage unique de 3000 sites Web sur un univers, à l'époque, d'un million, est totalement hors de propos. L'approche valide aurait dû être de reproduire l'opération, disons cent fois au moins, et de calculer la moyenne, la variance et d'autres attributs statistiques de la distribution obtenue. Le fait est cependant que cette approche par défaut a été réutilisée à deux reprises, (Lavoie, O'Neil, 1999) et (O'Neil et ale, 2003), et a véhiculé aux médias l'idée erronée que 80% des contenus Web était en anglais, sans aucun changement au cours de la période 1996-2003.

Dans la même période, l'Observatoire a amélioré sa méthode avec la collaboration de linguistes d'une institution partenaire, basée sur des vocabulaires équivalents dans différentes langues, en évitant autant que possible les biais potentiels. L'Observatoire a présenté des résultats montrant que l'anglais diminuait régulièrement, passant de 80 % des contenus Web, en 1996, à 50 %, en 2007. Cette approche, bien que limitée aux langues latines, anglais et allemand, a produit une série de mesures cohérentes pendant la période ; cependant, sa dépendance à la fiabilité du comptage des occurrences des moteurs de recherche a déclenché sa fin en 2007.

Deux autres initiatives ont eu lieu au cours de la période, toutes deux utilisant l'approche standard : le projet « Language Observatory » - LOP (Mikami et ale, 2005) et un projet de l'Institut statistique de Catalogne - IDESCAT (Monras, 2006). Le projet LOP, un consortium académique avec des partenaires réunissant leurs forces dans les deux exigences principales, exploration du Web à forte capacité et algorithme moderne de reconnaissance des langues, a présenté tous les attributs pour devenir la meilleure solution pour aborder le thème, combinant la rigueur des chercheurs et la puissance de la capacité d'exploration. Il a commencé à se concentrer sur les langues des pays asiatiques les moins peuplés et s'est progressivement étendu. Une collaboration a été mise en place avec l'Observatoire, sous l'égide du Réseau Mondial pour la Diversité Linguistique - MAAYA[19], lorsque LOP a produit des données pour les pays d'Amérique latine, mais malheureusement, ce projet, coordonné par l'Université de Nagaoka, a pris fin peu de temps après le tremblement de terre et le tsunami qui ont touché le Japon en 2011.

Quant au projet IDESCAT, qui s'est concentré spécifiquement sur la langue catalane, sa durée de vie fut courte. Cette période d'activités académiques autour du thème a été suivie

[17] https://web.archive.org/web/20011201133152/http://alis.isoc.org/palmares.en.html
[18] https://obdilci.org/lc2005/francais/L1.html
[19] https://web.archive.org/web/20190904002849/http://maaya.org/?lang=fr

pratiquement par un abandon de ce domaine aux sociétés de marketing, avec, comme conséquence, le règne de méthodologies non totalement transparentes et non revues par les pairs et, en même temps, un excellent marketing permettant grand impact public.

Après 2017, outre les initiatives de l'Observatoire, un consortium d'universités grecques (Giannakoulopoulos et al., 2020) a utilisé l'approche standard pour évaluer la présence de l'anglais sur les sites Web sous les domaines de premier niveau des pays de l'Union européenne (ccTLD). Leur échantillonnage comprend un peu plus de 100 000 sites et leur méthode a porté une attention toute particulière au multilinguisme des sites en vérifiant systématiquement la langue de tous les liens internes depuis la page d'accueil. À partir de leurs données de sortie, il est possible de calculer une valeur de 28 % de versions anglaises des sites Web pour tous les sites de l'Union européenne (y compris le Royaume-Uni, l'Irlande et Malte) ou 13 % pour les pays européens non anglophones (Pimienta, 2023).

W3Techs appliquait, jusqu'en mai 2022, son algorithme de reconnaissance des langues, au quotidien, sur une liste des 20 millions de sites Web les plus visités, fournie par Alexa.com, un service commercial d'analyse du trafic Web. Après mai 2022, lorsque le service Alexa a été arrêté, il a été appliqué sur la liste des millions de sites Web les plus visités fournie par Tranco[20] un service à but non lucratif orienté vers la recherche et qui se présente comme « robuste contre la manipulation ».

L'algorithme de W3Techs est appliqué sur la page d'accueil de chacun des sites de la liste Tranco et comptabilise une seule langue pour chacun d'entre eux, ignorant leur multilinguisme potentiel. L'absence d'alternative pendant une longue période, et aussi la réputation méritée de l'entreprise pour son service principal, les enquêtes sur les Technologies Web, a rendu cette source extrêmement populaire et bien souvent une référence, même pour la communauté des chercheurs. À différence des 26 autres technologies Web étudiées par l'entreprise, comme JavaScript, les langages de balisage ou les centres de données, les langues sont des *technologies Web* un peu particulières, avec la propriété que plusieurs de ces *technologies* peuvent être associées à une page Web unique ou un site Web unique. Le multilinguisme est une propriété de la Toile qui nécessite une attention particulière afin de fournir résultats non biaisés. Cette propriété est au cœur de la méthode exposée ci-après.

2. LA MÉTHODE

2.1 APERÇU

Il s'agit d'une approximation indirecte des contenus Web par langue, basée sur l'observation expérimentale systématiquement faite, depuis le début de l'Observatoire, que le rapport entre le pourcentage mondial de *contenus* et le pourcentage mondial de *locuteurs connectés* (défini comme la *productivité des contenus*) est toujours resté à l'intérieur de la fenêtre [0,5 --- 2], pour les langues à existence numérique.

Ce constat suggère l'existence d'une sorte de loi économique naturelle, qui lierait, pour chaque langue, **l'offre** (contenus et applications web dans la langue donnée) à **la demande** (locuteurs de cette langue connectés à l'Internet). Lorsque le nombre de personnes connectées augmente, le nombre de pages web augmente naturellement et en même temps, plus ou moins dans la

[20] https://tranco-list.eu

même proportion. Cela se produit parce que les gouvernements, les entreprises, les institutions éducatives, etc., et certains individus créent des contenus et des applications pour répondre à cette demande.

Il est important de noter, à l'appui de l'affirmation précédente, que les enquêtes et études sur le comportement des internautes rapportent systématiquement qu'ils préfèrent utiliser leur langue maternelle, lorsque des contenus sont disponibles, notamment pour le commerce électronique, et, en complément, sont désireux d'utiliser leur(s) deuxième(s) langue(s) (voir, à l'annexe 8, une sélection de sources sur ce sujet).

Ainsi, en fonction de chaque contexte linguistique, il y a une sorte de modulation du rapport mentionné, pour le rendre, plus ou moins, supérieur ou inférieur à un. Certaines langues ont une meilleure *productivité de contenus* que d'autres, en fonction d'un ensemble de facteurs propres à la langue ou liés au contexte des différents pays où une certaine proportion des locuteurs de cette langue se connecte à l'Internet. Les facteurs suivants ont été identifiés :

Propre à la langue :

- Évidemment, le nombre de locuteurs L2, car certaines personnes produisent, par exemple pour des raisons économiques, des contenus dans une langue différente de leur langue maternelle.
- Le support technologique de la langue pour le cyberespace, estimé à travers sa présence dans les interfaces d'applications et les programmes de traduction, ce qui faciliterait ou non la production de contenus.

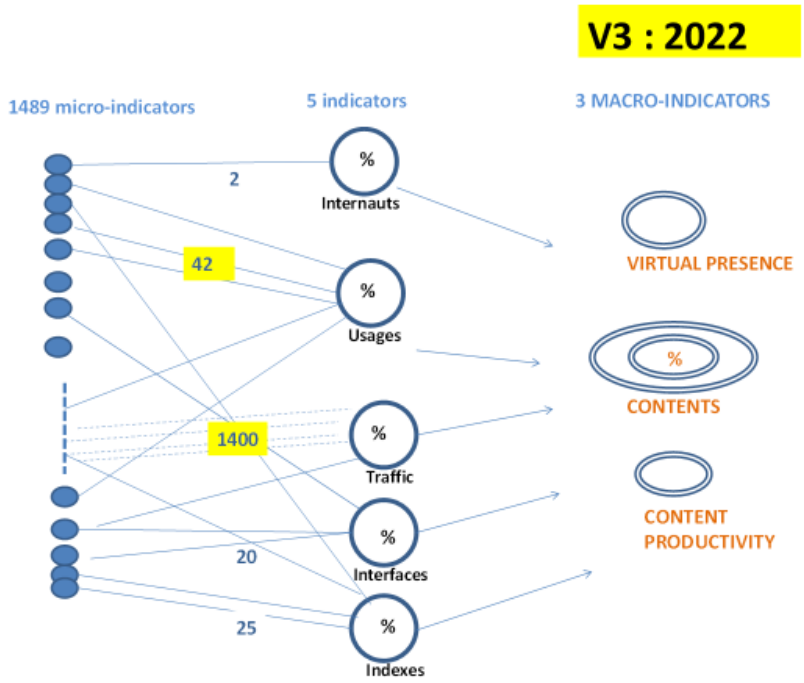
Mais aussi, selon chaque pays où il y a des locuteurs L1 ou L2 de cette langue :

- La quantité de trafic Internet, en fonction du tarif du pays, du contexte culturel ou éducatif.
- Le nombre d'abonnements aux réseaux sociaux et autres applications Internet.
- Le niveau de progrès du pays en termes de services de la société de l'information (commerce électronique, applications gouvernementales pour payer les impôts, etc.).

Par conséquent, s'il était possible de collecter suffisamment de données significatives sur chacun des facteurs mentionnés pour créer des indicateurs correspondants, on approcherait la valeur du ratio de productivité des contenus et, à partir de la proportion de *locuteurs connectés*, on déduirait la proportion de *contenus*.

C'est le cœur de la méthode et il est synthétisé dans le schéma suivant qui montre tous les indicateurs qui sont traités pour chaque langue et la quantité correspondante de sources que le modèle utilise.

Figure 1 : Des sources aux produits



2.2 DESCRIPTION DES ENTRÉES DU MODÈLE

Les entrées du modèle sont réparties en 5 types de sources : *internautes*, *usages*, *trafic*, *interfaces* et *indexes*.

2.2.1 Internautes

Il s'agit du pourcentage de locuteurs L1+L2 connectés à l'Internet pour chaque langue. La transformation des données de la source, exprimés par pays, en donnée requise, exprimée par langue, s'effectue par pondération :

CL(j) est le pourcentage de locuteurs connectés pour la langue j.

$$CL(j) = \sum_{j=1}^{j=P} LP(i, j) \times PC(i) / \sum_{j=1}^{j=P} LP(i, j)$$

Où :
P est le nombre total de pays
LP (i, j) = Le nombre de locuteurs L1+L2 de la langue j dans le pays i.
PC (i) = Le pourcentage de personnes connectées pour le pays i

Le produit matriciel CL = LP + . x PC, en notation APL[21], ou = SumProduct (LP;PC), en notation Excel, est une opération de pondération qui produit à partir d'un vecteur de la taille du nombre de pays, un nouveau vecteur, cette fois de la taille du nombre de langues.

La validité de ce calcul repose sur l'hypothèse implicite qu'au sein d'un même pays, tous les groupes linguistiques partagent le même pourcentage de personnes connectées. C'est l'un des biais fondateurs de la méthode, abordé dans le chapitre Biais.

Le vecteur CL(j) est un élément clé du modèle qui servira, toujours dans les opérations de pondération, avec différentes sources, à calculer la modulation de chaque indicateur.

La source de la matrice LP est Ethnologue ; le modèle utilise le Global Dataset #24 de mars 2021. Les sources de la matrice PC sont l'Union Internationale des Télécommunications (UIT) et la Banque mondiale ; l'UIT, la source historique de ces données[22], s'appuie sur des sources gouvernementales, et, lorsqu'ils ne sont pas disponibles, sur sa propre estimation. L'UIT ayant cessé, en 2017, de fournir ses propres estimations, la source est complétée par des données de la Banque Mondiale[23] qui comble cette lacune dans de nombreux cas. Lorsqu'aucune donnée récente n'est disponible, une extrapolation des données plus anciennes est réalisée.

2.2.2 Interfaces

Les chercheurs du réseau MetaNet[24] font un bon travail dans l'analyse du support technologique pour les langues européennes, mais il n'existe pas encore de métrique pour évaluer le support technologique pour toutes les langues dans le monde. Afin d'approximer ce paramètre, l'accent a été mis sur la présence de chaque langue dans les interfaces d'un ensemble d'applications populaires de l'Internet et comme l'une des paires dans un ensemble de services de traduction en ligne. Seize sources ont été identifiés pour lesquelles la liste des langues prises en charge est accessible (voir annexe 3).

2.2.3 Indexes

Le thème ici est d'évaluer les pays en fonction de leurs progrès selon les critères de la société de l'information. Une pondération avec des données démolinguistiques transformera cette donnée en un classement par langue. Dans la version 1, une liste de 4 sources était utilisée. À partir de la version 2, une recherche systématique a été réalisée et 27 sources ont été identifiées rendant la sélection quasi exhaustive (voir annexe 4).

2.2.4 Usages

Cinq sous-indicateurs ont été identifiés et les sources correspondantes ont été utilisées :

- Abonnés aux réseaux sociaux : 36 sources ont été exploitées, chacune liée à des réseaux sociaux comptant plus de 100 millions d'abonnés. Pour les principaux réseaux sociaux occidentaux, des sources sur le nombre d'abonnés par pays ont été identifiées ; pour les réseaux sociaux restants, principalement d'Asie, des données partielles de trafic par pays

[21]APL, « A Programming Language », qui est à la fois un formalisme mathématique et sa mise en œuvre sous forme de langage de programmation, conçu par Kenneth. Iverson. Pour plus de détails voir https://fr.Wikipédia.org/wiki/APL_(langue) .
[22]https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2021/December/PercentIndividualsUsingInternet.xlsx
[23]Source : https://data.worldbank.org/indicator/IT.NET.USER.ZS
[24]http://www.meta-net.eu

ont été obtenues, en utilisant SimilarWeb[25], extrapolées au reste des pays, proportionnellement au pourcentage de personnes connectées par pays.

- Commerce électronique : une seule source a été utilisée qui fait parfaitement le travail. Il s'agit de l'indicateur T-index de l'Imminent Translated Research Center[26]. Cet indicateur classe les pays en fonction de leur potentiel pour les ventes en ligne, estimant ainsi la part de marché de chaque pays par rapport au commerce électronique mondial. L'ensemble de pourcentage par pays est transformé par pondération avec les locuteurs connectés par langue en un ensemble de pourcentages par langue[27].
- Diffusion vidéo (streaming) : le modèle n'utilise que deux sources à ce stade : le pourcentage d'abonnés à Netflix par pays et la pénétration de YouTube par pays.
- Contenus ouverts : le modèle n'utilise qu'une seule source à ce stade : le pourcentage par pays de la somme des téléchargements OpenOffice 2012/21.
- Infrastructure : le modèle utilise trois données clés de la Banque mondiale qui sont fusionnées en deux indicateurs : % d'abonnés au haut débit fixe par pays et % d'abonnés au téléphone fixe + mobile par pays.

Les résultats finaux ont été pondérés, pour refléter la confiance actuelle dans les données[28]et ainsi réduire les biais, avec les valeurs suivantes :

- Abonnés aux réseaux sociaux : 0,3
- Commerce électronique : 0,3
- Diffusion vidéo : 0,05
- Contenu ouvert : 0,05
- Infrastructures : 0,3

qui sont par la suite transformées par pondération en répartition par langue.

La liste détaillée des sources pour l'indicateur *usages* se trouve à l'annexe 1.

2.2.5 Trafic

Des outils (tels que SimilarWeb, déjà mentionné) existent pour obtenir une estimation de la répartition du *trafic* par pays vers un site Web spécifique ; en général, ces outils proposent des données pour les sites Web classés parmi les premiers millions ou dix millions de sites les plus visités. Les deux défis sont : 1) d'évaluer ces outils et de comprendre leur biais potentiel et 2) d'établir une sélection de sites avec un minimum de biais, tout en restant dans une taille exploitable (disons autour de 1000 sites). De nombreux changements sont intervenus de la version 1 à la version 3 pour surmonter les biais ; ils sont décrits en 2.4.5. La liste des sites web utilisés pour *trafic* se trouve en annexe 5.

[25]Un service marketing fournissant une proportion du trafic par pays vers un large ensemble de sites Web : https://www.similarweb.com/

[26]https://imminent.translated.com/t-index

[27]Notez qu'Imminent fournit également l'ensemble des pourcentages par langue, faisant probablement une opération similaire. Il existe de légères différences entre Imminent et nos calculs, probablement en raison de données démolinguistiques différentes. Le modèle utilise nos calculs au lieu de la source directe Imminent car Imminent est limité à 89 langues alors que la technique d'extrapolation permet d'atteindre toutes les langues de l'étude.

[28]Une moyenne simple sans pondération sera utilisée dans la prochaine version lorsque chaque élément aura obtenu les sources nécessaires pour l'archiver.

2.2.6 Contenus

L'indicateur *contenus* était une entrée du modèle pour les deux premières versions, car la vision méthodologique originale était de collecter un maximum de sources. Une source de choix était Wikimedia, qui collecte, pour chacune de ses applications[29], et pour chaque langue traitée, des statistiques fiables et intéressantes par langue, nonobstant le fait qu'il s'agit probablement de l'application la plus multilingue du Web avec ses 327 versions linguistiques. La version 3 a conduit à supprimer cet indicateur de la liste d'entrée. Le chapitre 2.4.8 traite des biais de cet indicateur et donne la justification de cette décision.

2.3 DESCRIPTION DES SORTIES DU MODÈLE

Le modèle fournit les sorties suivantes, pour chaque langue :

Locuteurs : le pourcentage mondial de locuteurs L1+L2

Locuteurs connectés : le pourcentage de locuteurs de cette langue connectés à l'Internet

Internautes : le pourcentage mondial de locuteurs connectées

Contenus[30] : le pourcentage mondial de contenus

Présence virtuelle : le rapport *contenus* sur *locuteurs*.

La valeur mondiale (et moyenne) est 1 : une valeur supérieure à 1 signifie une présence virtuelle supérieure à la présence réelle et réciproquement.

Productivité des contenus : le rapport *contenu* sur *internautes*

La valeur mondiale (et moyenne) est de 1 : une valeur supérieure à 1 signifie une productivité élevée des locuteurs connectées.

Indice de cyber-mondialisation : ICM (l) = (L1 + L2) / L1(l) x P(l) x C(l)

où :

L1+L2/L1 (l) est le rapport du multilinguisme de la langue l (obtenu par la source Ethnologue)

P(l) est le pourcentage de pays du monde qui détiennent des locuteurs de la langue l (source Ethnologue)

C(l) est le % de locuteurs de la langue L connectés à l'Internet (calculé par le modèle)

C'est un indicateur des avantages stratégiques d'une langue dans le cyberespace[31].

De plus, en regroupant les résultats par famille de langue, le tableau exhibé auparavant *Cyber-géographie des langues* a été produit en regroupant les indicateurs par familles de langues[32], produisant une perspective globale intéressante sur la situation et les tendances.

2.4 ANALYSE DES BIAIS

Le tableau suivant montre l'évolution des biais de V1 à V3 en utilisant une note subjective de 0 (biais tellement énormes que les données n'ont aucun sens) à 20 (absolument sans biais), avec 10 (biais notables mais acceptables) au milieu.

---

[29]Wikipédia, Wiktionnaire, WikiBooks, WikiQuote, WikiVoyage, WikiSources, Wikimedia Commons, WikiSpecies, WikiNews, Wikiversity et WikiData.

[30]Dans les deux premières versions, les contenus étant un intrant, les indicateurs de sortie s'appelaient *Puissance*, *Capacité* et *Gradient*, avec exactement la même définition qu'aujourd'hui *Contenus*, *Présence virtuelle* et *Productivité de contensu*.

[31]En termes de pourcentage, l'anglais et le français détiennent ensemble près de 25 % du poids, suivis, de loin, par l'allemand, le russe, l'espagnol et l'arabe.

[32]La définition des familles de langues utilisée est celle d'Ethnologue.

Table 1 : Évaluation des biais

ÉVALUATION DES BIAIS Note sur 20	V1 2017	V2 2021	V3 2022
MÉTHODE DE BASE	**17**	**17**	**17**
MÉTHODE POUR L2	**13**	**19**	**19**
INTERNAUTES	**19**	**16**	**19**
INDEXES	**15**	**18**	**18**
CONTENUS	**5**	**8**	
TRAFIC	**13**	**11**	**17**
INTERFACES	**19**	**19**	**19**
USAGES	**12**	**12**	**16**

2.4.1 Méthode de base

Le biais implicite du cœur du modèle est de considérer que toutes les langues d'un même pays partagent le même taux de connectivité à l'Internet (la valeur nationale fournie par l'UIT). La réalité est évidemment différente car le concept de fracture numérique existe aussi au sein de chaque pays.

Cette hypothèse de travail provoque un biais positif pour les locuteurs de langues non européennes vivant dans les pays développés (qui sont probablement moins connectés que la moyenne) et réciproquement un biais négatif pour les locuteurs de langues européennes dans les pays en développement (qui sont probablement plus connectés que la moyenne). Étant un fondement de la méthode, cette hypothèse ne peut être changée et les décisions prises pour y faire face sont :

- Les comparaisons entre les performances des différentes langues à l'intérieur d'un pays ne sont pas possibles.
- Comme le risque de biais important croît de manière inversement proportionnelle à la taille de la population de locuteurs, l'étude a d'abord été limitée aux langues comptant plus de 5 millions de locuteurs L1, puis étendue aux langues comptant plus d'un million de locuteurs. Les futures versions pourront essayer d'étendre ce seuil mais probablement jamais en dessous de 100 000 car les biais pourraient devenir inévitables.

2.4.2 Méthode pour L2

Pour la première fois, en 2021, Ethnologue a étendu ses données démolinguistiques par pays aux locuteurs de L2. Cela a permis de supprimer l'un des biais les plus importants de la méthode (en V1) qui entraînait l'extrapolation des données (par exemple le pourcentage de locuteurs connectés) de L1 à L2, une méthode qui biaisait positivement les résultats des langues à forte présence dans les pays en développement, comme l'anglais et le français. En effet, ce procédé attribuait aux locuteurs L2 des pays en développement des taux de connexion Internet supérieurs à la réalité. À partir de la V2, avec l'existence de données démolinguistiques par pays aussi bien pour L2 que pour L1, le modèle travaille directement à partir des populations L1+L2 et ce biais d'extrapolation disparaît ; mais évidemment pas le biais de base qui se manifeste de la même manière pour L1 , L2 et L1+L2.

Il faut noter que les sources démolinguistiques ont un biais plus important pour les données L2 que pour les données L1, car il n'y a pas de définition parfaite du niveau de maîtrise d'une

deuxième langue requis pour être compté comme L2. De fait, les sources de données pour L2 varient dans des proportions énormes, notamment pour l'anglais[33].

2.4.3 Internautes

C'est, après les données démolinguistiques, le deuxième élément principal du modèle et il est important de s'assurer d'une source fiable. Comme mentionné au point 2.2.1, les données de l'UIT et de la Banque mondiale sont combinés pour obtenir les données les meilleures et les plus fiables à jour.

2.4.4 Indexes

Avec l'extension des sources dans la V2, atteignant une quasi-exhaustivité et une sélection d'institutions fiables (organisations internationales et organisations non gouvernementales), le biais de sélection est minime et la confiance dans les données est maximale.

2.4.5 Trafic

Les outils disponibles permettant d'obtenir la répartition du *trafic* par pays pour un large ensemble de sites (ceux considérés comme les plus visités) sont : Alexa.com, SimilarWeb.com, Ahrefs.com et Semrush.com. Tous sont issus de sociétés de marketing, pas totalement transparentes sur leur méthode. Par exemple, Alexa, la plus ancienne et la plus célèbre, bien qu'elle ait cessé ses activités en mai 2022, se produit à partir d'une bannière que les utilisateurs peuvent télécharger. Cette bannière, associée à un navigateur Web, signale à Alexa les sites visités par l'utilisateur depuis ce navigateur. Avec la collecte de toutes les données envoyées par toutes les bannières du monde, Alexa construit ses sorties, tant en termes de classement des sites que de répartition du trafic par pays. Il est évident que la répartition géographique des bannières pourrait être une indication de biais probables, mais malheureusement cette information n'est pas publiée.

Le travail pour surmonter les biais de cet indicateur a été le plus long. Dans la version 1, Alexa.com a été utilisé, avec une sélection de 450 sites Web. Il a été établi, en comparant les données de trafic par pays d'Alexa avec les données d'abonnés par pays, collectées auprès de diverses sources, qu'Alexa était positivement biaisée pour l'anglais et le français et fortement biaisé contre les pays asiatiques et le Brésil. Afin de lutter contre l'inévitable biais de sélection, le processus de l'indicateur n'a pas été réalisé par moyenne simple mais plutôt par une moyenne réduite avec un grand 20%, essayant d'atténuer ainsi les biais de sélection.

Les essais de la version 2 ont montré qu'Alexa semblait avoir corrigé le biais négatif asiatique mais un nouveau biais semblait affecter maintenant les pays européens. D'autres essais ont conduit à la découverte d'une erreur où le pays principal en termes de trafic n'était parfois pas répertorié et cela pourrait être la raison du biais observé dans les résultats, puisque cela arrive notamment avec les pays européens. Il a alors été décidé de n'utiliser Alexa que lorsque la somme des pourcentages offerts était supérieure à 70 %, un moyen simple d'éliminer ces cas erronés. Ahrefs et Semrush ont été essayés mais rejetés à cause d'un fort biais en faveur de l'anglais et pour l'un d'eux un total de pourcentages par pays souvent supérieur à 100%.

[33]La donnée d'Ethnologue pour l'anglais est de 1,348 milliard de locuteurs L1+L2 (L1 = 370 millions, L2 = 978 millions) tandis que d'autres sources proposent 1,18 milliard (https://en.wikipedia.org/wiki/List_of_countries_by_English-Speaking_population) ou 1,5 milliard (https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/ source réelle non citée). En 2008, David Crystal a exprimé la possibilité que cette donnée tende vers 2 milliards (https://www.cambridge.org/core/journals/english-today/article/two-thousand-million/68BFD87E5C867F7C3C47FD0749C7D417).

SimilarWeb a fourni des résultats relativement proches d'Alexa.com, après la correction mentionnée et il a été décidé pour la version 3 d'utiliser les deux outils en utilisant la demi somme des résultats.

Après les nombreux tests et expériences menés, il a été conclu que le biais de sélection était définitivement un problème sérieux qui devait être résolu de manière plus drastique qu'avec la moyenne réduite. La version 3 a abordé cette situation avec une nouvelle approche qui a permis de gérer une sélection de plus de 1000 sites Web où le biais a été réduit par tous les moyens possibles[34].

Pour atteindre l'objectif de sélection sans biais, il a finalement été décidé d'établir une sélection des sites Web les plus visités dans chaque pays, avec un nombre de sites Web proportionnel au trafic mondial du pays. Pour des raisons pratiques, l'algorithme n'a pas été défini pour cibler tous les pays mais a été limité aux 55 pays détenant les premières places en termes de contenus pour les langues parlées dans ces pays.

Table 2 : Liste des pays traités pour la sélection des sites nationaux

Afghanistan	Algérie	Allemagne	Angola	Arabie Saoudite
Argentine	Australie	Bangladais	Belgique	Brésil
Bulgarie	Cambodge	Chine	Hong Kong	Taïwan
Colombie	Corée du Sud	Égypte	Émirats Arabe unie	
Espagne	États-Unis	France	Inde	Indonésie
Iran	Irak	Italie	Japon	Kazakhstan
Koweït	Lituanie	Malaisie	Mexique	Maroc
Mozambique	Népal	Nigeria	Ouzbékistan	Pakistan
Pays-Bas	Philippines	Pologne	Portugal	Roumanie
Royaume-Uni	Russie	Singapour	Soudan	Sri Lanka
Tanzanie	Thaïlande	Turquie	Ukraine	Viêt Nam

La règle a été fixée pour sélectionner, pour chaque pays, au moins les trois premiers sites les plus visités, avec parmi eux, au moins, le premier domaine local (ccTld[35], comme .fr pour la France). Cette règle a été définie pour éviter que la sélection ne soit trop concentrée sur les sites mondiaux les plus visités (généralement .com). Évidemment, cela n'a pas empêché que les sites Web les plus mondiaux (tels que Google.com ou facebook.com) apparaissent dans la sélection de nombreux pays et une pondération a été effectuée pour respecter ce phénomène.

Pour réaliser la sélection, les quatre outils ont été utilisés (Ahrefs, Alexa, Semrush et SimilarWeb) bien qu'à certaines occasions, en raison du manque de données dans les pays à faible population, nous ayons dû collecter les données auprès d'autres sources.

Au total, 1421 sites Web ont finalement été sélectionnés automatiquement[36], dont 733 étaient des sites Web différents. Le nombre d'occurrences de chaque site Web dans la sélection de 1421 a été conservé pour une pondération ultérieure. Pour chaque pays, le nombre de sites Web

[34]Cette décision a obligé à renoncer à un résultat intéressant, mais statistiquement faible, des versions précédentes qui consistait à regrouper les sites Web par thème et à tirer pour certaines langues des conclusions provisoires sur sa force ou sa faiblesse par rapport à ces thèmes. La question de savoir si ces résultats reflétaient davantage les biais de sélection que certaines réalités thématiques de la présence linguistique sur Internet est restée non résolue.
[35]Domaine de premier niveau de code de pays.
[36]Le processus de sélection a été fait par programmation informatique pour éviter les erreurs ou les biais indésirables.

correspondant à sa part du trafic Internet mondial a également été calculé et conservé pour une pondération ultérieure, ceci afin de contrôler le biais de sélection.

Cette méthode a assuré la sélection la moins biaisée possible pour la mesure du trafic et a permis de surmonter l'énorme biais contre les pays asiatiques qui marquait cet indicateur depuis le début. Il a définitivement amélioré les résultats finaux pour le chinois, l'hindi et l'arabe ainsi que pour les autres langues asiatiques.

Un biais peut subsister qui pénalise les pays (et les langues associées) où le niveau général de littératie numérique est le plus élevé et pour lesquels il existe donc un trafic important vers des sites à contenus scientifique ou littéraire, et en tout cas hors réseaux sociaux et sites mondialement connus. C'est malheureusement le prix à payer pour obtenir des résultats exempts de biais majeurs. Il est clair que ce biais marginal ne favorisera pas les langues des pays développés, c'est-à-dire le plus souvent les langues européennes.

Une amélioration possible pour la version 4 pourrait être d'inclure un nouvel indicateur en établissant la proportion par pays des sites du domaine national par rapport aux sites du domaine générique ; cet indicateur pourrait être une première étape vers la mesure du degré global de littératie numérique par pays et pourrait même être utilisé à travers une nouvelle pondération pour compenser le biais résiduel en question. En attendant, les résultats bruts du modèle pourraient légèrement défavoriser le français et l'anglais et au contraire semblent désormais légèrement favoriser le chinois.

2.4.6 Interfaces

Pour chaque langue, un classement est établi en fonction du nombre de fois où la présence de cette langue existe dans la liste des applications sélectionnées (interface ou traduction en ligne). À partir de ce classement, l'opération de pondération avec le pourcentage de locuteurs connectées produit le pourcentage « modulé » attendu. Évidemment, cet indicateur est assez "agressif" car des centaines de langues sont totalement absentes de la liste et se voient donc attribuer une valeur de 0%, ce qui signifie qu'il n'y a absolument aucun support technologique. Cette mesure sévère est de toute façon le reflet de la réalité crue de ce domaine où trop de langues ont un niveau de support technologique numérique proche de zéro, malgré les efforts croissants des chercheurs en technologie linguistique[37].

4.7 Usages

L'élément *abonnés aux réseaux sociaux* était à l'origine d'un fort biais pro-occidental dans les versions 1 et 2, du fait de l'absence de réseaux sociaux non-occidentaux, et un effort particulier a été fait dans la version 3 pour compléter les 11 sources initiales[38] avec des applications analogues du reste du monde.

Le critère qui a été choisi pour compléter était de garder les réseaux sociaux avec plus de 100 millions d'abonnés. Lorsque les sources de données par pays ont été identifiées (généralement répartition des abonnés par pays) elles sont utilisées ; dans le cas contraire, la répartition des abonnés par pays a été établie à partir du trafic par pays, données obtenues par le service SimilarWeb, et étendues à tous les pays par extrapolation (voir 3.4).

[37]Voir les conférences et ateliers biannuels intenses de la communauté des chercheurs du LREC depuis 1998 : http://www.lrec-conf.org.

[38]Facebook, Linkedin, Twitter, Instagram, Reddit.

La répartition par pays, après extrapolation de chaque élément, est pondérée en fonction du nombre d'abonnés et est finalement transformée en pourcentage par langue par pondération avec la matrice démolinguistique.

Dans la version 3, la complémentation a considérablement réduit le biais contre les pays non occidentaux et indirectement contre les langues non européennes. La liste complète des réseaux sociaux traités est consultable en Annexe 1.

Pour l'élément *commerce électronique*, comme mentionné précédemment, la source est unique mais tout à fait fiable.

Pour le *streaming vidéo*, le modèle n'utilise que deux sources à ce stade : le pourcentage d'abonnés Netflix par pays. Ce sous-indicateur doit clairement être étendu dans la prochaine version du modèle avec des applications de streaming alternatives au-delà de YouTube et Netflix, avec un effort particulier pour les pays non occidentaux. D'ici là, l'élément reçoit une faible pondération.

Pour les *contenus ouverts*, ce sous-indicateur doit également être étendu dans la prochaine version du modèle avec plus de données liées à l'ouverture, en particulier dans le domaine des MOOC. L'élément reçoit une faible pondération pour le moment.

Pour les *infrastructures*, les données de la Banque mondiale sur les lignes fixes, mobiles et haut débit par pays sont fiables et offrent une base solide pour l'indicateur. La somme des lignes fixes et mobiles dans une seule donnée équilibre la situation entre les pays développés avec une forte pénétration des lignes fixes et les pays en développement avec une forte pénétration mobile.

Il reste qu'à ce stade, l'indicateur *usages* est celui qui a reçu le moins d'attention et il doit être amélioré pour la prochaine version, quoique que l'objectif principal qui était de surmonter le biais occidental a pu être atteint. Les réseaux sociaux non-occidentaux ont été intégrés pour le volet réseaux sociaux et cela a produit les effets attendus sur les résultats, révélant la présence en plein essor des pays et des langues asiatiques.

2.4.8 Contenus

Cet indicateur est celui, avec *trafic* et *usage*s, qui a reçu la plus grande attention dans les travaux contre les biais. C'est aussi celui dont les biais, hérités de la galaxie Wikimédia, ont eu l'influence majeure sur les résultats des deux premières versions, donnant un avantage notable, pour des indicateurs indépendants de la population de locuteurs, aux résultats des langues majoritairement présentes dans Wikimédia.

Les deux principaux défis avec Wikimedia sont, premièrement, que malgré ses efforts notables et son succès pour être véritablement mondial, il souffre d'un parti pris occidental et, deuxièmement, que certaines langues particulières ont beaucoup investi pour participer à l'encyclopédie en ligne et présentent des présences extrêmement disproportionnées par rapport à la réalité de leur nombre de locuteurs connectées[39], tandis que d'autres langues ont vu leurs résultats dans les premières versions dopés par leur forte présence dans les services Wikimédia[40]. De plus, certaines langues ont augmenté artificiellement leur nombre d'articles en

[39]C'est le cas du cebuano, du malgache et du tagalog.
[40]Comme l'hébreu, le suédois ou le serbo-croate.

les traduisant à partir d'autres versions linguistiques tout en maintenant un taux de mises à jour extrêmement bas.

Dans la version 2, une formule a été définie et utilisée comme indicateur, à la place du nombre d'articles Wikipédia, pour supprimer efficacement l'avantage artificiel mentionné :

W (i) = Articles (i) x Éditions (i) x Éditeurs (i) x Profondeur (i) / L1+L2 (i) ²

Où :
Articles (i) = le nombre d'articles Wikipédia pour la langue i
Éditions (i) = le nombre d'éditions des articles pour la langue i
Éditeurs(i) = le nombre d'éditeurs pour les articles de la langue i
Profondeur (i) = un indicateur de la fréquence des mises à jour des articles pour la langue i[41]
L1+L2(i) = le nombre de locuteurs première et deuxième langue de i.

Tous les éléments de la formule sont fournis dans les statistiques de Wikipédia, pour plus de détails, voir (Pimienta 2021).

Pour la version 3, un effort profond et systématique a été consacré à équilibrer les données de Wikipédia avec des données équivalentes pour d'autres langues. Le tableau exposé en annexe 2 liste les encyclopédies en ligne traitées avec les données recueillies, principalement en nombre d'articles. À partir de ce tableau, l'indicateur *contenus* a été construit avec une représentation plus juste des langues en cumulant, par langues, les différents nombres d'articles. La conclusion de cet effort lourd, nécessaire, mais finalement frustrant, c'est que certaines langues (comme le chinois ou le turc) ont investi massivement dans les encyclopédies en ligne alors que d'autres ne semblent pas s'intéresser à la question.

C'était un vrai dilemme d'abandonner les merveilleuses statistiques de Wikimedia ; cependant, la suppression de l'indicateur *contenus* come donnée d'entrée a conduit au renouvellement positif de la conception de l'approche vers un modèle cohérent où les biais sont maitrisés.

Au lieu d'appeler *puissance* la sortie principale, elle a été renommée directement *contenus* et *capacité* et *gradient*, avec la même opération arithmétique sont devenus *indicateur de présence virtuelle* et *productivité des contenus*, des concepts beaucoup plus naturels et compréhensibles. Par ailleurs, toutes les opérations de pondération qui étaient développées à l'intérieur du modèle à partir de la version 1 étaient désormais reflétées de manière cohérente dans la conceptualisation de l'approche, comme une modulation de la *productivité des contenus*. Dans le même temps, les anomalies mentionnées pour les deux premières versions du modèle, qui étaient motivées par les particularités de Wikimedia, ont disparu pour laisser place à des résultats plus fiables et prévisibles[42].

[41]Voir la définition précise dans https://meta.wikimedia.org/wiki/Wikipedia_article_depth

[42]Le meilleur symptôme est que le japonais est monté en première place pour la *présence virtuelle* et *la productivité des contenus* ce qui est cohérent avec la réalité de l'usage omniprésent de l'Internet au Japon. Certaines des langues qui ont été favorisées par leur forte présence dans Wikipédia restent dans des positions élevées, mais pas aux premières places, ce qui permet de maintenir l'affirmation selon laquelle les langues des pays (ou régions) les plus performants dans les paramètres de la société de l'information bénéficient de bonnes places dans le des indicateurs de *présence virtuelle* ou de *productivité des contenus*.

2.5 MODÉLISATION

2.5.1 Pré-traitement

L'essentiel des données fournies par Ethnologue se présente sous la forme d'une matrice Excel de 11 500 lignes au format suivant : "ISO639[43], *Nom de la langue, Nom du pays, nombre de locuteurs L1, nombre de locuteurs L2 »*, ainsi qu'un grand nombre de paramètres associés non utilisés pour cette méthode et qui ont été supprimés.

Afin d'obtenir le format requis par le modèle (une matrice avec tous les pays considérés en colonnes et toutes les langues considérées en lignes), une série d'étapes a été mise en place avec le support de différents programmes écrits sous forme de macros VBA[44]. Une des étapes les plus complexes a été de fusionner toutes les données des langages appartenant à la même macro-langue. Ce processus a impliqué 60 macro-langues comprenant 434 langues différentes[45] (voir détails en annexe 6).

Après avoir terminé cette étape, le processus consistait à réduire la liste complète des langues pour ne garder que celles qui sont gérées par le modèle, en additionnant soigneusement tous les nombres restants par pays sur une seule ligne pour le reste des langues.

Il est important de comprendre que l'adoption des données d'Ethnologue implique l'acceptation de ses règles de présentation, qui reposent sur des considérations purement linguistiques :

- Regroupement de macro-langues[46]
- Liste des pays et dénominations anglaises correspondantes.

La liste des pays traités par Ethnologue est plus longue que celle traitée par l'UIT pour la fourniture des pourcentages de connexion à l'Internet par pays : l'UIT, en tant qu'entité des Nations Unies, ne sépare pas, par exemple, la Martinique de la France. Dans ce cas, la règle de l'UIT est celle qui prévaut et l'exigence a été de rassembler soigneusement les données Ethnologue pour les 29 pays non considérés par l'UIT (pour la liste complète, voir l'annexe 7) dans une seule colonne "Autres pays"[47].

2.5.2 Gestion des sources pour les micro-indicateurs

L'ensemble du processus de gestion des sources des micro-indicateurs est la tâche la plus lourde et la plus difficile du projet, avec une forte consommation de ressources humaines. De nombreuses étapes sont nécessaires :

1. Pour chaque indicateur, vérifier que les sources pour 2017 sont toujours disponibles et à jour, sinon rechercher dans l'Internet d'autres sources comparables.

[43]Le code ISO à 3 caractères attribué à chacune des 7486 langues identifiées.
[44]Virtual Basic Applications, un langage utilisé pour créer des macros exécutables dans Excel.
[45]Par exemple, la macro-langue arabe réunit 29 langues comme l'arabe égyptien ou l'arabe marocain.
[46]Un exemple significatif est le cas de la macro-langue serbo-croate dont la définition comprend, par ordre alphabétique, le bosnien, le croate, le monténégrin et le serbe. Ce regroupement ne répond pas du tout aux critères géopolitiques et pourrait même être considéré comme controversé de ce point de vue. De plus, certaines sources séparant clairement les langues et les pays concernés, cela entraîne un risque d'erreur dans les résultats, même si la saisie des sources a été transformée pour tenir compte de cette situation (le risque survient lorsque les données ne doivent pas être additionnées mais plutôt moyennées comme dans l'indicateur de profondeur de Wikipédia).
[47]À noter que le Kosovo dispose de données fournis par l'UIT mais est absent de la liste des pays Ethnologue : pour cette raison, il n'apparaît pas dans les résultats.

2. Sélectionner de nouvelles sources en fonction de leur fiabilité et de leur applicabilité au processus[48].
3. Rassembler les sources sélectionnées dans un format permettant une introduction simplifiée dans le modèle.
4. Introduire les sources validées dans le modèle.
5. Évaluer les biais de la source.

En annexe 5, la liste complète des sources est présentée, pour chaque indicateur.

Pour effectuer l'étape 4, les données doivent être transformées au format Excel, avec les noms de pays et de langue correspondant à ceux du modèle et dans le même ordre séquentiel.

À l'étape 3, toutes les sources sont collectées à partir d'une URL spécifique (voir l'annexe 5 pour la liste complète des URL) et la plupart des sources sont obtenues au format HTML. Certaines sources sont au format PDF et un sous-ensemble limité (principalement celui de l'UIT et de la Banque mondiale) est au format Excel, celui choisi pour transformer toutes les sources. Le processus de conversion de PDF vers Excel peut être relativement simple dans la plupart des cas, lorsque les tableaux sont bien structurés, mais dans certains cas, il y a une incompatibilité et quelques astuces sont nécessaires, comme passer par un format .doc intermédiaire.

Le processus de transformation de HTML vers Excel peut souvent devenir un véritable cauchemar nécessitant beaucoup d'imagination, y compris dans certains cas la nécessité d'aller chercher les données à l'intérieur de la source HTML et d'essayer à partir de là de construire un tableau en utilisant la fonction de conversion d'Excel, après nettoyage du code HTML entourant les données.

Dans un nombre croissant de cas, la source offre un accès géographique aux données (cartes cliquables) lequel, sauf lorsque le nombre de pays ou de langues est limité et que la copie à la main n'est pas trop lourde, rend impossible un traitement automatisé ou nécessite l'externalisation d'un travail manuel de collecte qui est fastidieux et demande une grande concentration et discipline pour éviter les erreurs.

Il convient de remercier les institutions (généralement des organisations internationales ou des ONG) qui fournissent les données dans un format lisible par ordinateur (Wikimedia fournit, par exemple, dans sa version anglaise, des tableaux HTML qui peuvent être transformés directement au format Excel sans perte de structure).

L'obtention d'une copie de la source au format Excel ou compatible (généralement un tableau de noms de pays ou de langues avec des valeurs ou des pourcentages associés) n'est pas la fin du processus. Avec 215 pays et 329 langues à traiter et, au lieu d'utiliser un code ISO sans ambiguïté, l'usage courant de noms littéraux qui peuvent être dans différentes langues et dans des orthographes non-standard, l'intégration des données dans le modèle ne peut se faire à la main. Deux programmes ont été écrits pour ce processus, qui nécessitaient tous deux un réglage récursif[49] pour s'adapter aux différentes orthographes. Les sorties du programme sont des fichiers Excel qui peuvent être utilisés directement pour intégrer les données dans le modèle. Outre le gain de temps appréciable de cette méthode informatisée, elle garantit d'obtenir les données sans erreur.

[48]Il peut arriver que des données fiables soient dans un format qui interdit l'exploitation automatisée.

[49]Le processus récursif reconnaît les nouvelles orthographes et se termine lorsque la vérification des erreurs n'identifie plus d'orthographes inconnues.

À noter également que la gestion des macro-langues a rendu ce processus encore plus complexe, car le regroupement des langues doit être effectué dans les données sources avant traitement par la macro. Pour prendre quelques exemples, les occurrences fréquentes de l'arabe égyptien ou marocain dans les sources ont été cumulées dans la macro-langue arabe et celles du serbe, du bosniaque, du croate et du monténégrin ont été fusionnées dans le serbo-croate (le nombre de cas similaires étant assez haut). Pour le traitement manuel des orthographes inconnues signalées par le programme (incorporation des orthographes comme synonymes ou rejet dans l'autre catégorie), la page Ethnologue descriptive de chaque code langue a été utilisée en support[50].

2.5.3 Structure du modèle et processus

Le modèle est implémenté dans un fichier Excel de 17 onglets qui sont présentées ci-dessous, avec le processus correspondant.

UIT : une copie de la source UIT, modifiée en fonction du pré-traitement.

SP : (pour « Speakers ») la matrice des locuteurs L1+L2 par pays.
En lignes, les 329 langues, triées par code ISO à 3 digits (ISO369), en commençant par la ligne 9 avec la somme du reste des langues non traitées.
E,n colonnes, les 215 pays traités, triés par code ISO à 2 digits, en commençant dans la colonne I par la somme du reste des pays non traités.

Les 8 premières lignes et colonnes sont réservées aux informations de contrôle :
Lignes de contrôle : code pays 3 caractères, code pays 2 caractères, nom du pays, total des locuteurs L1+L2 du pays, % de personnes connectées, nombre de personnes connectées, % de connexion mondial, total ou moyenne (nombre de langues parlées par pays), langues restantes.
Colonnes de contrôle : ISO639, nom de la langue[51], total de locuteurs L1+L2, % mondial de locuteurs L1+L2, % mondial de locuteurs L1+L2 connectés, nombre de pays avec locuteurs, nombre de locuteurs L1, ratio L1+L2/L1, pays restants.

Cette fiche est protégée de la lecture car elle contient des informations propriétaires d'Ethnologue qui ne peuvent être rendues publiques.

SP2 : (pour SP numéro 2) données démolinguistiques secondaires calculées à partir de SP.
Pour les 329 langues en lignes, et le reste des langues : % mondial de locuteurs L1+L2, nombre de locuteurs L1+L2, % mondial de locuteurs L1+L2 connectés, nombre de locuteurs L1+L2 connectés, % mondial de locuteurs L1+L2 connectés, % mondial de locuteurs L1 connectés, % mondial d'internautes L1+L2.

PL : (pour « Percentage Language ») Matrice parallèle à SP où PL(i,j) = % d'internautes de langue i du pays j connectés, calculée à partir de SP et SP2. Il s'agit d'une information redondante utilisée pour simplifier l'opération de pondération effectuée dans l'onglet **Wut**.

Mil : (pour « Micro-Indicator Language ») Contient la liste des langues en lignes et la valeur 0 ou 1 selon l'absence ou la présence de la langue dans l'une des 16 applications utilisées pour l'indicateur d'interface, renseignée à partir des sources des langues.

[50] https://www.ethnologue.com/language/srp
[51] Suivi de « macro » si c'est une macro-langue.

MIc : (pour « Micro-Indicator Country ») Contient la liste des pays en colonnes et les données des sources par pays, successivement pour les entrées *indexes*, *usages* et *trafic*. Pour la version 3 il y a 786 lignes.

À noter qu'un pré-traitement est nécessaire pour *usages* afin d'intégrer les réseaux sociaux non occidentaux ; cela se fait dans l'onglet **MIcU**.

À noter qu'un post-traitement est nécessaire pour le trafic afin d'effectuer la pondération avec le nombre optimal de sites en fonction du % de personnes connectées par pays ; cela se fait dans **MIcT** et **MIcT1**.

Les colonnes de contrôle sont les suivantes :
Col. A : indique le type d'indicateur à partir d'une recherche du nom dans **MATRIX**.
Col. B : indique le nom de l'indicateur
Col. C : selon le type de données, calcule la moyenne ou le total, ou un produit matriciel avec le nombre de personnes connectées par pays des entrées dans chaque ligne
Col. D : indique le type de données, soit un pourcentage mondial par pays, soit une quantité par pays, soit un pourcentage au sein de chaque pays
Col. E : indique si une extrapolation est requise et, le cas échéant, lequel des deux types d'extrapolation
Colonne F : calcule le nombre de pays disposant de données sources
Col. H : contient l'URL de la source sauf pour le trafic où elle indique le nombre de fois que le site a été cité, afin de permettre la pondération correspondante en **Wut**.

Les lignes de contrôle sont les suivantes :
La ligne 8 indique pour chaque pays le nombre de sites web qui ont été mesurés.
La ligne 9 indique le ratio du nombre de sites qu'il aurait fallu utiliser pour respecter la proportionnalité de personnes connectées (le produit ligne 8 par ligne 9 pour chaque pays représente le nombre exact de sites requis pour ce pays dans l'hypothèse du total réel de sites Web (cellule C7).Ceci sera utilisé comme facteur de pondération pour obtenir une représentation équitable des mesures de trafic dans **MIcT1** et **MIcT** avant la pondération par le nombre d'occurrences de sites Web effectuées dans **Wut** (cela a été ajouté dans V3.c pour corriger une erreur en V3 où la pondération se faisait en parallèle avec la pondération démolinguistique, ce qui était une erreur aux conséquences très marginales).

MicU : (pour « Micro-Indicator Country Usage ») Le dernier onglet ajouté pour le nouveau traitement V3 de *usages*. Comprend une copie complète des sources d'utilisation migrées de **MIc**, aux mêmes lignes, complétées par T-Index et la liste des nouveaux réseaux sociaux rajoutés pour la V3. Pour cette nouvelle liste, les mesures de trafic par pays obtenues à partir de SimilarWeb sont définies, suivies du processus d'extrapolation (voir **EX**). Le résultat est une nouvelle et dernière ligne appelée "*Réseaux sociaux pondérés*" qui est obtenue en pondérant la liste complète des réseaux sociaux avec le total correspondant d'abonnés, équilibrant finalement de manière les réseaux sociaux occidentaux avec ceux du reste du monde.

Ma : (Masque d'absence) Un onglet parallèle à **MIc** contenant la valeur 1 lorsqu'une donnée est absente pour le couple (pays, source). Utilisé pour l'extrapolation.

Mp : (Masque de présence) Un onglet parallèle à **MIc** contenant la valeur 1 lorsqu'une donnée existe pour le couple (pays, source). Utilisé pour l'extrapolation.

EX : (Extrapolation) Un onglet parallèle à **MIc** où le processus d'extrapolation est effectué. Deux processus différents sont utilisés selon le type de données.

Lorsque les données sont exprimées en pourcentage mondial par pays, le complément à 100% est réparti entre les pays qui n'ont pas reçu de données, au prorata de leur pourcentage mondial de personnes connectées à l'Internet. C'est typiquement le cas de la mesure de trafic où les outils utilisés, Alexa et SimilarWeb, ne couvrent pas tous les pays.

Lorsque les données sont une notation par pays, la technique du quartile est utilisée: quatre valeurs de quartile sont placées en fonction du pourcentage de personnes connectées dans l'intervalle compris entre : 0 %, 15 %, 35 %, 65 %, 85 % et 100. %.C'est typiquement le cas des données *Indexes*.

À noter que lorsqu'il apparaît que l'une ou l'autre des méthodes ne peut pas fournir d'extrapolation signifiante, la source des données est exclue du modèle.

Dans les rares cas où tous les pays sont renseignés par la source, aucune extrapolation n'est évidemment nécessaire, comme c'est le cas pour les données NapoleonCat de pourcentage d'abonnés par réseau social.

Noter que pour le processus d'utilisation en V3, l'extrapolation pour les réseaux sociaux n'est pas effectuée dans **EX** et a été répliquée dans **MIcU**.,

Noter que la somme des valeurs du T-Index pour les pays répertoriés est de 99,78 %, si proche de 100 % qu'aucune extrapolation n'a été effectuée.

MicT1 : (« Micro-Indicateur Trafic1 ») L'onglet est parallèle à **MIc** et n'est rempli que pour les indicateurs de trafic. Chaque cellule (pays, site Web) contient le produit du trafic source de **MIc** ajouté au trafic extrapolé d'**EX**, multiplié par le facteur de pondération pour le pays (ligne 10 de **MIc**. La somme des pourcentages est calculée et placée dans la colonne G pour une normalisation ultérieure à 100 % en **MicT**.

MicT : L'onglet est parallèle à **MIcT1** et sert à la normalisation des données à 100% pour chaque site en divisant chaque cellule par le total. Les résultats seront utilisés dans **Wut** pour calculer la répartition finale du trafic par langue.

Wut : (« Weighting Usages and Traffic » - Poids *usages* et *trafic*) Dans cet onglet sont traités les indicateurs *usages* et *trafic*. Le procédé consiste à pondérer les valeurs avec le pourcentage de locuteurs connectés par pays, issus de **PL**, après application de l'extrapolation. Pour l'indicateur de *trafic* l'extrapolation a déjà été effectuée dans **MIcT** mais il y a une pondération supplémentaire à effectuer avec les données calculées dans **MIc** (colonne H) pour le nombre d'occurrences de chaque site web.

Wi : (« Weighting Indexes » – pondération indexes) Dans cette feuille, la pondération est effectuée avec les données démolinguistiques afin d'obtenir des données par langue pour l'indicateur *indexes*, à partir de la colonne BA, suivi, à partir de la colonne 10, de la normalisation à 100%.

Pi1 : (« Process indicator language » – traite l'indicateur langue) Dans cet onglet, la pondération avec les données démolinguistiques est effectuée, afin d'obtenir des données par langue, pour

les indicateurs par pays *usages* et *trafic*. Pour *usages*, une pondération supplémentaire est effectuée avec le poids attribué à chaque composante de cet indicateur (voir 2.2.4). Pour *trafic*, une pondération supplémentaire est effectuée avec le nombre d'occurrences de chaque site web dans l'échantillon (voir 2.2.5)

RES : (Résultats) Les résultats finaux de chaque indicateur par langue (*usages*, *trafic*, *indexes*, *interfaces*) sont calculés.

FINAL : Cet onglet présente les résultats finaux avec tous les paramètres associés et propose les résultats triés par *contenus*, *présence virtuelle*, *productivité des contenus* et *locuteurs connectées*. Il présente également les 20 premières positions de contenu et crée la cyber-géographie du résultat linguistique (voir tableau 1). Une copie sans formule de cet onglet est rendue publique en tant que produit du modèle (voir http://obdilci/Results).

À noter qu'un accès à ces résultats sous la forme de base de données est prévu avant fin 2022, avec les codes ISO 639-2 comme clé d'accès.

MATRIX : La liste de tous les micro-indicateurs utilisés dans le modèle pour chaque type (*indexes, interfaces, usages, trafic*).

3. RÉSULTATS

Le modèle élaboré produit, pour chaque langue, les indicateurs suivants :

A. Part des locuteurs L1+L2 dans le monde
B. Pourcentage de locuteurs L1+L2 connectées
C. Pourcentage de locuteurs connectées L1+L2
D. Pourcentage de contenus Internet
E. Indicateur de présence virtuelle (défini comme le rapport D/A)
F. Indicateur de productivité de contenus (défini comme le rapport D/C)

Des constructions plus élaborées sont réalisées à partir de l'agrégation de ces indicateurs, comme le tableau suivant "*Cyber-Géographie des Familles de Langues*", qui donne une perspective globale de la situation des différentes familles de langues[52] et montre que les langues asiatiques sont en passe de prendre le pas sur les langues européennes alors que les langues africaines sont dans une situation difficile, du fait de la fracture numérique qui prévaut, traduite en fracture linguistique[53].

[52]Les familles de langues comprennent, pour chaque région, les langues qui sont originaires de cette région. L'anglais, le français et l'espagnol sont des langues européennes et, selon la classification Ethnologue que nous utilisons, le russe est classé comme langue européenne tandis que le turc et l'hébreu sont des langues asiatiques.
[53]Moins de 30% des locuteurs de langues africaines connectés à l'Internet et une *présence virtuelle* et une *productivité de contenus* très faibles sont obtenues.

Table 3 : Cyber Géographie des familles linguistiques

Langues de	Afrique	Amériques	monde arabe	Asie	Europe	Pacifique	Non inclus	TOTAL
Locuteurs L1+L2	9,21 %	0,31 %	3,53 %	48,24 %	30,91 %		7,81 %	**100%**
Internautes %	29,8 %	56,7 %	64,0 %	49,3 %	82,6 %		47,06 %	**56,91 %**
% des internautes	5,21 %	0,32 %	3,89 %	44,63%	39,51%		6,36 %	**100%**
Contenus	2,89 %	0,22 %	3,09 %	44,77 %	45,39%		3,64 %	**100%**
Présence virtuelle	0,31	0,71	0,88	0,93	1,47		0,47	**1**
Productivité des contenus	0,56	0,69	0,79	1,00	1.15		0,57	**1**
Nombre de langues	138	8	1	135	47	0		**329**

Les résultats du modèle peuvent être consultés en mode CC-BY-SA-4.0, à https://obdilci.org/lc2022 et peuvent être lus dans (Pimienta, 2022). Les résultats des mesures antérieures et postérieures sont accessibles à https://obdilci.org/Results.

Dans un souci de contrôle par recoupement des résultats, le modèle a été exécuté séparément avec les données L1 uniquement et avec les données L2 uniquement (voir l'annexe 9 pour les résultats correspondants qui représentent un contrôle indirect tout à fait positif de la méthode).

4. DISCUSSION

L'observation de la présence des langues dans l'Internet a connu une forte activité dans la période 2000-2007 (Pimienta, 2009) mais, après cette période, comme mentionné en introduction, seules deux options sont restées disponibles pour le grand public : InternetWorldStats et W3techs.

Les deux donnent quelques points saillants de leur méthodologie respective, mais aucun article scientifique évalué par des pairs n'a abordé leurs biais respectifs ; leur présence de longue date, sans données alternatives, leur a assuré un grand nombre de citations dans divers articles nécessitant ces données, trop souvent sans la prudence nécessaire qui exigerait la réalité de leurs biais.

4.1 Biais d'InternetWorldStats (IWS)

Les données produites par IWS diffèrent légèrement de celles de l'Observatoire, principalement parce que les sources de données démolinguistiques ne sont pas les mêmes, et que, surtout pour les données L2, les écarts entre sources peuvent être énormes (voir note 26). Une autre différence existe cependant sur la gestion des données L2. L'Observatoire calcule les pourcentages L1 + L2 de langues dans le monde par rapport au nombre total de locuteurs L1 + L2. C'est-à-dire une valeur 43% supérieure à la population mondiale, selon la source Ethnologue[54]. Quant à IWS, il calcule les pourcentages pour L1+L2 par rapport à la population

[54]Dans les données de 2021, celles que nous utilisons, Ethnologue compte la population mondiale (nombre total de locuteurs L1) à 7 231 699 136 et le nombre total de locuteurs L1+L2 à 10 361 716 756.

mondiale (méthode appelée *approche à somme nulle*[55]). À moins qu'il y ait une astuce cachée quelque part dans les calculs, l'approche "à somme nulle" semble provoquer une erreur grossière en surévaluant les 10 langues mentionnées, erreur cachée dans le pourcentage pour les langues restantes, qui deviendra négatif à un moment donné si le nombre de langues est étendu jusqu'au point où la somme des locuteurs L1+L2 croise la valeur L1.

4.2 Biais de W3Techs

La méthode utilisée par W3Techs consiste à appliquer un algorithme de reconnaissance de la langue à la page d'accueil de 10 millions de sites Web qui sont sélectionnés par certains services d'analyse du trafic Web (Alexa.com ou tranco-list.eu, jusqu'à fin 2022) comme les plus visités.

Les différences entre les résultats de W3Techs et ceux de l'Observatoire sont énormes souvent dans un rapport de 1 à 3, parfois, comme pour le chinois et l'hindi, dans un rapport supérieur à 1 pour 10). L'une des deux sources, au moins, doit être extrêmement biaisée ! Le tableau suivant expose ces différences en utilisant les données W3Techs du 24/8/22 et les données de l'Observatoire de la V3.1 en 8/2022.

Table 4 : Comparaison des données W3Techs vs Observatoire

	W3TECHS		OBSERVATOIRE	
LANGUE	**Rang**	**Contenus[56]**	**Rang**	**Contenus**
Anglais	**1**	**61.4 %**	**1**	**19,92 %**
Russe	**2**	**5,6 %**	**4**	**3,86%**
Espagnol	**3**	**3,9 %**	**3**	**8,09 %**
Turc	**4**	**3,2 %**	**12**	**1,15 %**
Allemand	**5**	**3,1%**	**dix**	**2,38 %**
Français	**6**	**3,0 %**	**6**	**3,43 %**
Persan	**7**	**2,7 %**	**16**	**0,89 %**
Chinois	**9**	**1,7 %**	**2**	**19,82 %**
Arabe	**13**	**1,1 %**	**8**	**3,14 %**
Hindi	**35**	**0,1 %**	**5**	**3,67 %**

Les différences les plus importantes se trouvent dans les pourcentages de contenus pour l'hindi et le chinois, en plus de la différence concernant l'anglais (plus de 60 % contre environ 20 %).

En août 2022, l'agrégateur de statistiques Statista[57], s'appuyant sur les données de W3Techs, affirme que « *l'anglais est la langue universelle de l'Internet,* » alors que l'Observatoire, dans le même temps, affirme : « *La transition de l'Internet entre la domination des langues européennes, l'anglais en tête, vers les langues asiatiques et l'arabe, le chinois en tête, est bien avancée et le vainqueur est le multilinguisme, mais les langues africaines tardent à prendre leur place* ». Encore une fois, ces deux affirmations ne sont pas compatibles, une au moins est fausse.

[55]Citation en provenance du site IWS *: En effet, de nombreuses personnes sont bilingues ou multilingues, mais ici nous n'attribuons qu'une seule langue par personne afin que tous les totaux des langues s'additionnent à la population mondiale totale (approche à somme nulle).*
[56]Notez que W3Techs propose des valeurs avec un seul chiffre après la virgule.
[57]https://www.statista.com/chart/26884/languages-on-the-internet/

On pourrait discuter du biais vers l'anglais des algorithmes de reconnaissance de langue, du biais vers l'anglais de la sélection des 10 millions de sites Web les plus visités[58]; mais ce sont des biais marginaux qui ne pourraient pas expliquer des différences aussi importantes. Le problème principal est dans le **manque de considération du multilinguisme**, une caractéristique du Web qui est ignorée par la méthode W3Techs, alors que le Web est probablement encore plus multilingue que l'humanité[59].

En arrière-plan de cette discussion, il est important de rappeler le point énoncé et documenté à l'annexe 8, à savoir que les internautes préfèrent utiliser leur langue maternelle sur le Net comme première option et sont désireux d'utiliser leur(s) seconde(s) langue(s) en complément.

Le problème est ainsi dans la décision de mesurer uniquement les pages d'accueil et de compter une seule langue pour chacune. De nombreux sites Web non anglais peuvent avoir des résumés en anglais ou quelques mots anglais dans leurs pages d'accueil et sont probablement comptés comme anglais. De nombreux sites Web en anglais ont plusieurs autres versions linguistiques qui doivent également être comptées (si, comme c'est probablement le cas, l'algorithme est défini dans un environnement informatique en anglais, le site Web est compté comme étant uniquement en anglais).

W3Techs produirait des données assez différentes (et, espérons-le, plus proches de ceux de l'Observatoire) si les règles suivantes étaient rajoutées à son algorithme :

- Le comptage est effectué sur les pages Web et non sur les sites Web.
- L'algorithme vérifie l'existence d'options linguistiques dans la page d'accueil et compte chaque langue proposée en option.
- S'il n'y a pas d'options linguistiques, l'algorithme vérifie l'existence d'une autre langue que l'anglais dans la page d'accueil, si tel est le cas, il compte ce site Web dans cette langue au lieu de l'anglais.
- L'algorithme évalue un nombre approximatif de pages du site Web et multiplie chaque nombre de langues par ce nombre après avoir divisé par le nombre d'options de langue.

Dans Pimienta (2023) une tentative est faite pour *débiaiser* le chiffre de W3Techs pour les contenus en anglais, en évaluant le taux de multilinguisme de l'échantillonnage Tranco utilisé par W3Tech, et à partir de cela, en établissant la correction dans les résultats.

C'est une équation simple : $P' = (P - Err) / Rm$, où :

- ✓ P est le pourcentage donné W3Techs pour les contenus en anglais
- ✓ P' est le pourcentage *débiaisé* pour les contenus en anglais
- ✓ Err est le pourcentage de sites Web comptabilisés par erreur en anglais
- ✓ Rm est le taux de multilinguisme de l'échantillon.

À partir des données calculées, la fenêtre du pourcentage de contenus en anglais glisserait des 50 % - 60 %, annoncés par W3Techs vers les 20% - 30 % annoncés par l'Observatoire ou le consortium universitaire grec qui a étudié les ccTLD de l'UE.

[58]Selon https://news.netcraft.com/archives/category/web-server-survey, il y a en mai 2022 1,16 milliard de sites web dont 270 millions sont actifs. La couverture des plus visités est alors inférieure à 4% du total.
[59]Il en serait ainsi si les 270 millions de sites actifs proposaient ensemble plus de 400 millions d'interfaces linguistiques différentes, soit une moyenne de l'ordre de 1,5 langues par site.

L'Observatoire a encouragé les collègues grecs à appliquer leur algorithme à la liste de sites Tranco, avec une réponse prometteuse. Cela contribuerait de manière définitive à ce débat puisque leur méthode fait honneur au multilinguisme du web. C'est une perspective optimiste pour les mois à venir pour quiconque s'intéresse à ce sujet.

5. CONCLUSION

Pour la première fois dans l'histoire de l'Internet, une méthode est capable d'offrir une variété d'indicateurs significatifs sur la présence de 329 langues dans le Web. Le modèle fournit des résultats cohérents avec les études précédentes réalisées par l'Observatoire mais est en forte contradiction avec les résultats fournis par la source unique qui couvre le sujet depuis 2011 ; il montre en particulier que les contenus en anglais sur le Web sont aujourd'hui au même niveau que les contenus en chinois, autour de 20%, alors que les médias continuent de rapporter des contenus en anglais bien au-dessus de 50%.

La méthode utilisée pour obtenir ces résultats est exposée de manière complète et transparente et ses biais sont ouvertement discutés afin que la communauté scientifique puisse les analyser.

Ces résultats reflètent simplement une étape logique de l'évolution du Web, qui est passé d'une première phase centrée sur l'anglais (1992-2000), vers une deuxième étape centrée sur les langues européennes, avec un leadership anglais (2000-2010), suivie par une étape internationalisée, avec l'essor des langues asiatiques et arabes et encore un écart important laissant derrière elles les langues africaines, avec un Web chaque jour plus multilingue (2010-2020). L'étape à venir (2020-2030) verra probablement un Web plus homogène en termes de représentation des langues, avec, espérons-le, la fracture numérique commençant à être surmontée en Afrique, ouvrant l'espace des langues locales d'Afrique. L'enracinement du multilinguisme dans le Web est en marche et pourrait bien dépasser celui de l'humanité, si ce n’est déjà le cas. Cependant, les différences de productivité de contenue prévaudront, avec le maintien de certains avantages pour certaines langues ayant une combinaison d'une grande population L2 et d’une présence dans un grand nombre de pays (comme l'anglais et le français).

La surprise ne devrait pas venir des données produites par l’Observatoire qui ne sont que le reflet, dans sa composante cyber, de l'évolution naturelle du monde ; elles devraient provenir du fait que des données fortement biaisées ont été admises dans la dernière décennie sans grande réaction de la communauté scientifique.

Espérons que la transparence totale de la méthode aidera davantage d'esprits scientifiques à contester les résultats fournis par le monde du marketing et à repositionner ce thème là où il devrait appartenir : la communauté scientifique. Évidemment, cela inclut la remise en question de la méthode exposée ci-avant et la détection et la discussion d'éventuels biais qui n'ont pas été détectés par les auteurs. Laissons l'approche scientifique primer sur le marketing !

RÉFÉRENCES

Bauböck, R. (2015). The political value of languages. Crit. Rev. Int. Soc. Pol. Phil. 18, 212–223. doi: 10.1080/13698230.2015.1023635

Flint, C. (2021). Introduction to Geopolitics. Milton-park: Routledge.

Gazzola, M. (2015). Il Valore Economico Delle Lingue (The Economic Value of Languages). Available online at: https://ssrn.com/abstract=2691086 (accessed April 22, 2023).

Giannakoulopoulos, A., Pergantis, M., Konstantinou, N., Lamprogeorgos, A., Limniati, L., Varlamis, I. (2020). Exploring the dominance of the English language on the websites of EU countries. Fut. Int. 12, 76. doi: 10.3390/fi12040076

Grefenstette, G., and Noche, J. (2000). Estimation of English and Non-English Language use on the WWW. Rhone-Alpes: Xerox Research Centre Europe. Available online at: http://arxiv.org/ftp/cs/papers/0006/0006032.pdf

Grin, F., and Vaillancourt, F. (1997). The economics of multilingualism: overview and analytical framework. Annu. Rev. Appl. Linguist. 17, 43–65. doi: 10.1017/S0267190500003275

Heller, M. (2010). The Commodification of Language. Ann. Rev. Anthropol. 39, 101–114. doi: 10.1146/annurev.anthro.012809.104951

Lavoie, B. F., and O'Neill, E. T. (1999). How "World Wide" is the Web? Annual Review of OCLC Research.

Mikami, Y., Zavarsky, P., Rozan, M. Z. A., Suzuki, I., Takahashi, M., Mak, T., et al. (2005). The language observatory project (LOP). In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web. 990–991. Available online at: http://eprints.utm.my/id/eprint/3405/1/The_Language_Observatory_Project_%28LOP%29.pdf

Monrás, F., Medina, M., Cabré, S., Canto, P., Melendez, V., Ripoll, E., et al. (2006). Estadística de la presència del català a la xarxa d'Internet i de les característiques dels Webs Catalans, in Llengua i ús: Revista tècnica de política lingüística. Núm. 37, 62–66. Available online at: https://raco.cat/index.php/LlenguaUs/article/view/128275

O'Hara, K., and Hall, W. (2018). Four Internets: The Geopolitics of Digital Governance.Waterloo: Centre for International Governance Innovation.

Oliveira, G. M. (2010). O lugar das línguas. A América do Sul e os mercados linguísticos na nova economia. Brazil: Synergies Brésil. 21–30.

O'Neill, E. T., Lavoie, B. F., and Bennett, R. (2003). Trends in the Evolution of the PublicWeb: 1998 - 2002. Reston : D-Lib Magazine.

Pimienta, D. (2014). Le français dans l'Internet, Rapport 2014 "La langue française dans le monde". Nathan: OIF. 501.

Pimienta, D. (2021). Internet and Linguistic Diversity: The Cyber-Geography of Languages with the Largest Number of Speakers, LinguaPax Review 2021. Barcelona : Language Technologies and Language Diversity. 9–17. Available online at: https://www.linguapax.org/wp-content/uploads/2022/02/LinguapaxReview9-2021-low.pdf

Pimienta, D. (2022). Resource: Indicators on the Presence of Languages in Internet In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, Marseille. European Language Resources Association. 83–91. Available online at: https://aclanthology.org/2022.sigul-1.11/

Pimienta, D. (2023). Is it true that more than half the Web contents are in English? If Web multilingualism is paid due attention then no! ReseachGate Preprint. doi: 10.13140/RG.2.2.20767.43683

Pimienta, D., and Oliveira, G. M. (2022a). Cyber-Geography of Languages. Part 2: The Demographic Factor and the Growth of Asian Languages and Arabic. Alberta: International Review of Information Ethics. 32. Available online at: https://informationethics.ca/index.php/irie/article/view/488

Pimienta, D., and Oliveira, G. M. (2022b). Cyber-Geography of Languages. Part 1: Method, Results and Focus on English. Alberta: International Review of Information Ethics. 32. Available online at: https://informationethics.ca/index.php/irie/article/view/491

Pimienta, D., and Prado, D. (2016). Medición de la presencia de la lengua española en la Internet: métodos y resultados. Revista Española de Documentación Científica 39, e141. doi: 10.3989/redc.2016.3.1328

Pimienta, D., Prado, D., and Blanco, Á. (2009). Twelve Years of Measuring Linguistic Diversity on the Internet: Balance and Perspectives. Paris: UNESCO publications for the World Summit on the Information Society. Available online at: http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Simons, G. F., Thomas, A. L., and White, C. K. (2023). Assessing Digital Language Support on a Global Scale, In Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics. 4299–4305. Available online at: https://aclanthology.org/2022.coling-1.379.pdf

ANNEXE 1 : SOURCES POUR L'INDICATEUR USAGES

Table 5 : Réseaux sociaux sélectionnés et nombre total d'abonnés

RÉSEAU SOCIAL	TOTAL DES ABONNÉS (Million)
Whatsapp	2000
Wechat	1225
Tiktok	732
Douyin	600
Telegram	600
QQ	595
Snapchat	528
Weibo	521
Qzone	517
Kuaishou	481
Quora	300
Skype	300
Tieba	300
Viber	260
IMO	200
LINE	169
picsart	150
Likee	150
Discord	140
Twitch	140
Stack Exchange	100
VK	650
Odnoklassniki	200
Douban	200
MOJ	160
JOSH	115
ShareChat	160
FACEBOOK %utilisateurs par pays (NapoleonCat 2021)	1455
INSTAGRAM %utilisateurs par pays (NapoleonCat 2021)	1200
MESSENGER % d'utilisateurs par pays (NapoleonCat 2021)	1300
LINKEDIN % d'utilisateurs par pays (NapoleonCat 2021)	155
FACEBOOK Monde% de l'IWS 2021	1455
Linkedin %utilisateur par pays (ApolloTech 2021)	155
Twitter %utilisateurs par pays (Statista 2021)	396
% d'audience Pinterest (Statista 2021)	460
% d'utilisateurs de REDDIT par pays (Statista 2021)	430

Table 6 : Sources de données pour les réseaux sociaux

RÉSEAUX SOCIAUX	SOURCE
FACEBOOK %utilisateurs par pays (NapoleonCat 2021)	https://napoleoncat.com/stats/
INSTAGRAM %utilisateurs par pays (NapoleonCat 2021)	https://napoleoncat.com/stats/
MESSENGER % d'utilisateurs par pays (NapoleonCat 2021)	https://napoleoncat.com/stats/
LINKEDIN % d'utilisateurs par pays (NapoleonCat 2021)	https://napoleoncat.com/stats/
Linkedin %utilisateur par pays (ApolloTech 2021)	https://www.apollotechnical.com/linkedin-users-by-country/
Twitter %utilisateurs par pays (Statista 2021)	https://www.statista.com/statistiques/242606/ nombre-d-utilisateurs-actifs-de-twitter-dans-les-pays-sélectionnés/
FACEBOOK Monde% de l'IWS 2021	https://www.internetworldstats.com/stats1.htm + stats2.htm+…stats6.htm
Audience Facebook % (Statista 2021)	https://www.statista.com/statistiques/268136/ top-15-pays-basé-sur-nombre-d-utilisateurs-facebook/
YouTube % de connectés dans le pays (Statista 2021)	https://www.statista.com/statistics/1219589/youtube-penetration-worldwide-by-country/
% d'abonnés Netflix par pays (CompariTech 2020)	https://www.comparitech.com/tv-streaming/netflix-subscribers/
% d'audience Pinterest (Statista 2021)	https://www.statista.com/statistics/328106/pinterest-penetration-markets/
% d'utilisateurs de REDDIT par pays (Statista 2021)	https://backlinko.com/reddit-users
Cumul. 2012/21 % de téléchargements OpenOffice par pays	http://www.openoffice.org/stats/countries.html
# Serveurs Internet sécurisés	https://data.worldbank.org/indicator/IT.NET.SECR
% Abonné haut débit fixe dans le pays (WB 2021)	https://data.worldbank.org/indicator/IT.NET.BBND.P2
% Tél. fixe+abonnement mobile dans le pays (WB 2021)	https://data.worldbank.org/indicator/IT.MLT.MAIN.P2 + https://data.worldbank.org/indicator/IT.CEL.SETS.P2

ANNEXE 2 : ENCYCLOPÉDIES EN LIGNE ANALYSÉES

Table 7 : Encyclopédies en ligne

LANGUE	ENCYCLOPÉDIE	NOMBRE D'ARTICLES (Des millions)	LES AUTRES INFORMATIONS
Diverses	Encyclopédie de la vie (fin de vie)	0,75 (2010) 1.9 aujourd'hui	Langues prises en charge : arabe, portugais brésilien, anglais, finnois, français, macédonien, piémontais, chinois traditionnel et turc Langues d'interface : les mêmes plus l'allemand, l'espagnol, le néerlandais, le turc et l'ukrainien.
Diverses	thefreedictionary.com/ Gratuit avec publicité ou payant	pas de statistiques	Anglais, Espagnol, Allemand, Français, Italien, Chinois, Portugais, Néerlandais, Norvégien, Grec, Arabe, Polonais, Turc, Russe, Hébreu Il n'est pas clair s'il s'agit d'une version parallèle ou d'une langue spécifique.
Diverses	fr.metapedia.org/ version néonazie de wikipedia	Marginal (5000 articles en anglais)	Tchèque, Danois, Allemand, Espagnol, Anglais, Hongrois, Néerlandais, Portugais, Roumain, Slovène, Suédois, Estonien, Croate, Islandais, Norvégien, Macédonien
Chinois	Baidu Baiké	24,5	194 millions de modifications 7,5 millions d'éditeurs
Chinois	Baïké (Hudong)	18	5,8 millions d'éditeurs (2013)
Chinois	SogouBaike	????[60]	
Arabe	Marefa	0,136636	2,4 millions de pages
Arabe	Mawdoo3	0,15	45 (2018)
Bengali et Anglais	bengaldie	0,0057	1450 éditeurs
Croate	enciklopedija.hr	0,067	Données de la version d'impression
Croate	proleksis.lzmk.hr	0,062	
Danois	Den Store Danske	0,161	1100 éditeurs 1 million d'utilisateurs
Néerlandais	winklerprins.com	0,0115	par abonnement
Anglais	britannica.com		accès gratuit limité
Anglais	Everipedia Articles copiés de wikipedia	?	7000 éditeurs actifs (2019) Utilisateurs 3M (2017) accès libre mais aussi marché blockchain
Anglais	Citoyenneté	0,017	Statistiques arrêtées en 2014 proches de l'arrêt
Anglais	Conservapedia	0,0518	800 millions de pages vues 1,5 million de modifications
Anglais	Scholarpedia	0,0018	Données marginaux
Anglais	Encyclopédie.com	0,3	Agrégateur d'encyclopédie formelle
Anglais	Encyclopédie de la Colombie		Agrégé par Encyclopedia.com
Anglais	digitaluniverse.net		hors ligne
Français	Larousse	0,317	
Allemand	bavoir rétro	0,3	
Hébreu Anglais	Hamichlol	0,28	Version censurée de Wikipédia pour un public hyper-religieux
Coréen	Doopédia	0,588	
Sundanais Malais	Superpédia	0,02	

[60]Sogou Baike est considéré comme au moins aussi important que Baidu Baké et la même valeur du nombre d'articles a été supposée.

Javanais			
Italien	Treccani	0,9	
Malayalam	Sarvavijnanakosam	0,007	
Marathi	Viswakosh	0,016	
Norvégien Bokmal et Nynorsk	Magasin norske leksikon	0,2 (2019)	3 millions d'utilisateurs/mois lisent 500 000 articles
Polonais	encyclopedia.interia.pl	0,12 (2006)	
Polonais	encyclopedia.pwn.pl	0,08	
Russe	Grande Encyclopédie Russe	0,012 (2016)	
Russe	Krugosvet	0,012	
Espagnol	https://www.ecured.cu/ cubain	0,237	73 000 537 éditeurs actifs
Espagnol	Encyclonet	0,185	
Espagnol	enciclopedia.us.es/	0,053	https://wikiapiary.com/wiki/
Suédois	ne.se/	0,26 (2005)	
Tamil	pas en ligne		
Turc	Ekşi Sozluk	8M d'entrées en 2009[61]	400 000 utilisateurs 110 000 éditeurs 4M de nouvelles admissions/an en 2013[62] Ouverte à la publication, chaque entrée est conservée après modération.
Vietnamien	semble avoir disparu		Allez sur archive.org - https://bachkhoatoanthu.vass.gov.vn

[61] https://www.researchgate.net/publication/242100750_Web_Based_Authorship_in_the_Context_of_User_Generated_Content_An_Analysis_of_a_Turkish_Web_Site_Eksi_Sozluk

[62] https://www.researchgate.net/publication/271521393_SOCIAL_MEDIA_IN_A_DICTIONARY_FORMAT_ONLINE_COMMUNITY_OF_eksisozlukcom/figures?lo=1

ANNEXE 3 : SOURCES POUR L'INDICATEUR INTERFACE

Table 8 : Sources pour indicateur interface

Langues de traduction de Bing Translator	https://www.bing.com/translator/
Langues prises en charge par Amazon Kindle direct Publishing	https://kdp.amazon.com/en_US/help/topic/G200673300
Langues prises en charge par Cortana	https://en.wikipedia.org/wiki/Cortana
Langues de WordReference prises en charge	https://www.wordreference.com
Langues de traduction WordLingo	http://www.worldlingo.com/en/languages/
Langues prises en charge par Facebook	https://www.facebook.com/language.php
Langues des publicités Facebook InStream prises en charge	https://www.facebook.com/business/help/267128784014981
Langues du Free-translator prises en charge	http://www.free-translator.com
Langues prises en charge par la console Google Play	https://support.google.com/googleplay/android-developer/table/4419860?hl=fr
Langues prises en charge par Google Cloud	https://cloud.google.com/translate/docs/languages?hl=fr
Langues prises en charge par Google Traduction	https://en.wikipedia.org/wiki/Google_Translate
Langues prises en charge par Google Scholar	https://scholar.google.com/scholar_settings?sciifh=1&hl=fr&as_sdt=0,5#1
Langue prise en charge par Paralink Translator	http://paralink.com
Langues prises en charge par oinline-Translator	https://www.online-translator.com/traduction
Langues prises en charge par le traducteur Reverso	https://www.reverso.net/text_translation.aspx?lang=EN
Langues prises en charge par Free-Translations	https://www.freetranslations.org
Langues prises en charge par Skype	https://support.skype.com/en/faq/FA34781/what-languages-are-supported-in-skype
Langues prises en charge par Systran	https://support.systran.net/systranlinks/faq/

ANNEXE 4 : SOURCES POUR L'INDICATEUR INDEXES

Table 9 : Sources pour l'indicateur indexes

Index du gouvernement électronique	https://publicadministration.un.org/egovkb/Data-Center
Indice de participation électronique	https://publicadministration.un.org/egovkb/Data-Center
Index des services en ligne	https://publicadministration.un.org/egovkb/Data-Center
Indice du capital humain	https://publicadministration.un.org/egovkb/Data-Center
Indice des infrastructures de télécommunications	https://publicadministration.un.org/egovkb/Data-Center
Indice mondial de préparation numérique de Cisco 2019	https://www.cisco.com/c/dam/en_us/about/csr/reports/global-digital-readiness-index.pdf
Indice de préparation du gouvernement à l'IA 2020	https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf
Scores de liberté de l'Internet,	https://freedomhouse.org/countries/freedom-net/scores
Indice de connectivité mondiale	https://www.huawei.com/minisite/gci/en/country-rankings.html
Indice mondial de la cybersécurité 2018	https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf
Indice du commerce électronique B2C de la CNUCED, 2020	https://unctad.org/system/files/official-document/tn_unctad_ict4d17_en.pdf
L'indice mondial des données ouvertes	https://index.okfn.org/place/
Classement mondial de la compétitivité numérique 2020	https://www.imd.org/globalassets/wcc/docs/release-2020/digital/digital_2020.pdf
Indice de préparation pour Frontier Technologies	https://unctad.org/system/files/official-document/tir2020_en.pdf
Indice mondial de l'innovation	https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2020.pdf
Accès aux connaissances de base	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à l'information et aux communications	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à l'enseignement supérieur	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à l'électricité (% de la pop.)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à une éducation de qualité (0=inégal ; 4=égal)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à la gouvernance en ligne (0=faible ; 1=élevé)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Censure des médias (0=fréquent ; 4=rare)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Liberté d'expression (0=pas de liberté ; 1=pleine liberté)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Universités pondérées par la qualité (points)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Documents citables	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Femmes ayant fait des études supérieures	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Années d'études supérieures	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx

ANNEXE 5 : SÉLECTION DES SITES WEB POUR L'INDICATEUR TRAFIC

Table 10 : Sélection de sites Web pour l'indicateur trafic

WEBSITE	NUMBER OF TIMES
10086.cn	1
10jqka.com.cn	1
122.gov.cn	1
12306.cn	1
12371.cn	1
12377.cn	1
12388.gov.cn	1
1688.com	1
17ok.com	1
189.cn	1
1c-bitrix.ru	1
22.cn	1
24.kg	1
24h.com.vn	1
2gis.ru	1
300.cn	1
360.cn	1
4.cn	1
6.cn	1
66law.cn	1
999.md	1
abc.com.py	1
abc-communication.dz	1
abril.com.br	1
accuweather.com	2
activemind.de	1
actuniger.com	1
ad.iq	1
admin.ch	1
adminbuy.cn	1
Adobe.com	1
afrikmag.com	1
agenziaentrate.gov.it	1
ah.gov.cn	1
ahoraeg.com	1
ahram.org.eg	1
aktuality.sk	1
alakhbar.info	1
aliexpress.com	1
alipay.com	1
allegro.pl	1
allevents.in	1
almasryalyoum.com	1
alraziuni.edu.ye	1
alwakeelnews.com	1
alwatanvoice.com	1
amazon.ae	1
amazon.ca	1
amazon.cn	1
amazon.co.jp	1
amazon.co.uk	1
amazon.com	20
amazon.com.br	1
amazon.de	1
amazon.eg	1
amazon.es	1
amazon.fr	1
amazon.in	1
amazon.it	1
ameblo.jp	1
amritmahotsav.nic.in	1
andersnoren.se	1
anpc.gov.ro	1
anyxxx.com	1
ap.gov.in	1
aparat.com	2
apple.com	5
arabiaweather.com	1
argentina.gob.ar	1
aruba.it	1
autohome.com.cn	1
avaz.ba	1
babytree.com	1
baharain.bh	1
baidu.com	4
band.us	1
bangladesh.gov.bd	1

bankmandiri.co.id	1
bayern.de	1
bb.com.br	1
bbc.co.uk	1
bbc.com	1
bbc.in	1
bcel.com.la	1
beian.gov.cn	1
beijing.gov.cn	1
belgium.be	1
belizebank.com	1
belonnanotservice.ga	2
bet365.com	1
bgeneral.com	1
bih.nic.in	1
Bing.com	3
biobiochile.cl	1
bitrix24.ru	1
bjx.com.cn	1
blogger.com	3
bnonline.fi.cr	1
boc.cn	1
Bongacams.com	3
borneobulletin.com.bn	1
bri.co.id	1
britannica.com	2
bshare.cn	1
bt.bt	1
bt.cn	1
bukalapak.com	1
bund.de	1
businessday.ng	1
businessinsider.in	1
businesstoday.in	1
businessworld.in	1
cac.gov.cn	1
cafebazaar.ir	1
caixa.gov.br	1
cambridge.org	1
Canva.com	5
cao.ir	1
careers.sl	1
cas.cn	1
cbec.gov.in	1
cbic.gov.in	1
cbos.gov.sd	1
cbse.gov.in	1
cbse.nic.in	1
ccdi.gov.cn	1
ccgp.gov.cn	1
ccm.gov.cn	1
ce.cn	1
centrafrique-presse.over-blog.com	1
chase.com	1
chaturbate.com	2
china.cn	1
china.com.cn	1
chinadaily.com.cn	1
chinanews.com.cn	1
chinatax.gov.cn	1
chsi.com.cn	1
cib.com.cn	1
cmbc.com.cn	1
cmseasy.cn	1
cnil.fr	1
cninfo.com.cn	1
cnipa.gov.cn	1
cnnindonesia.com	1
cnpq.br	1
cnr.cn	1
cntv.cn	1
coinmarketcap.com	2
comores-infos.net	1
conac.cn	1
consultant.ru	1
coremail.cn	1
correios.com.br	1
corriere.it	1
coupa.ng	1
court.gov.cn	1
covid19.go.id	1
cowin.gov.in	1
cpdp.bg	1
cq.gov.cn	1
creditchina.gov.cn	1
cri.cn	1
cricbuzz.com	1
cro.ma	1
csdn.net	1
csrc.gov.cn	1

customs.gov.cn	1
cvc.nic.in	1
cyberpolice.cn	1
dahe.cn	1
dailypost.ng	1
dakaractu.com	1
daraz.pk	1
data.gov.in	1
dataprotection.gov.cy	1
daum.net	1
defimedia.info	1
detik.com	1
dg-datenschutz.de	1
dictionary.com	1
digikala.com	1
digitalindia.gov.in	1
dinesh-ghimire.com.np	1
discord.com	1
ditaduraeconsenso.blogspot.com	1
dlszywz.cn	1
dns4.cn	1
docdro.id	1
dpboss.net	1
dr.dk	1
draugiem.lv	1
duckduckgo.com	1
dwz.cn	1
e.gov.kw	1
ebay.com	1
ebay.de	1
ebs.org.cn	1
eci.gov.in	1
education.gov.in	1
elcomercio.com	1
eldeber.com.bo	1
elnuevodia.com	1
elpais.com	1
elsalvador.com	1
eluniverso.com	1
emansion.gov.lr	1
emploi.cg	1
ems.com.cn	1
enamad.ir	1
enimerotiko.gr	1
eol.cn	1

e-recht24.de	1
ernet.in	1
espn.com	1
espncricinfo.com	1
estadao.com.br	1
eta.gov.lk	1
ethiojobs.net	1
etnet.com.hk	1
facebook.com	80
facebook.com.br	1
fandom.com	3
fazenda.gov.br	1
fijivillage.com	1
file-upload.com	1
findlaw.cn	1
firefox.com.cn	1
fiverr.com	1
flipkart.com	1
flydubai.com	1
fmprc.gov.cn	1
focus.cn	1
follow.it	1
Force.com	1
free.fr	1
freebitco.in	1
freeindianporn2.com	1
freepik.com	1
fs.fed.us	1
ftc.go.kr	1
fujian.gov.cn	1
gansu.gov.cn	1
garanteprivacy.it	1
gd.gov.cn	1
geni.us	1
gesetze-im-internet.de	1
ghanaweb.com	1
gismeteo.ru	1
globo.com	1
gmw.cn	1
gogo.mn	1
gome.com.cn	1
goo.ne.jp	1
google,com	1
google.ad	1
google.ae	1

google.at	1
google.az	1
google.be	1
google.bf	1
google.bg	1
google.ca	2
google.cd	1
google.cg	1
google.ch	1
google.ci	1
google.cl	1
google.cn	1
google.co.id	1
google.co.il	1
google.co.in	1
google.co.jp	1
google.co.ke	1
google.co.kr	1
google.co.ma	1
google.co.mz	1
google.co.nz	1
google.co.th	1
google.co.tz	1
google.co.ug	1
google.co.uk	1
google.co.uz	1
google.co.ve	1
google.co.za	1
google.co.zm	1
google.co.zw	1
google.com	146
google.com.af	1
google.com.ar	1
google.com.bd	1
google.com.bn	1
google.com.bo	1
google.com.br	1
google.com.bz	1
google.com.co	1
google.com.cu	1
google.com.do	1
google.com.eg	1
google.com.hk	2
google.com.jm	1
google.com.kw	1
google.com.lb	1
google.com.ly	1
google.com.mm	1
google.com.mt	1
google.com.mx	1
google.com.na	1
google.com.ng	1
google.com.ni	1
google.com.np	1
google.com.om	1
google.com.pa	1
google.com.pe	1
google.com.pg	1
google.com.ph	1
google.com.pk	1
google.com.pr	1
google.com.py	1
google.com.qa	1
google.com.sa	1
google.com.sb	1
google.com.sg	1
google.com.sl	1
google.com.sv	1
google.com.tj	1
google.com.tr	1
google.com.tw	1
google.com.ua	1
google.com.uy	1
google.com.vn	1
google.de	1
google.dj	1
google.dk	1
google.dz	1
google.ee	1
google.es	2
google.fr	3
google.ge	1
google.gr	1
google.gy	1
google.hn	1
google.ht	1
google.ie	1
google.iq	1
google.is	1
google.it	1

google.jo	1
google.kg	1
google.kz	1
google.la	1
google.lk	1
google.lt	1
google.lu	1
google.lv	1
google.md	1
google.me	1
google.mg	1
google.mk	1
google.ml	1
google.mn	1
google.mw	2
google.nl	1
google.no	1
google.pl	1
google.ps	1
google.pt	1
google.ro	1
google.rs	1
Google.ru	3
google.rw	1
google.se	1
google.si	1
google.sk	1
google.sn	1
google.so	1
google.sr	1
google.st	1
google.td	1
google.tg	1
google.tl	1
google.tm	1
google.tn	1
google.tt	1
gosuslugi.ru	1
gov.bw	1
gov.ls	1
govtrack.us	1
grid.id	1
grupobancolombia.com	1
gst.gov.in	1
gsxt.gov.cn	1

guardian.co.tt	1
guardian.ng	1
gujarat.gov.in	1
gxzf.gov.cn	1
gz.gov.cn	1
haberler.com	1
hainan.gov.cn	1
haosou.com	1
hatena.ne.jp	1
hd315.gov.cn	1
hdfcbank.com	1
healthline.com	1
heartland.us	1
henan.gov.cn	1
herald.co.zw	1
hi.is	1
hindustantimes.com	1
homedepot.com	1
hoster.kz	1
hotlog.ru	1
hotpepper.jp	1
hotstar.com	2
huanqiu.com	1
hubei.gov.cn	1
hunan.gov.cn	1
hurriyet.com.tr	1
ibps.in	1
ibw.cn	1
icbc.com.cn	1
icicibank.com	1
icio.us	1
ico.org.uk	1
idnes.cz	1
iitb.ac.in	1
iitkgp.ac.in	1
ijavhd.com	1
imageshack.us	1
imdb.com	2
imjo.in	1
in.gr	1
incometax.gov.in	1
incometaxindia.gov.in	1
incometaxindiaefiling.gov.in	1
index.hr	1
index.hu	1

india.com	1
india.gov.in	1
indiamart.com	1
indianrailways.gov.in	1
indianvisaonline.gov.in	1
indiapost.gov.in	1
indiatimes.com	1
indiatoday.in	1
inflibnet.ac.in	1
instagram.com	47
instructure.com	1
intoday.in	1
iol.co.za	1
ionos.de	1
iplt20.com	1
irctc.co.in	1
irembo.gov.rw	1
irna.ir	1
is.fi	1
isna.ir	1
itau.com.br	1
jamaica-gleaner.com	1
japanpost.jp	1
jc001.cn	1
Jd.com	1
jiangsu.gov.cn	1
jiangxi.gov.cn	1
jiji.ng	1
jl.gov.cn	1
jne.co.id	1
jotform.us	1
jrj.com.cn	1
jumia.ci	1
jumia.com.ng	1
juraforum.de	1
justindianporn.me	1
kancloud.cn	1
kar.nic.in	1
karnataka.gov.in	1
kaskus.co.id	1
kemdikbud.go.id	1
kemenag.go.id	1
kemkes.go.id	1
kenh14.vn	1
kerala.gov.in	1

khaberni.com	1
knet.cn	1
knetreg.cn	1
kominfo.go.id	1
kompas.com	1
kriesi.at	1
kuaishang.cn	1
kuenselonline.com	1
kumparan.com	1
kupujemprodajem.com	1
lanouvelletribune.info	1
laodong.vn	1
laprensa.com.ni	1
laprensa.hn	1
lawtime.cn	1
lazada.co.id	1
leader.ir	1
lefigaro.fr	1
legifrance.gouv.fr	1
legit.ng	1
lemonde.fr	1
lex.uz	1
licindia.in	1
line.me	2
linkd.in	1
linkedin.com	13
liputan6.com	1
list.am	1
listindiario.com	1
live.com	19
liveinternet.ru	1
livroreclamacoes.pt	1
lnkd.in	1
ltn.com.tw	1
ltt.ly	1
m.in	1
macaodaily.com	1
mahaonline.gov.in	1
maharashtra.gov.in	1
mail.ru	2
mana.pf	1
mastercard.us	1
mayoclinic.org	2
medcol.mw	1
mediacongo.net	1

mercadolibre.cl	1
mercadolibre.com.co	1
mercadolibre.com.ve	1
mercadolivre.com.br	1
merdeka.com	1
merriam-webster.com	1
meskerem.net	1
meteo.nc	1
metruyenchu.com	1
mhlw.go.jp	1
microsoft.com	25
microsoftonline.com	4
milliyet.com.tr	1
mk.by	1
mof.gov.tl	1
moh.go.tz	1
moip.gov.mm	1
mol.gov.om	1
monetizze.com.br	1
msn.com	3
myshopify.com	5
namibian.com.na	1
namnak.com	1
naver.com	1
ncdc.gov.ng	1
nessma.tv	1
netafrique.net	1
netflix.com	13
nethouse.ru	1
nettruyengo.com	1
news24.com	1
niagahoster.co.id	1
notion.so	1
novinky.cz	1
nsw.gov.au	1
nzherald.co.nz	1
odnoklassniki.ru	1
office.com	8
ok.ru	3
okezone.com	1
onlinehome.us	1
orange.fr	1
orient.tm	1
otr.tg	1
ouest-france.fr	1

oxu.az	1
ozon.ru	1
pagcor.ph	1
pagesjaunes.fr	1
paypal.com	1
paystack.com	1
pikiran-rakyat.com	1
pinterest.com	11
pinterest.de	1
pinterest.es	1
pinterest.fr	1
pinterest.it	1
pixnet.net	1
planalto.gov.br	1
pornhub.com	4
portaldoconhecimento.gov.cv	1
post.ir	1
postcourier.com.pg	1
postimees.ee	1
premierbet.co.ao	1
premierleague.com	1
prensa-latina.cu	1
prensalibre.com	1
presidence.gov.bi	1
president.ir	1
president.tj	1
prom.st	1
prom.ua	1
public.lu	1
pulse.ng	1
punchng.com	1
qq.com	2
r01.ru	1
rae.es	1
rakuten.co.jp	1
rambler.ru	1
reddit.com	9
reg.ru	1
repubblica.it	1
republika.co.id	1
ria.ru	1
rijksoverheid.nl	1
rt.com	1
rte.ie	1
rtvslo.si	1

s.id	1
sabay.com.kh	1
sacoronavirus.co.za	1
sahibinden.com	1
sakura.ne.jp	1
salesforce.com	1
salla.sa	1
sana.sy	1
sante.gov.dz	1
sante.gov.gn	1
sapo.pt	1
sapp.ir	1
saude.gov.br	1
scielo.br	1
sekolahku.web.id	1
seneweb.com	1
serveriai.lt	1
service-public.fr	1
setn.com	1
seznam.cz	1
shopee.co.id	1
shopee.co.th	1
shopee.tw	1
shopee.vn	1
shop-pro.jp	1
singaporepools.com.sg	1
smarturl.it	1
sohu.com	2
solomonstarnews.com	1
soy502.com	1
spiegel.de	1
stackoverflow.com	1
standardmedia.co.ke	1
state.co.us	1
state.fl.us	1
state.il.us	1
state.ma.us	1
state.md.us	1
state.mn.us	1
state.nj.us	1
state.nm.us	1
state.nv.us	1
state.ny.us	1
state.oh.us	1
state.or.us	1

state.pa.us	1
state.tx.us	1
suara.com	1
sucursalelectronica.com	1
suribet.sr	1
sympla.com.br	1
syri.net	1
t.me	2
taobao.com	2
theguardian.com	1
thethao247.vn	1
tiktok.com	10
time.mk	1
times.co.sz	1
timesofmalta.com	1
timeweb.ru	1
tmall.com	1
tokopedia.com	1
t-online.de	1
tradingview.com	2
trendyol.com	1
tribunnews.com	1
tripadvisor.com.br	1
tripadvisor.fr	1
tripadvisor.it	1
turkiye.gov.tr	1
twitch.tv	5
twitter.com	32
ucoz.ru	1
uem.mz	1
ultimahora.com	1
uol.com.br	1
ura.go.ug	1
usp.br	1
vanguardngr.com	1
vg.no	1
vk.com	7
vkontakte.ru	1
vnexpress.net	1
walmart.com	1
wbs-law.de	1
weather.com	1
webmd.com	1
whatsapp.com	22
wikipedia.org	29

wiktionary.org	2
wildberries.ru	1
wizard.id	1
www.gob.mx	1
www.gob.pe	1
www.gov.br	1
www.gov.pl	1
www.gov.uk	1
xhamster.com	1
xnxx.com	3
xosodaiphat.com	1
xvideos.com	6
yahoo.co.jp	1

yahoo.com	25
yandex.ru	5
yasour.org	1
yelp.com	1
ynet.co.il	1
youm7.com	1
youtu.be	1
youtube.com	103
youtube.org	1
zalo.me	1
zambiaimmigration.gov.zm	1
zhizhuchi.cm	1
zoom.us	15

ANNEXE 6 : MACRO-LANGUES

Tel que défini par Ethnologue.

Table 11 : Liste des macro-langues

CODE ISO	*MACRO LANGUES*	*NOMBRE DE LANGUES FUSIONNÉES*
ara	**Arabe**	***29***
aym	**Aymara**	***2***
aze	**Azerbaïdjanais**	***3***
bal	**Baloutche**	***3***
bik	**Bikol**	***8***
bnc	**Bontok**	***5***
bua	**Bouriate**	***3***
chm	**Mari**	***2***
cre	**Cri**	***6***
del	**Delaware**	***2***
den	**Slave (Athapascan)**	***2***
din	**Dinka**	***5***
doi	**Dogri**	***2***
est	**Estonien**	***2***
fas	**Persan**	***2***
ful	**Fulfulde**	***9***
gba	**Gbaya**	***6***
gon	**Gondi**	***3***
grb	**Grébo**	***5***
grn	**Guarani**	***5***
hai	**Haïda**	***2***
hbs	**Serbo-croate**	***4***
hmn	**Hmong**	***25***
iku	**Inuktitut**	***2***
ipk	**Inupiatun**	***2***
jrb	**Judéo-arabe**	***5***
kau	**Kanuri**	***3***
kln	**Kalenjin**	***9***
kok	**Konkani**	***2***
kom	**Komis**	***2***
kon	**Congo**	***3***
kpe	**Kpell**	**2**
kur	**Kurde**	***3***
lah	**Lahnda**	***7***
lav	**Letton**	***2***
luy	**Luiya**	***14***
man	**Mandingue**	***6***
mlg	**Malgache**	***11***
mon	**Mongol**	***3***
msa	**Malais**	***36***
mwr	**Marwari**	***6***
nep	**Népalais**	***2***
oji	**Ojibwé**	***sept***
ori	**Oriya**	***2***
orm	**Galla**	***4***
pus	**Pachtou**	***3***
que	**Quechua**	***42***
raj	**Rajasthan**	**6**
rom	**Romani**	***6***
sqi	**Albanais**	***4***
srd	**Sarde**	***4***
swa	**Swahili**	***2***
syr	**Syriaque**	***2***
tmh	**Tamashek**	***4***
uzb	**Ouzbek**	***2***
yid	**Yiddish**	***2***
zap	**Zapotèque**	***57***
zha	**Zhuang**	***16***
zho	**Chinois**	***15***
zza	**Dimli**	***2***

ANNEXE 7 : LISTE DES PAYS OU TERRITOIRES SANS DONNÉES UIT

Table 12 : Liste des pays sans données UIT

Code ISO	NOM DU PAYS	POPULATION
AX	**Île d'Åland**	**27 652**
AS	**Samoa américaines**	**55 990**
IO	**Territoire britannique de l'océan Indien**	**4 000**
QB	**Pays-Bas caribéens**	**18 740**
CX	**L'île de noël**	**1 170**
CC	**Îles Cocos (Keeling)**	**630**
CK	**Les Îles Cook**	**15 000**
CW	**Curaçao**	**140 000**
GF	**Guyane Française**	**366 590**
GP	**Guadeloupe**	**454 800**
GU	**Guam**	**139 550**
IM	**Île de Man**	**88 085**
QM	**Martinique**	**377 100**
NC	**L'ile de Norfolk**	**1 500**
KP	***Corée du Nord***	***25 579 000***
PM	**Îles Mariannes du Nord**	**53 280**
PW	**Palaos**	**17 550**
PN	**Pitcairn**	**36**
RE	**Réunion**	**751 580**
BL	**Saint Barthélemy**	**7 850**
FM	**Saint Martin**	**28 500**
PM	**Saint-Pierre-et-Miquelon**	**6 340**
SX	**Saint Martin**	**33 470**
CT	**Îles Turques-et-Caïques**	**30 170**
GO	***État du Vatican***	***330***
HE	***Sahara occidental***	***544 150***
	TOTAL	**28 689 463**

Il existe deux raisons possibles pour lesquelles le pays ou le territoire est exclu des données de l'UIT :

1) C'est un territoire dont les données sont reprises avec celle d'un autre pays
2) Il n'y a pas de source ou d'estimation du pourcentage de personnes connectées à l'Internet (en italique dans le tableau).

ANNEXE 8 : SOURCES SUR LE COMPORTEMENT LINGUISTIQUE DES INTERNAUTES

https://motsdici.be/wp-content/uploads/2019/04/Article-cant-read-wont-buy.pdf
Rapport consultatif de Common Sense 2006 « *Si je ne peux pas lire je n'achète pas* ».

https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556
Rapport d'enquête 2011 de l'Union européenne "*Agenda numérique : plus de la moitié des internautes de l'UE utilisent une langue étrangère lorsqu'ils sont en ligne*"
Citation : "*Alors que 90 % des internautes de l'UE préfèrent accéder aux sites Web dans leur propre langue, 55 % utilisent au moins occasionnellement une langue autre que la leur lorsqu'ils sont en ligne, selon un Eurobaromètre paneuropéen*".

https://hbr.org/2012/08/speak-to-global-customers-in-t
Harvard Business Review 2012 : « Parlez aux clients internationaux dans leur propre langue »
Citation : "72,1 % des consommateurs passent la plupart ou la totalité de leur temps sur des sites Web dans leur propre langue"

https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf
Étude KPMG/Google 2017 "*Langues indiennes - Définir l'Internet indien*"
Citation : « *Les utilisateurs de l'Internet, en langues indiennes devraient représenter près de 75 % de la base d'utilisateurs de l'Internet, en Inde en 2021*. »

https://insights.csa-research.com/reportaction/305013126/Marketing
https://csa-research.com/Blogs-Events/CSA-in-the-Media/Press-Releases/Consumers-Prefer-their-Own-Language
Rapport de recherche CSA 2020 "*Si je ne peux pas lire, je n'achèterai pas - Analyse B2C des préférences linguistiques et des comportements des consommateurs dans 29 pays*"
Citation : « *Une enquête menée auprès de 8 709 consommateurs dans 29 pays révèle que 76 % préfèrent acheter des produits contenant des informations dans leur propre langue ».*

https://octopustranslations.com/e-commerce-and-the-impact-of-language-on-consumer-behavior/
Rapport Octopus Translation 2021 *Le commerce électronique et l'impact de la langue sur le comportement des consommateurs*
Citation : « *55 % des consommateurs dans le monde effectuent leurs achats en ligne uniquement dans leur langue maternelle ».*

https://www.businesswire.com/news/home/20211026005375/en/Unbabel%E2%80%99s-2021-Global-Multilingual-CX-Survey-Reveals-68-of-Consumers-Prefer-to-Speak- avec-les-marques-dans-leur-langue-natale
Rapport BusinessWire 2021 "*L'enquête mondiale CX multilingue 2021 d'Unbabel révèle que 68 % des consommateurs préfèrent parler aux marques dans leur langue maternelle*"

https://www.prweb.com/releases/2014/04/prweb11725995.htm
2022.PRWeb Market Research « *Une enquête auprès de 3 000 acheteurs en ligne dans 10 pays révèle que 60% achètent rarement ou jamais sur des sites Web uniquement en anglais*».

ANNEXE 9 : RESULTATS SÉPARÉS POUR L1 ET L2

Comme méthode de recoupement de la validité du modèle, qui est basé sur des données démolinguistiques L1+L2, deux passages supplémentaires ont été réalisés, un avec les données L1 uniquement et un autre avec les données L2 uniquement.

Table 13 : Modèle exécuté avec L1 uniquement

		INTERNAUTES L1	POPULATION L1	LOCUTEURS L1	CONTENUS L1	PRESENCE VIRTUELLE	PRODUCTIV. CONTENUS
1	Chinois	22,34 %	18,33 %	71,18%	**25,55 %**	**1.39**	**1.14**
2	Anglais	7,82 %	5,12 %	89,24 %	**12,96 %**	**2.53**	**1,66**
3	Espagnol	8,14 %	6,52 %	72,95 %	**8,76 %**	**1.34**	**1.08**
4	Arabe	5,33%	4,80 %	64,91 %	**4,15 %**	**0,86**	**0,78**
5	Portugais	3,91 %	3,21 %	70,99 %	**3,91 %**	**1.22**	**1,00**
6	Japonais	2,77 %	1,75 %	92,63%	**3,47 %**	**1,99**	**1.25**
7	Russe	3,00 %	2,13%	82,36%	**3,22 %**	**1.51**	**1.07**
8	Hindi	3,35 %	4,73 %	41,34%	**2,93 %**	**0,62**	**0,88**
9	Français	1,59 %	1,10 %	84,59%	**2,08 %**	**1,89**	**1.31**
10	Allemand	1,62 %	1,06 %	89,51%	**1,96 %**	**1,85**	**1.21**

Si l'on ne considère que les locuteurs de première langue, le français serait en position 9 et assez logiquement le chinois afficherait un gros avantage sur l'anglais, malgré sa très grande présence virtuelle et la productivité de son contenu. La présence virtuelle et la productivité des contenus pour le français sont très élevées, malgré cette neuvième place.

Table 14 : Modèle exécuté avec L2 uniquement

		INTERNAUTES L2	POPULATION L2	LOCUTEURS L2	CONTENUS L2	PRESENCE VIRTUELLE	PRODUCTIV. CONTENUS
1	Anglais	32,53%	31,25%	55,64 %	**37,91 %**	**1.21**	**1.17**
2	Chinois	8,68 %	6,38 %	72,65%	**10,68%**	**1,67**	**1.23**
3	Français	6,47 %	5,99 %	57,81 %	**6,90 %**	**1.15**	**1.07**
4	Hindi	6,32 %	8,25 %	40,93 %	**5,96 %**	**0,72**	**0,94**
5	Espagnol	3,37 %	2,28%	78,82 %	**5,47 %**	**2.39**	**1.62**
6	Russe	4,82 %	3,33%	77,32%	**5,12 %**	**1,54**	**1.06**
7	Malais	5,37 %	5,21 %	55,08 %	**4,52 %**	**0,87**	**0,84**
8	Allemand	3,10 %	1,87 %	88,72 %	**3,61 %**	**1,93**	**1.17**
9	Thaïlandais	1,86 %	1,28 %	77,84 %	**1,55 %**	**1.21**	**0,83**
10	Ourdou	1,81 %	5,15 %	18,86 %	**1,15 %**	**0,22**	**0,63**
11	Portugais	0,68 %	0,81 %	44,81%	**0,89 %**	**1.10**	**1.32**

Si l'on ne considère que les locuteurs de langue seconde, l'anglais prend logiquement la première place et le français la troisième place devant l'espagnol.

Pour rappel, voici les résultats pour L1+L2.

Table 15 : Résultats du modèle pour L1+L2

		INTERNAUTES L1+L2	POPULATION L1+L2	LOCUTEURS L1+L2	CONTENUS L1+L2	PRESENCE VIRTUELLE	PRODUCTIV. CONTENUS
1	Chinois	18,46 %	14,72 %	71,38%	**21,60%**	**1,47**	**1.17**
2	Anglais	14,83 %	13,01 %	64,86 %	**19,60 %**	**1.51**	**1.32**
3	Espagnol	6,79 %	5,24 %	73,72%	**7,85 %**	**1,50**	**1.16**
4	Hindi	4,19 %	5,80 %	41,16%	**3,76 %**	**0,65**	**0,90**
5	Russe	3,51 %	2,49 %	80,32%	**3,76 %**	**1.51**	**1.07**
6	Français	2,98 %	2,58 %	65,80 %	**3,33%**	**1.29**	**1.12**
7	Portugais	2,99 %	2,49 %	68,43%	**3,13%**	**1.26**	**1.05**
8	Arabe	3,97 %	3,53 %	63,99 %	**3,09 %**	**0,87**	**0,78**
9	Japonais	1,99 %	1,22 %	92,63%	**2,66 %**	**2.18**	**1.34**
10	Allemand	2,04 %	1,30%	89,17%	**2,37 %**	**1,82**	**1.16**

Un contrôle de cohérence entre les 3 résultats est effectué, le troisième devant découler logiquement des deux premiers.

Table 16 : Contrôle des résultats L1 et L2

	POP. Mondiale	POP. Connectée	% Pop. Connect	Pop. Anglais	Pop. Connect. Anglais	% Pop. Connect. Anglais	Contrôle
L1	**7 231 699 136**	**4 223 428 027**	**58,40%**	5,12 %	7,82 %	89,24 %	89,24 %
L2	**3 130 017 620**	**1 673 121 762**	**53,45%**	31,25%	32,53%	55,64 %	55,64 %
L1+L2	**10 361 716 756**	**5 896 549 789**	**56,91%**	13,01 %	14,83 %	64,86 %	64,86 %
Contrôle			56,91 %	13,01 %	14,83 %	64,86 %	

En vert les vérifications sont effectuées : il s'agit de calculer directement les mêmes valeurs et donc de vérifier que les deux modèles L1 et L2 ont fonctionné correctement : la preuve est faite.

La deuxième série de contrôles est plus complexe et il ne faut pas s'attendre à des correspondances parfaites (car la modélisation n'est pas un processus linéaire par rapport aux données démolinguistiques).

Table 17 : Vérification des résultats L1 et L2 (suite)

	Anglais	Chinois	Espagnol	Français	Hindi	Portugais	Russe	Allemand
Contenus L1	12,96 %	25,55 %	8,76 %	2,08 %	2,93 %	3,91 %	3,22 %	1,96 %
Contenus L2	37,91 %	10,68%	5,47 %	6,90 %	5,96 %	0,89 %	5,12 %	3,61 %
Contenus L1+L2	19,60 %	21,60%	7,85 %	3,33%	3,76 %	3,13%	3,76 %	2,37 %
Contrôle	20,04 %	21,33%	7,83 %	3,45 %	3,79 %	3,05 %	3,76 %	2,43 %

Les trois premières lignes montrent les résultats des trois modèles respectifs. La ligne de contrôle en vert est calculée en pondérant les pourcentages respectifs L1 et L2 par rapport aux populations connectées respectives. Ainsi, pour l'anglais, 20,04 % est obtenu par la formule suivante : ((12,96 x 4 233 428 027) + (37,91 x 1 673 121 762)) / 5 896 549 789

Il est à la fois remarquable et très rassurant, quant à la validité du modèle, que les résultats obtenus par les deux méthodes (le modèle L1+L2 ou le prorata des résultats des modèles L1 et L2 par rapport aux populations connectées respectives) soient si proches.