Nouvelle version améliorée d'une approche alternative pour produire des indicateurs de la présence des langues dans l'Internet

Daniel Pimienta
Observatoire de la diversité linguistique et culturelle dans l'Internet
http://funredes.org/lc

Août 2021

Crédits: Les travaux menant à cette version améliorée ont été rendus possibles grâce au soutien du *Département culturel et éducatif du ministère brésilien des Affaires étrangères* dans le cadre l'Institut international de la langue portugaise et sous la coordination de la Chaire UNESCO sur les politiques linguistiques pour le multilinguisme. Des crédits sont également accordés à Daniel Prado, qui le premier a eu l'idée de collecter des sources multiples pour mesurer la présence des langues dans l'Internet ainsi que pour transformer des données par pays en données par langue.

Remerciements: Au professeur Gilvan Müller de Oliveira pour le soutien sur les questions linguistiques et la coordination avec les bailleurs de fonds; à Alvaro Blanco pour la rédaction des programmes qui ont changé radicalement la gestion de tant de sources et d'orthographes de langues et de pays et à David Pimienta qui a écrit les programmes nécessaires pour transformer le format Ethnologue au format requis pour cette étude et pour le traitement des macro-langues.

Avertissement: L'étude qui suit est essentiellement un travail statistique basé sur une grande variété de sources. L'adoption d'une source majeure dans ce type de travaux implique également et logiquement l'adoption des règles soutenant les données de cette source. L'auteur n'est pas responsable de la liste des pays et territoires considérés, établie par l'UIT, une agence des Nations Unies, ni de la liste des langues comptant plus de cinq millions de locuteurs L1, selon Ethnologue, ainsi que pour le regroupement en macro-langues, adopté par Ethnologue, en accord avec la norme ISO 693.3.

RESUMÉ

Dans un contexte de rareté de données fiables sur l'espace des langues dans l'Internet, l'approche alternative de 2017 pour la production d'indicateurs du comportement dans l'Internet des 140 langues avec plus de 5 millions de locuteurs, a été enrichie et actualisée. Les améliorations de cette approche, laquelle est basée sur la collecte d'une large série de micro-indicateurs des langues ou des pays dans divers espaces ou applications de l'Internet (ou en relation avec l'Internet) sont exposés. L'utilisation des dernières données produites par Ethnologue a permis de disposer des chiffres démolinguistiques les plus fiables et actualisés et également de fournir les éléments pour surmonter l'un des biais majeurs de la méthode, lié au traitement des locuteurs L2. Les six indicateurs de la présence des langues dans l'Internet qui ont été définis et instruits en 2017 (i*nternautes, trafic, usage, contenus, index sociétaux et interfaces*), et les quatre macro-indicateurs qui en sont déduits *(puissance, capacité, gradient et productivité des contenus*) sont reproduits, après mises à jour pour 2021 de toutes les sources. Les résultats montrent une présence de l'anglais en diminution relative, autour de 25 % (versus 30% en 2017) et le chinois en forte augmentation, tandis que l'espagnol se conforte en troisième position. Le français partage désormais la quatrième place avec l'hindi, avec une avance réduite, par rapport à 2017, sur un groupe de langues aux positions très proches : portugais, russe, arabe et allemand. Comme dans l'édition 2017, tous les biais possibles dérivés de la méthode, des hypothèses ou des sources sont examinés et une estimation est proposée qui prend en compte ces biais, pour les langues de majeure puissance. Il est prévu pour fin 2021 un nouvel ensemble d'améliorations avec la possibilité d'étendre les résultats pour les 332 langues avec plus de 1 million de locuteurs L1.

Mots clés: Langues, Internet, diversité linguistique, indicateurs, biais

Contenu

LISTE DES TABLEAUX ET DES FIGURES

LES TABLES

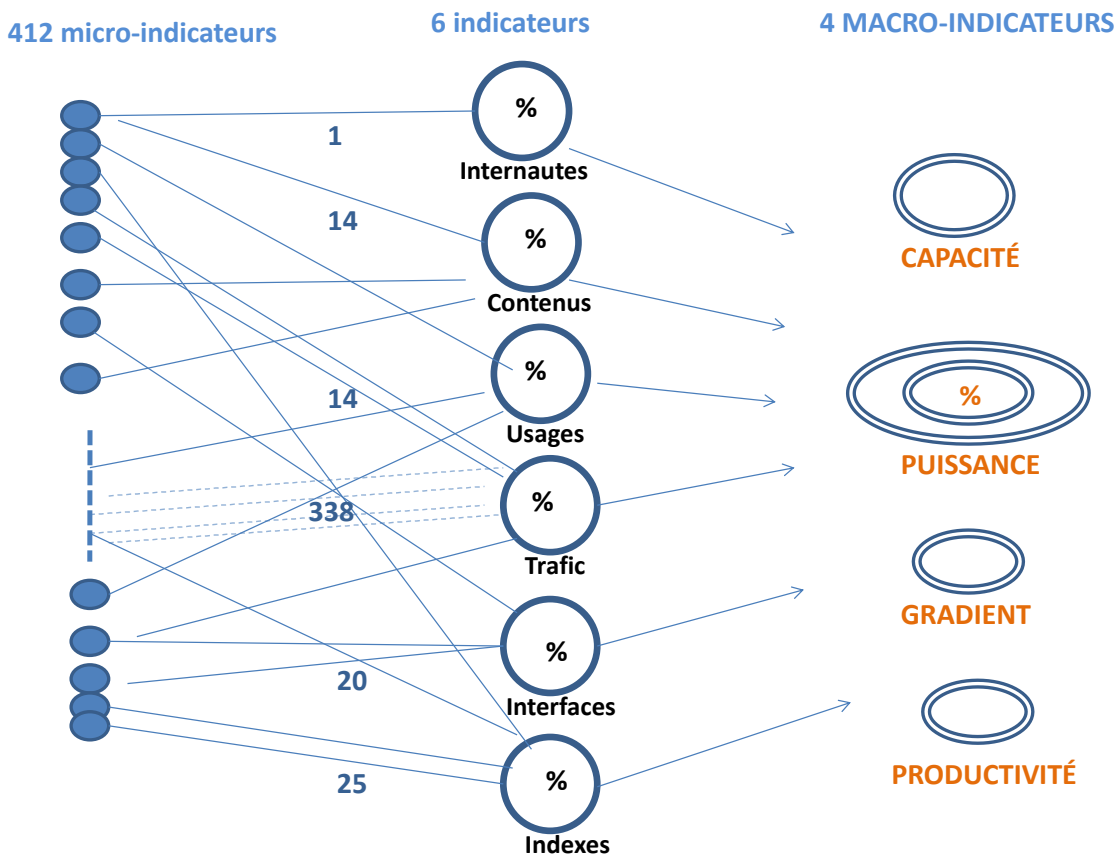
LES FIGURES

CONTEXTE

La première édition de cette méthode pour produire des indicateurs de la présence des langues dans l'Internet a été réalisée en 2017 et documentée sous le titre « *Une approche alternative pour produire des indicateurs des langues dans l'Internet* » (voir référence [1]), accessible sur le site de l'Observatoire, en 4 versions linguistiques (anglais, français, portugais et espagnol)[1]. Le lecteur est invité à consulter ce document préalablement à la lecture de cet article, lequel est rédigé en complément de cette première version, laquelle présentait à la fois la méthode et les résultats; cet article présente les différences de méthode et les nouveaux résultats.

Pour rappel, la méthode aborde les 138 langues dont le nombre de locuteurs L1[2] est supérieur à 5 millions[3] et produit des indicateurs pour chacune d'entre elles, selon le schéma suivant (dont les valeurs sont mis à jour pour la deuxième version).

Figure 1: Des micro-indicateurs aux macro-indicateurs



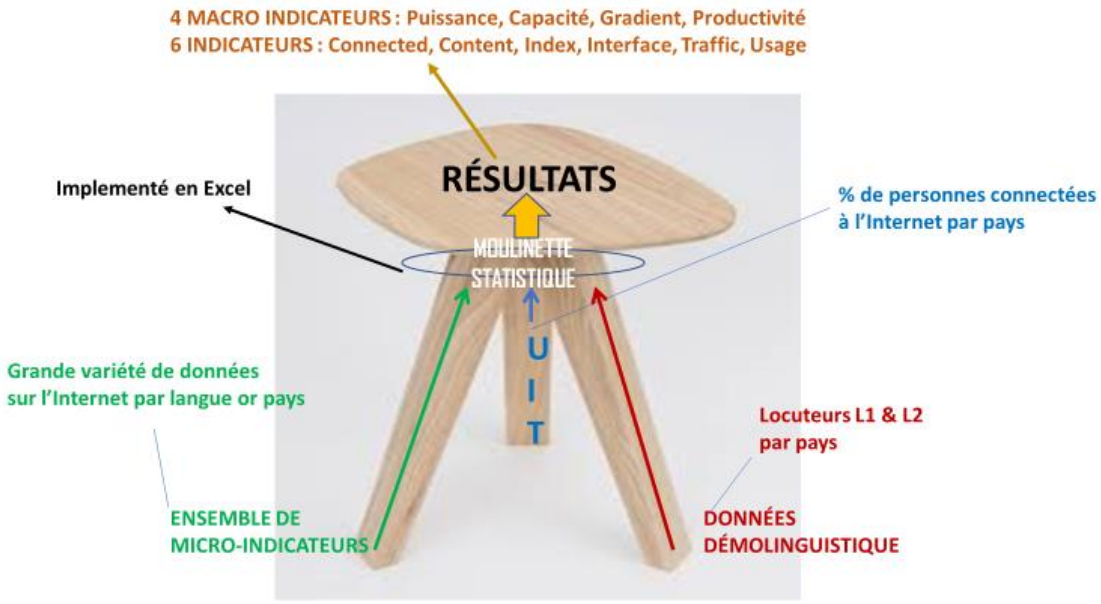
La méthode se compose de 3 types d'entrées et une dizaine d'indicateurs en sortie, tels que représentés dans la figure suivante.

[1] http://funredes.org/lc2017

[2] La convention est d'appeler L1 la langue maternelle (ou langue première) et L2 les langues secondes, étant entendu qu'une maitrise suffisante d'une langue est nécessaire pour être qualifiée de cette manière.

[3] En fait, le total est de 128 : afin de pouvoir faire des comparaisons avec l'étude 2017, 10 langues avec moins de 5 millions de locuteurs ont été laissées, car elles apparaissaient dans l'étude 2017, afin de pouvoir faire des contrôles et comparaisons. Ces langues sont : l'awhadi, le biélorusse, le bikol, le bugis, le dugri, l'arménien, le kimbundu, le luyia, le flamand occidental et le thaï du sud.

Figure 2: Le processus d'entrée/sortie du modèle



Le processus du modèle repose sur des mécanismes de pondération capables de transformer des données par pays en données par langue, des techniques d'extrapolation pour compléter les sources offrant des valeurs pour un nombre limité de pays, et des mécanismes de pondération avec la répartition mondiale des personnes connectées à l'Internet, afin de produire des pourcentages mondiaux à partir des différentes sources.

Tableau 1 : Les 2 types de pondérations utilisées.

	Démolinguistique	Internautes par langue
ENTRÉE	% par pays ---> % par langue	% Critère ---> % mondial
SORTIE	Données par pays	Donnée en % selon des critères spécifiques
PRODUCTION	Données par langue	Données par rapport à population L1+L2
DONNÉES DE PONDÉRATION	Locuteurs L1+L2 par pays	% de personnes connectées à l'Internet par pays
PORTÉE	Toutes les sources par pays	Indicateurs d'index et d'interfaces.
HYPOTHÈSE IMPLICITE	Indépendance des langues dans le pays	Taux de modulation de connexion à Internet selon le critère

Le modèle est implémenté sous Excel au sein d'un tableur de 7 Mo avec 17 feuilles de calcul corrélées, organisées autour des 215 pays considérés, des 138 langues traitées et des 412 micro-indicateurs collectés. Le modèle ainsi mis en place permet de vérifier en une fraction de seconde l'impact de toute hypothèse (y compris l'analyse prospective).

1. INTRODUCTION

Cette seconde version de la méthode référencée pour créer des indicateurs de présence des langues dans l'Internet apporte un ensemble d'améliorations tangibles qui impactent positivement la fiabilité de la méthode et réduisent les biais.

Les principales améliorations découlent de l'adoption de l'*Ethnologue Global Dataset 24*[4], de mars 2021, qui non seulement met à jour les données démolinguistiques (la quantité de locuteurs de chaque langue dans chaque pays) mais fournit également les données les plus fiables sur le sujet, même si une exactitude parfaite est inaccessible, et de plus, dans cette dernière version, offre la première source historique de répartition du nombre de locuteurs L2 par pays, pour chaque langue.

2. DIFFÉRENCES PAR RAPPORT À LA PREMIÈRE VERSION

De nombreuses différences sur la méthode ou les sources ont été réalisées par rapport à la version 1, dans l'esprit d'améliorer la qualité de la méthode et des produits.

2.1 Adoption de la base de données d'Ethnologue comme source démolinguistique

La partie principale de la source fournie par Ethnologue se présente sous la forme d'une matrice Excel de 11500 lignes au format suivant : « ISO639[5], Nom de la langue, Nom du pays, nombre de locuteurs L1, nombre de locuteurs L2, plus un grand nombre de paramètres associés non utilisés pour cette méthode ».

Afin d'obtenir le format requis par le modèle (une matrice avec tous les pays considérés en colonne et toutes les langues considérées en lignes), un ensemble d'étapes prudentes a été implémenté avec le support de différents programme rédigés sous la forme de macros VBA[6]. L'une des étapes les plus complexes a été de fusionner toutes les données des langues appartenant à une même macro-langue. Ce processus a impliqué 60 macro-langues regroupant 434 langues différentes[7] (voir le détail en annexe 2).

Après avoir terminé cette étape, le processus a consisté à réduire la liste complète des langues pour conserver seulement celles qui sont traitées par le modèle[8], en additionnant soigneusement tous les chiffres restants par pays sur une seule ligne « RESTE ».

Il est important de comprendre que l'adoption des données d'Ethnologue entraine l'acceptation des règles de présentation, lesquelles sont basées sur des considérations purement linguistiques :

- Regroupement des macro-langues[9]

[4] https://www.ethnologue.com/product/ethnologue-global-dataset-0
[5] Le code ISO à 3 caractères attribué à chacune des 7486 langues identifiées.
[6] Virtual Basic Applications, un langage utilisé pour créer des macros exécutables sous Excel.
[7] Par exemple la macro-langue arabe contient 29 langues telles que l'arabe égyptien ou l'arabe marocain.
[8] À ce stade 138 langues avec un nombre de locuteurs L1 supérieur à 5 millions.
[9] Un exemple significatif est le cas de la macro-langue serbo-croate dont la définition regroupe, par ordre alphabétique, le bosniaque, le croate, le monténégrin et le serbe. Ce groupement ne répond pas du tout à des critères géopolitiques et pourrait même être considéré comme polémique de ce point de vue. De plus, comme

- Liste des pays et dénomination anglaise correspondante.

La liste des pays traités par Ethnologue est plus grande que celle traitée par l'UIT[10] pour la fourniture des taux 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 réunir soigneusement les données d'Ethnologue pour les 29 pays non considérés par l'UIT (pour la liste complète, voir l'annexe 3) dans une seule colonne « Autres pays ».

2.2 Gestion des L2 et du multilinguisme

L'inclusion des dernières données d'Ethnologue dans le modèle a permis, en sous-produit, d'éliminer le biais majeur de la méthode qui était lié au traitement des L2. Pour la première fois, il existe une source fiable qui complète, pour chaque langue, le nombre de locuteurs L1 par pays avec le nombre de locuteurs L2 par pays. Dans la version 2017, les pourcentages de connexion à l'Internet pour les populations L2 ont été calculés en appliquant le pourcentage obtenu par le modèle pour les locuteurs L1. Un biais important résulte du fait que pour certaines langues majeures (comme par exemple le français et l'anglais) une forte proportion de locuteurs L2 appartient aux pays en développement où le taux moyen de connexion est bien inférieur à ce qui est obtenu en moyenne pour les locuteurs de L1.

Une autre conséquence positive de l'utilisation des données d'Ethnologue est la possibilité d'obtenir un « chiffre officiel » pour le multilinguisme. Le ratio mondial (L1+L2)/L1 a été établi dans l'édition 2017 en projetant les données disponibles pour les pays traités : il est ressorti à environ 1,25. Maintenant, le chiffre est fourni indirectement par les données Ethnologue et sa valeur est de 1,43.

Les données mondiales d'Ethnologue sont les suivantes:

- ✓ Population mondiale (somme des L1) : 7 231 699 136
- ✓ Total mondial de locuteurs L1+L2 : 10 361 716 756
- ✓ Le « ratio de multilinguisme » est donc de 10 361 716 756/7 231 699 136 = 1,4328 (en d'autres termes, 43 % de la population mondiale est au moins bilingue).

Ce chiffre de 43% est bien meilleur que les 25% utilisés dans la première version et ce n'est pas un élément anecdotique du modèle mais bien un des éléments clés. Comme le montre la première étude, le biais le plus courant et le plus critique des chiffres proposés pour les langues est le fait qu'ils ne considèrent pas correctement les locuteurs L2 (problème qui s'exprime pleinement dans l'Internet où la plupart des internautes utilisent leur(s) deuxième(s) langue(s) et où de nombreux sites Web sont multilingues[11]). Ne pas prêter attention à cela conduit à d'énormes erreurs, souvent cachées dans "le reste des langues", car les pourcentages mondiaux qui devraient être calculés pour une population de 10 milliards (les locuteurs L1 + L2) le sont sur un total de 7 milliards (la population mondiale).

---

certaines sources séparent 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 chiffres ne doivent pas être additionnés mais plutôt moyennés comme dans l'indicateur de profondeur de Wikipedia).

[10] L'Unité des Télécommunications Internationales (http://itu.int), l'organe des Nations Unies qui fournit des statistiques sur les télécommunications, y compris le pourcentage de personnes connectées à l'Internet par pays.

[11] En effet, les 6 indicateurs traités par l'étude sont par nature multilingues : les internautes visitent des sites et génèrent du trafic dans les différentes langues qu'ils maitrisent, souvent les sites internet sont multilingues, les interfaces sont multilingues, les services de traduction couvrent différentes langues…

Dans cette deuxième édition de la méthode, le principe de tout mesurer en termes de population L1+L2 (au lieu de la population mondiale) a été pleinement adopté pour assurer la cohérence des résultats. Pour cela (et aussi en raison d'autres améliorations), la comparaison entre les résultats 2017 et les résultats 2021 doit être conduite avec prudence. En effet, tous les macro-indicateurs, *puissance* mais aussi *capacité* et *gradient*, suivent désormais la règle d'être calculés sur la population L1+L2 au lieu de la population L1 et apparaîtront donc avec des valeurs inférieures en comparaison avec la version 2017.

2.3 Source pour les personnes connectées à l'Internet

Jusqu'en 2017, l'UIT fournissait, chaque année, une mise à jour de ses données[12] sur *le pourcentage d'individus qui utilisent l'Internet par pays*, y compris de ses propres estimations, quand qu'il n'y avait pas de source officielle pour certains pays. Cette donnée, qui est de plus un élément central de la méthode, était considérée comme parmi les plus fiables. Malheureusement, après 2017, l'UIT a décidé de ne plus fournir ses propres estimations, ce qui laisse de nombreux pays (presque tous les pays en développement[13]) avec, pour 2021, les anciennes valeurs de 2017.

Cela a posé un sérieux problème à cette étude et, après quelques itérations, a conduit à la décision de transgresser, dans ce cas, un principe fondamental dans ce type de travaux statistiques : celui de ne jamais modifier les données des sources.

La Banque mondiale fournit ses propres chiffres[14] pour le même indicateur, lesquelles sont clairement repris de l'UIT, mais, dans de nombreux cas, dépassent la limitation mentionnée et proposent des valeurs actualisées là où l'UIT a conservé les données de 2017. C'est un progrès appréciable, mais de nombreux pays restent encore en dehors de la mise à jour, ce qui aurait un impact négatif sur les indicateurs produits par le modèle pour les langues parlées dans ces pays et pourrait oblitérer d'éventuels progrès.

Finalement, il a été décidé d'utiliser les données de la Banque mondiale, lorsqu'elles complémentent celles de l'UIT et, pour les nombreux cas restants non actualisés, de réaliser, pour chaque pays concerné, une recherche dans l'Internet pour des données crédibles. Le projet a fourni ensuite ses propres estimations, basées, sauf argument contraire, sur la progression linéaire approximative des données des années précédentes.

Un cas continue malgré tout de poser un problème : l'Inde présente en 2021 un chiffre officiel de 20,1 % de personnes connectées à l'Internet alors que l'estimation de l'UIT pour 2017 indiquait une valeur bien 32 % et que de nombreuses sources rapportent un *boom* de l'Internet dans ce pays, avec des chiffres aux alentours de 50%[15]! N'ayant pu obtenir de réponse de la source officielle ni des collègues indiens consultés, il a été décidé, en raison de l'importance primordiale de l'Inde dans le contexte de l'étude[16], de transgresser, exceptionnellement, un autre principe encore plus fort : celui de ne pas modifier les sources officielles. L'hypothèse de travail qui motive ce changement est que le chiffre fourni par le *Ministère indien des statistiques et de la mise en œuvre des programmes* ne concerne que

[12] https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2021/PercentIndividualsUsingInternet.xlsx
[13] Seuls 80 pays ont fourni des chiffres officiels en 2019.
[14] Source : https://data.worldbank.org/indicator/IT.NET.USER.ZS
[15] Voir par exemple https://www.statista.com/statistics/255146/number-of-internet-users-in-india/ ou alors https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users
[16] Avec 34 langues qui font partie de la liste des langues considérées, dont des langues majeures telles que l'hindi et le bengali.

les connexions fixes et n'inclut pas les connexions mobiles à l'Internet. Sur la base de cette hypothèse, le chiffre conservateur de 40 % a été fixé.

Il faut noter que la sensibilité de ce chiffre sur les résultats n'est pas du tout marginale. Ci-après, les différents résultats finaux pour l'hindi et le bengali, selon la valeur du pourcentage de personnes connectées à l'Internet en Inde (avec 50% l'hindi passerait devant le français).

Tableau 2: Sensibilité des chiffres de l'Inde pour le pourcentage de personnes connectées à l'Internet

% Personnes connectées en Inde	*20,08 %*	30%	40%	50%
Hindi Puissance (classement)	*2,42% (10)*	2,91 % (8)	3,38 % (5)	3,81 % (4)
Bengali Puissance (classement)	*0,75% (17)*	0,82% (15)	0,88% (15)	0,95% (14)

2.4 Gestion des sources pour les micro-indicateurs

L'ensemble du processus de gestion des sources pour les 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, recherchez des sources dans l'Internet.
2. Sélectionner les sources en fonction de leur fiabilité et de l'applicabilité au processus[17].
3. Collecter les sources sélectionnées dans un format permettant une introduction simplifiée dans le modèle.
4. Introduire les sources validées dans le modèle et leur associer un thème.
5. Évaluer les biais des sources.

En annexe 1, 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 langues correspondant à ceux du modèle et dans le même ordre séquentiel.

Pour l'étape 3, toutes les sources sont collectées à partir d'une URL spécifique (voir en Annexe 1 la liste complète des URLs) 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 qui est visé pour transformer toutes les sources. Le processus de transformation du format PDF vers Excel peut être relativement simple dans la plupart des cas, quand les tableaux sont bien structurés, mais dans certains cas, il existe une incompatibilité et certaines astuces sont nécessaires, comme passer d'abord par un format DOC intermédiaire.

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

[17] Il peut arriver que des données fiables soient dans un format qui interdit une exploitation automatisée.

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 le traitement automatique ou nécessite la sous-traitance d'un travail de collecte manuelle qui est fastidieux et requiert une concentration et une discipline élevées pour éviter les erreurs. La collecte de données de trafic impliquant des centaines de micro-indicateurs a ainsi été sous-traitée.

Il faut rendre crédit aux institutions (en général, organisations internationales ou ONG) qui fournissent les données dans un format informatiquement exploitable (Wikimedia fournit par exemple, dans sa version anglaise, des tableaux HTML qui se transforment directement au format Excel sans perte de structure).

L'obtention d'une copie de la source au format Excel ou compatible (en général, un tableau de nom de pays ou de langues avec des valeurs ou des pourcentages associés) n'est pas la fin du processus. Avec 250 pays ou des centaines de langues à traiter et au lieu de l'utilisation du code ISO, non ambigu, celle de noms littéraux qui peuvent être dans différentes langues et dans des orthographes non standardisées, l'intégration des données dans le modèle n'est pas réalisable à la main. Deux programmes ont été écrits pour ce processus qui, dans les deux cas, nécessitent un réglage récursif[18] afin d'intégrer les différentes orthographes. Les sorties des programmes sont des fichiers Excel directement utilisables pour copier les sources entièrement, ou ligne par ligne, dans le fichier Excel du modèle. Outre l'énorme gain de temps de cette méthode, elle garantit d'obtenir les données sans erreurs.

Notez également que la gestion des macro-langues a rendu ce processus encore plus complexe, car le regroupement des langues doit être réalisé dans les données sources avant le 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, bosniaque, croate et monténégrin ont été fusionnées en serbo-croate (le nombre de cas similaires étant assez élevé). Pour le traitement manuel des orthographes inconnues rapportées par le programme (incorporation des orthographes comme synonymes ou rejet dans la catégorie autres), il a été utilisé en soutien la page d'Ethnologue descriptive de chaque code langue[19].

2.4.1 INDEX

L'échéance est arrivée trop tôt lors de la production de la version 2017 et cet indicateur s'est avéré insuffisant avec une source unique fournissant 5 micro-indicateurs. Cette fois, l'attention requise a été accordée et une collecte quasi exhaustive a été réalisée pour cet indicateur. Une grande variété de paramètres caractérisant les progrès des pays dans la société de l'information a été inclus, avec maintenant 25 micro-indicateurs, de la stabilité électrique à l'intelligence artificielle en passant par la gouvernance et de nombreux autres paramètres (voir l'annexe 1 pour la gamme complète).

[18] Le processus récursif se termine lorsque le contrôle d'erreur n'identifie plus d'orthographes inconnues.
[19] https://www.ethnologue.com/language/srp

2.4.2 CONTENUS

Comme expliqué auparavant, les sources de données sur les langues dans l'Internet sont extrêmement rares et en conséquence cet indicateur dépend fortement des statistiques exceptionnelles de Wikimedia. Cependant, dans la section Biais du document il est établi que la présence des langues dans Wikimedia n'est pas proportionnelle à leur présence dans la Toile et ne peut donc être considérée comme un indicateur fiable de la répartition linguistique des contenus de l'Internet.

Un moyen d'équilibrer les résultats des statistiques de Wikimedia a été introduit dans le modèle mais le diagnostic douloureux reste que *contenus* est l'indicateur le plus faible de cette méthode, tout en étant de plus un élément très sensible (des changements relativement marginaux des valeurs dans cet indicateur peuvent provoquer un impact important dans les valeurs des macro-indicateurs). Si l'un des principaux objectifs du projet est bien sûr de connaître la répartition linguistique des contenus de l'Internet, il existe une difficulté frustrante à pondérer correctement cette information, le macro-indicateur holistique *puissance* en restant probablement à ce stade la meilleure approximation[20].

Pour essayer de mieux contrôler l'influence excessive des chiffres de la galaxie Wikimédia sur cet indicateur, deux décisions ont été prises. La première concerne exclusivement l'encyclopédie Wikipédia : au lieu d'avoir un micro-indicateur pour chacun des chiffres fournis (*nombre d'articles, éditeurs actifs, éditions et profondeur*[21]), une formule a été mise en place pour définir un seul micro-indicateur:

$$W(i) = \text{Articles}(i) \times \text{Éditions}(i) \times \text{Éditeurs}(i) \times \text{Profondeur}(i) / \text{L1+L2}(i)^2$$

Cette formule exprime plus précisément l'activité globale de Wikipédia par langue, rabaissant la part des langues pour lesquelles des *bots* (programme informatique simulant le

[20] Comme le montre la première édition, l'effort louable et apprécié de W3Techs pour proposer des chiffres pour les contenus est caractérisé par des biais très importants et à de nombreux niveaux (le plus fort mais pas unique étant le manque de considération du multilinguisme et le fait que les sites Web multilingues qui incluent l'anglais sont probablement comptabilisés en anglais uniquement). Ainsi cette source projette des valeurs pour les contenus en anglais qui sont extrêmement exagérées (au-dessus de 50 % alors que la réalité est probablement aujourd'hui inférieure à 25 %). L'absence de pluralité de sources entretient ainsi dans les médias le mythe que plus de la moitié des sites Web sont en anglais. C'était le cas entre 2007 et 2009 (voir [3]), mais, depuis 2009, la croissance exponentielle du chinois, de l'hindi, de l'arabe, du turc, du bengali, du vietnamien, de l'ourdou, du persan et du marathi, pour citer seulement les langues qui sont maintenant dans les 20 premières places et qui pèsent ensemble près de 28% des contenus, a changé radicalement la donne et **l'anglais représente seulement un quart des contenus**. Entre 2000 et 2007, le mythe persistant était que l'anglais occupait 80% de la Toile et cette désinformation a finalement disparu après 2009 avec la publication par l'UNESCO de rapports (voir [3] et [4]) qui établissaient une présence de l'anglais autour de 50%. Comment l'anglais aurait-il pu rester stable à 50% des contenus depuis 14 ans malgré l'intense internationalisation qu'a connu l'Internet et un nombre d'anglophones connectés (L1+L2) qui est passé de 32% du total des personnes connectées en 2007 (source : https://web.archive.org/web/20120511104604/http://dtil.unilat.org/LI/2007/es/resultados_es.htm) à seulement 13% aujourd'hui?

[21] Cité de Wikimedia : *Profondeur*, qui est défini comme [Modifications/Articles] × [Non-Articles/Articles] × [1 − Stub-ratio]), est un indicateur approximatif de la qualité d'un Wikipédia, montrant à quelle fréquence ses articles sont mis à jour. Il ne fait pas référence à la qualité académique.

comportement d'un humain) sont utilisés pour créer des articles en traduisant des articles dans d'autres langues[22], sans trop se préoccuper par la suite de les mettre à jour.

Le tableau suivant montre comment cette formule parvient à mieux refléter la réalité. La dernière colonne (*présence*), qui est le rapport entre le nombre d'articles et la population L1+L2 (nombre d'articles par locuteur), est une démonstration claire de pourquoi la présence des langues dans Wikipédia n'est pas un bon indicateur de la présence globale des langues dans l'Internet…

Notez que la valeur de *profondeur* pour le vietnamien n'a pas été renseignée et nous avons mis une valeur de 1 pour éviter une formule nulle[23].

Tableau 3: Facteurs Wikipédia et la formule

Langue	Articles	Modifications	Utilisateurs actifs	Profondeur	PRÉSENCE	FORMULE
Anglais	6332139	1027716498	125399	1073	0,47	481775
Cebuano	5853095	32075254	186	2	36,71	275
Suédois	3050759	49330695	2148	12	23,37	22759
Allemand	2593827	212207089	18119	93	1,92	50897
Français	2342875	183969129	18054	242	0,88	26424
Néerlandais	2060512	59302602	3933	17	8,45	13742
Russe	1736736	115035192	10425	137	0,67	4286
Italien	1703284	121418801	8085	172	2,51	62435
Espagnol	1698331	136390848	15694	210	0,31	2590
Polonais	1480982	63723938	4235	32	3,64	7742
Japonais	1277204	84188217	15173	85	1,01	8683
Vietnamien	1266628	65110373	2476	1	1,65	35
Chinois	1208732	66159632	8940	202	0,08	62
Arabe	1123561	54279052	5189	227	0,31	536
Ukrainien	1100281	32831286	2773	53	3,32	4823
Portugais	1067241	61371751	9508	176	0,41	1651

Dans le chapitre discutant les biais, une analyse approfondie des statistiques de Wikimedia est présentée.

La deuxième décision prise pour équilibrer l'influence de Wikimedia sur l'indicateur *contenus* est un système de pondération qui donne plus d'importance au *T-Index* de Translated[24] qu'à toute la collection d'indicateurs de Wikimedia. Jouer avec différentes configurations des facteurs de pondération a montré la grande sensibilité de cet indicateur, essentiellement due

[22] Sans cette formule, le cebuano avec un grand nombre d'articles, mais une très faible, profondeur apparait avec le score de *gradient* le plus élevé.

[23] La faible valeur de profondeur est le reflet du fait que 67% des articles sont créés par des bots, pas par des humains (source : https://www.wikiwand.com/en/Vietnamese_Wikipedia).

[24] Cet index, accessible sur https://translated.com/les-langues-qui-comptent, est une tentative de mesurer le potentiel des langues dans le commerce électronique à partir du nombre d'internautes par langue multiplié par les dépenses en ligne estimées. Il a utilisé les chiffres de la Banque mondiale et de l'UIT et propose une projection 2021 qui est le chiffre retenu pour le modèle. C'est, outre les données de Wikimedia, l'une des très rares sources sérieuses de langues disponibles sur l'Internet.

au très faible nombre de sources et au fait que certaines langues ont une présence disproportionnée par rapport à leur nombre de locuteurs.

La configuration de pondération finalement mise en œuvre est la suivante :

Tableau 4: Pondération des indicateurs de contenu

OBJET	POIDS
Amazon US - nombre de livres 2017[25]	0,5
Formule Wikipédia	1
Nombre de WikiBooks par langue	0,5
Articles WikiQuote par langue	0,1
Nombre d'articles WikiSource par langue	0,1
Nombre d'articles Wikiversité par langue	0,1
Nombre d'articles Wiktionnaire par langue	0,1
Nombre d'articles WikiNews par langue	0,1
Nombre d'articles WikiVoyages par langue	0,1
T-Index pour le commerce électronique Projection 2021	3

2.4.3 TRAFIC

Le travail pour l'indicateur *trafic* a également été très dense avec beaucoup d'essais et d'erreurs. En 2017, il avait été établi que les données d'Alexa (pourcentage de trafic par pays vers une série de sites) étaient extrêmement biaisées en défaveur des pays asiatiques (en particulier l'Inde et la Chine) et du Brésil et quelque peu biaisées en faveur du français et de l'anglais. Quatre ans plus tard, la collecte de données d'Alexa a montré des situations étranges (absence de trafic dans le pays de création du site[26]) et l'impression d'une tendance forte à sous-estimer les trafics des pays européens ; par contre l'Inde apparaît maintenant bien placée dans tous les sites, pas tellement la Chine.

Une étude comparant les données de trafic avec les données d'abonnement pour cinq réseaux sociaux importants a confirmé les impressions empiriques. En résumé, le trafic en provenance du Brésil semble largement sous-estimé par rapport au niveau d'abonnement, de même pour Allemagne, Espagne, France, Italie et Royaume-Uni ; par contre Inde, Japon, Corée apparaissent largement surestimés par Alexa. Voir le chapitre sur les biais pour plus de détails.

Devant ces résultats peu encourageants, il a été décidé de rechercher un outil de mesure alternatif. SimilarWeb semblait être la meilleure alternative et le test était prévu avant l'achat d'un abonnement. Malheureusement, il a été impossible d'accéder aux données de trafic par pays et malgré de nombreuses tentatives de communication vers l'entreprise, via différents canaux, y compris le chat interactif, aucune réponse n'a jamais été obtenue.

[25] Le manque de données accessibles équivalentes pour 2021 et la situation avec Wikimedia ont conduit à la décision de conserver ce micro-indicateur bien qu'il ne soit pas actualisé.
[26] À titre d'exemple, theses.fr a montré un trafic nul en France, de même que spip.net, une application principalement utilisée en France.

Face à cette situation de blocage, un autre fournisseur, Semrush.com, a été testé et des chiffres par pays ont été collectés pour la même série de sites Web. Semrush, à la différence d'Alexa, fournit, pour chaque site mesuré, les résultats pour tous les pays, ce qui était une perspective intéressante, car supprimant le besoin d'extrapolation. Cependant, il arrive que dans certains cas le total est inférieur à 100 % (ce qui n'est pas un problème) et d'autres fois il dépasse 100 % (ce qui est un problème). Les chiffres ont été normalisés pour être exacts à 100 % en utilisant une règle au prorata avant l'introduction dans le modèle.

Après avoir exécuté le modèle, transformées les données nationales en données linguistiques, les résultats n'étaient pas convaincants : la valeur du chinois était bien trop faible, de même pour l'hindi et l'arabe et pour les « langues restantes ».

Les différences extrêmes entre les résultats obtenus à partir des données d'Alexa et de Semrush pour un échantillon identique de sites sont un signal d'alarme quant à la fiabilité de ces outils et une inquiétude pour les futurs plans visant à étendre le nombre de sites Web étudiés et permettre des résultats de différenciation thèmatique pour certaines langues.

2.4.4 INTERFACES

La liste des langues acceptées dans les interfaces d'applications importantes ou comme cible possible pour les services de traduction en ligne ne pose pas de problème particulier. La liste des applications sélectionnées peut être consultée en Annexe 1. À noter qu'afin de réduire l'importance des données de Wikimédia sur le modèle il a été décidé de supprimer de cet indicateur les sources Wikimédia.

2.4.5 USAGES

Pas de difficultés particulières non plus pour cet indicateur, si ce n'est de trouver des données gratuites (essentiellement nombre d'abonnés par pays) pour les principaux réseaux sociaux. Finalement, la couverture a réussi à inclure les applications suivantes : Facebook, Instagram, Linkedin, Messenger, Pinterest, Reddit et Twitter. De plus, certaines sources non liées aux réseaux sociaux ont été incluses, comme par exemple le nombre de téléchargements d'OpenOffice par pays (voir la liste complète en Annexe 1).

2.5 Résumé des indicateurs

Le tableau suivant résume la description de chacun des indicateurs et comment il est construit à partir des micro-indicateurs.

Tableau 5: Description des indicateurs

INDICATEUR	DÉFINITION	TECHNIQUE	FIABILITÉ/BIAIS
R : INTERNAUTES	Mono indicateur, à partir des chiffres de l'UIT et de la Banque mondiale, du % de personnes connectées par pays, extrapolé là où les chiffres font défaut.	pondération pays -> langue sans extrapolation	Grande fiabilité Biais très marginal bien qu'en augmentation en raison du manque de mise à jour pour de nombreux pays.

B : USAGES	Comprend 14 micro-indicateurs avec des données 2021 : - Fixe + mobile % par pays - Haut débit % par pays - Téléchargement cumulatif d'OpenOffice - Facebook, Instagram, LinkedIn, Messenger, Netflix, Pinterest Twitter, YouTube, % d'abonnés par pays	pondération pays -> langue extrapolé en proportion taux connectivité Moyenne des micro-indicateurs	Forte fiabilité. Faible biais.
C : TRAFIC	Alexa a mesuré le trafic par pays vers une sélection de 338 sites Web.	pondération pays -> langue extrapolé proportionnellement Moyenne tronquée à 20 %	Fiabilité relativement bonne Mais de forts biais négatifs européens d'Alexa confirmés par des comparaisons de trafic et de nombre d'abonnés par pays.
D : INDEXES	Comprend 25 indexes provenant de diverses sources mesurant des paramètres tels que : - E. gouvernement - Accès universel - E. participation - Infrastructures générales (Voir l'annexe 1 pour la liste complète)	pondération pays -> langue extrapolé par la méthode des quartiles. Puis transformation en données mondiales par pondération en pourcentage avec l'UIT Moyenne des micro-indicateurs	Bonne fiabilité et biais marginal (données subjectives quantifiées par un organisme compétent).
E : CONTENUS	Comprend 13 micro-indicateurs avec pondération associée. T-Index of Translated, une mesure du potentiel de commerce électronique d'une liste de langues (2021) - Nombre de livres sur Amazon (2017) - 11 micro-indicateurs de langue issus de Wikimedia : articles, utilisateurs ou éditeurs ; tous les indicateurs Wikimedia sont synthétisés avec une formule.	Utilisation directe de chiffres par langue pondérés pour équilibrer l'importance de Wikimedia. Fusion de 4 indicateurs Wikipedia avec une formule. Moyenne tronquée à 20% du micro indicateur	Très bonne fiabilité pour Wikimedia et Amazon. Mais assez biaisé en raison de la très faible présence de certaines langues asiatiques majeures. Le nombre de micro-indicateurs devrait être augmenté pour donner plus de force à la moyenne.
F : INTERFACES (et langues de traduction)	Comprend 23 micro-indicateurs binaires	% de présence sur les 23 micro-indicateurs. % mondial pondération avec les chiffres de l'UIT.	Parfait.

3. RÉSULTATS

Les prochains tableaux présentent les résultats pour les langues en tête de chaque macro-indicateur, hors productivité[27]. Le tableau suivant présente tous les résultats récapitulatifs pour les 15 langues les plus "*puissantes*" dans l'Internet. Les résultats sont des pourcentages réalisés sur la base de la population L1+L2. Conn.M signifie pourcentage mondial de personnes connectées, Pop.M, population mondiale et L. Conn. Pourcentage de locuteurs connectés.

Tableau 6 : Indicateurs pour les 15 premières langues en termes de puissance

	Conn.M	Pop.M	TRAFIC	L. Con.	USAGE	CONT.	INTERF.	INDEX	PUISS.	Capac.	Grad.
Anglais	15,30%	13,01%	37,44%	64,33%	27,92%	38,61%	21,73%	17,87 %	**26,48%**	**2,04**	**1,73**
Chinois	17,65%	14,72%	7,79 %	65,59%	5,47%	8,18%	25,07%	19,38%	**13,92%**	**0,95**	**0,79**
Espagnol	7,00%	5,24%	10,72%	73,08 %	11,74%	5,42%	9,94 %	7,59%	**8,73 %**	**1,67**	**1,25**
Français	3,00%	2,58%	2,64 %	63,67 %	3,75%	5,40%	4,26%	3,21%	**3,71%**	**1,44**	**1,24**
Hindi	4,26%	5,80%	4,81%	40,18%	3,16%	0,28%	4,03%	3,71%	**3,38%**	**0,58**	**0,79**
Portugais	3,05%	2,49%	1,42%	67,16%	5,53%	3,30%	3,85%	2,92%	**3,35%**	**1,35**	**1,10**
Russe	3,51%	2,49%	1,81%	77,20%	2,28%	3,38%	3,88%	3,78%	**3,11%**	**1,25**	**0,88**
Arabe	3,89%	3,53%	2,30%	60,14%	3,02%	2,05%	4,29%	3,01%	**3,09%**	**0,88**	**0,80**
Allemand	2,09 %	1,30%	1,32%	87,65%	1,95 %	5,84%	2,97%	2,98%	**2,86%**	**2,19**	**1,37**
Japonais	2,07 %	1,22%	1,98%	92,62%	1,76%	3,55%	2,77%	3,01%	**2,52%**	**2,07**	**1,22**
Malais	2,20%	2,36%	0,89%	51,00%	2,79%	0,79%	1,91 %	1,99%	**1,76%**	**0,75**	**0,80**
Italien	0,91%	0,66%	0,51%	75,65 %	0,97%	3,39%	1,22%	1,20%	**1,37%**	**2,09**	**1,51**
Turc	1,21%	0,85%	1,03%	77,98%	1,59%	0,94%	1,43%	1,22%	**1,24%**	**1,46**	**1,02**
Coréen	0,93%	0,79%	0,93%	64,73 %	0,99%	0,85%	1,10%	0,95%	**0,96%**	**1,22**	**1,03**
Bengali	1,14%	2,58%	1,22%	24,15%	1,13%	0,26%	0,72%	0,84%	**0,88%**	**0,34**	**0,78**
RESTE	**31,79 %**	**40,39%**	**23,19%**		**25,95%**	**17,77%**	**10,81%**	**26,34%**	**22,64%**		
TOTAL	**100%**	**100%**	**100%**		**100%**	**100%**	**100%**	**100%**	**100%**		

La ligne RESTE représente l'ensemble complet de toutes les langues du monde, à l'exception des 15 langues répertoriées dans le tableau. Il doit rester clair que le classement en termes de *puissance* privilégie les langues qui ont le plus grand nombre de locuteurs. Les macro-indicateurs *capacité* et *gradient* offrent des résultats indépendamment du nombre de locuteurs.

Rappel:
Puissance[28] a été définie comme la moyenne des 5 indicateurs.
Capacité[29] est la valeur de la puissance divisée par le % de locuteurs L1+L2
Gradient[30] est la valeur de la puissance divisée par le % de locuteurs L1+L2 connectés

Le tableau suivant montre les langues les plus connectées.

[27] Cet indicateur sera réévalué dans le chapitre Correction des biais. L'indicateur de *puissance*, qui intègre tous les éléments serait probablement, à ce stade, une meilleure approximation de la répartition des contenus par langue, donnée qui reste très difficile à obtenir de manière fiable à ce jour.
[28] Le terme *puissance* a été utilisé au lieu de *poids* qui paraissait plus naturel pour éviter toute confusion avec l'utilisation transversale importante de la pondération dans la méthode. Il représente la présence absolue d'une langue dans l'Internet intégrant tous les facteurs.
[29]La *capacité* est la présence relative d'une langue dans l'Internet, indépendamment de son nombre de locuteurs; il indique le dynamisme d'une langue dans l'Internet.
[30] Le *gradient* indique le dynamisme des locuteurs connectés ; le terme *gradient*, exprimant une dérivée et donc une tendance a été choisi car un fort gradient est une promesse d'augmentation de capacité.

Tableau 7 : Langues triées par pourcentage de personnes connectées

TRI PAR INTERNAUTES	Internautes	Capacité	Gradient
Danois	97,82 %	2,19	1,22
Suédois	93,49 %	2,61	1,53
Japonais	92,62%	2,07	1,22
Néerlandais	92,02%	2,26	1,34
Suisse allemand	91,56 %	1,21	0,72
Flamand occidental	90,43%	1,12	0,68
Finlandais	89,67%	3,42	2,09
Bavarois	87,68%	0,97	0,61
Allemand	87,65%	2,19	1,37
Hébreu	85,46%	5,24	3,35
Slovaque	82,47%	1,30	0,86
Biélorusse	82,27%	1,00	0,66
Tchèque	81,37%	1,70	1,14
Polonais	81,17%	1,88	1,26
Hongrois	79,92%	1,79	1,22
Tatar	78,05%	0,87	0,61
Turc	77,98%	1,46	1,02
Serbo-croate	77,78%	3,14	2,21
Grec	77,71 %	1,75	1,23
Russe	77,20%	1,25	0,88
Kazakh	76,98%	0,90	0,64
Roumain	75,66 %	1,18	0,86
Italien	75,65 %	2,09	1,51
Albanais	75,48%	1,12	0,81
Azerbaïdjanais	74,76%	0,94	0,69
Napolitain-calabrais	74,39%	0,84	0,62
Espagnol	73,08 %	1,67	1,25
Kurde	73,02 %	0,89	0,67
Bulgare	70,34%	1,18	0,92
Arménien	69,86%	1,41	1,11
Vietnamien	69,04 %	1,07	0,85
Guarani	68,83%	0,64	0,51
Portugais	67,16%	1,35	1,10

Le tableau suivant est trié par *capacité*.

Tableau 8 : Langues triées par capacité

TRI PAR CAPACITÉ	Internautes	Capacité	Gradient
Hébreu	85,46%	5,24	3,35
Finlandais	89,67%	3,42	2,09
Serbo-croate	77,78%	3,14	2,21
Suédois	93,49 %	2,61	1,53
Néerlandais	92,02%	2,26	1,34
Allemand	87,65%	2,19	1,37
Danois	97,82 %	2,19	1,22
Italien	75,65 %	2,09	1,51

Japonais	92,62%	2,07	1,22
Anglais	64,33%	2,04	1,73
Polonais	81,17%	1,88	1,26
Hongrois	79,92%	1,79	1,22
Grec	77,71 %	1,75	1,23
Tchèque	81,37%	1,70	1,14
Espagnol	73,08 %	1,67	1,25
Turc	77,98%	1,46	1,02
Français	63,67 %	1,44	1,24
Arménien	69,86%	1,41	1,11
Portugais	67,16%	1,35	1,10
Slovaque	82,47%	1,30	0,86
Russe	77,20%	1,25	0,88

Et enfin, le dernier tableau, trié par gradient, met en évidence le dynamisme des personnes connectées. La présence en troisième position du malgache[31] est une conséquence du dynamisme de ses locuteurs dans certains indicateurs Wikimedia.

Tableau 9 : Langues triées par gradient

TRI PAR GRADIENT	Internautes	Capacité	Gradient
Hébreu	85,46%	5,24	3,35
Serbo-croate	77,78%	3,14	2,21
Malgache	9,79%	0,40	2,21
Finlandais	89,67%	3,42	2,09
Anglais	64,33%	2,04	1,73
Suédois	93,49 %	2,61	1,53
Italien	75,65 %	2,09	1,51
Allemand	87,65%	2,19	1,37
Néerlandais	92,02%	2,26	1,34
Polonais	81,17%	1,88	1,26
Espagnol	73,08 %	1,67	1,25
Français	63,67 %	1,44	1,24
Grec	77,71 %	1,75	1,23
Danois	97,82 %	2,19	1,22
Hongrois	79,92%	1,79	1,22
Japonais	92,62%	2,07	1,22
Tchèque	81,37%	1,70	1,14
Arménien	69,86%	1,41	1,11
Portugais	67,16%	1,35	1,10

Au-delà du fait prévisible que les langues nationales des pays reconnus pour leurs politiques volontaristes en faveur de la société de l'information figurent aux premières places, il est remarquable que plusieurs langues se classent au-dessus de l'anglais, malgré son avantage stratégique dans l'Internet

[31] Un tel classement pour le malgache, une langue avec moins de 10 % de locuteurs connectés, et une capacité très faible, peut légitimement provoquer la surprise : c'est le résultat d'un « accident mathématique » dû à une présence extrêmement disproportionnée dans l'un des micro-indicateurs de contenu et c'est un symptôme de la faiblesse de cet indicateur qui est discuté ci-après.

(langue de choix pour les contenus multilingues et croyance de beaucoup que toujours c'est la lingua franca de l'Internet).

Ces résultats sont à prendre avec certaines réserves en raison des biais mentionnés dans le document, en particulier les difficultés avec l'indicateur de *contenus* dont les variations peuvent avoir un impact considérable sur ces macro-indicateurs[32].

4. ANALYSE DES RÉSULTATS

Bien que les comparaisons avec les résultats de 2017 doivent être prudente, en raison de l'importance et de la nature des changements (notamment le choix d'exprimer des pourcentages par rapport à la population mondiale totale L1+L2), certains phénomènes peuvent être mis en évidence.

La croissance attendue de l'hindi qui rivalise avec le français pour la 4ème place et l'apparition du turc dans la liste des langues les plus puissantes. Comme prévu également, les écarts entre les langues qui suivent le français et l'hindi sont trop faibles pour considérer que les résultats sont au-delà de l'intervalle de confiance : portugais, russe, arabe et allemand. Cependant, la démographie pourrait dans un futur proche séparer les positions respectives au rythme de la réduction de la fracture numérique.

Quant aux macro-indicateurs indépendants du nombre de locuteurs, l'apparition du serbo-croate est à prendre avec précaution en raison des risques d'erreurs dans la gestion des sources résultant de la décision d'adopter la classification Ethnologue pour les macro-langues. Et clairement, l'indicateur c*ontenus*, et sa forte dépendance aux statistiques de Wikimédia, malgré l'effort fait pour le contrebalancer, favorisent clairement les langues dont les locuteurs ont investi pour une forte présence dans Wikimédia. Ces langues apparaissent dans le tableau ci-dessous, d'abord trié par le ratio 1000 x Nombre d'articles/L1+L2 locuteurs, puis trié par le résultat de la formule mise en place.

Tableau 10: Présence des langues dans Wikipédia

Langue	Articles	Éditions	Utilisateurs actifs	Profondeur	FORMULE	%FACTEUR/L1+L2	%FACTEUR/CONN	ART/L1+L2
Suédois	3050759	49330695	2148	12	22759	1,74	1,86	**233,68**
Finlandais	512026	19813368	1752	40	21354	3,70	4,13	**88,74**
Néerlandais	2060512	59302602	3933	17	13742	0,56	0,61	**84,51**
Serbo-croate	1514114	78699318	1959	92	53779	2,69	3,46	**75,77**
Biélorusse	281379	6093511	384	61	2620	0,67	0,81	**71,87**
Danois	267641	10777444	767	64	4486	0,80	0,82	**47,64**
Hongrois	489514	23958462	1561	59	6871	0,55	0,69	**39,04**
Polonais	1480982	63723938	4235	32	7742	0,19	0,23	**36,44**
Tchèque	484445	20095461	2242	46	5593	0,42	0,51	**36,16**
Ukrainien	1100281	32831286	2773	53	4823	0,15	0,23	**33,16**
Bulgare	273163	11023721	789	27	942	0,11	0,16	**33,10**
Hébreu	298053	31660591	3335	258	92147	9,82	11,49	**31,75**
Italien	1703284	121418801	8085	172	62435	0,92	1,22	**25,10**
Allemand	2593827	212207089	18119	93	50897	0,38	0,43	**19,21**
Japonais	1277204	84188217	15173	85	8683	0,07	0,07	**10,11**

[32] Avant l'introduction de la formule Wikipédia et de la pondération Wikimedia, le cebuano, la deuxième langue en nombre d'articles Wikipédia, proche de l'anglais, avec un nombre d'articles deux ordres de grandeur supérieur à son nombre de locuteurs apparaissaient au premier rang de *gradient*…

Persan	816984	32472834	5416	172	3534	0,04	0,07	9,77
Français	2342875	183969129	18054	242	26424	0,10	0,16	8,78
Anglais	6332139	1027716498	125399	1073	481775	0,36	0,56	4,70

Le tableau suivant montre clairement pourquoi certaines langues, comme l'hébreu, le finnois et le serbo-croate, ont obtenu un avantage dans les résultats finaux classés par gradient.

Tableau 10: Présence Wikipédia triée par formule

Langue	FORMULE	%FORMULE/L1+L2	%FORMULE/CONN
Hébreu	92147	9,82	11,49
Finlandais	21354	3,70	4,13
Serbo-croate	53779	2,69	3,46
Suédois	22759	1,74	1,86
Italien	62435	0,92	1,22
Danois	4486	0,80	0,82
Biélorusse	2620	0,67	0,81
Hongrois	6871	0,55	0,69
Néerlandais	13742	0,56	0,61
Anglais	481775	0,36	0,56
Tchèque	5593	0,42	0,51
Allemand	50897	0,38	0,43
Polonais	7742	0,19	0,23
Ukrainien	4823	0,15	0,23
Bulgare	942	0,11	0,16
Français	26424	0,10	0,16
Japonais	8683	0,07	0,07
Persan	3534	0,04	0,07

Ces considérations conduisent naturellement à la discussion sur les biais.

5. ANALYSE DES BIAIS

Il existe trois grandes catégories de biais susceptibles d'affecter les résultats :

- Biais propres à la méthode
- Biais de sélection des sources
- Biais des sources

5.1 Les biais propres à la méthode

L'un des principaux biais propre à la méthode, qui consiste à donner la même valeur de pourcentage de locuteurs L1 connectés à l'Internet pour les locuteurs L2, a été éliminé avec le passage aux données Ethnologue, grâce à la répartition des locuteurs L2 par pays. Ce biais important affectait particulièrement les langues à forte population L2 dans les pays à faible taux de connectivité (français et anglais). C'est un progrès primordial pour la confiance dans les chiffres produits par le modèle établi.

Le deuxième biais de la méthode est de considérer qu'au sein d'un pays donné tous les locuteurs ont le même pourcentage de connectivité (en d'autres termes, le pourcentage national moyen de personnes connectées à l'Internet est appliqué à tous les locuteurs). Ce biais interdit de distinguer les locuteurs de langues différentes au sein d'un pays avec la méthode (par exemple, les locuteurs du catalan en Espagne reçoivent le même pourcentage de connectivité que les locuteurs de l'espagnol et aucun avantage de différenciation ne peut être analysé ; il en est de même avec le créole martiniquais ou le corse en France ou avec les nombreuses langues de l'Inde. On comprend intuitivement que cette hypothèse ne se vérifie pas dans de nombreux cas (la fracture numérique nationale est souvent liée à des considérations linguistiques) et que l'impact de ce biais est d'autant plus fort que la population considérée est faible. Un effet marginal est attendu si la méthode est appliquée à une population de locuteurs supérieure à 5 millions (bien que dans le cas de l'Inde ce ne soit peut-être pas si évident). Le prochain lancement du modèle, qui devrait se terminer avant la fin de 2021, tentera de repousser la limite vers les langues comptant plus d'un million de locuteurs.

D'autres biais marginaux du modèle peuvent résulter de l'adoption de structures impliquées par les sources principales. Par exemple, la division en pays a été dérivée de la classification de l'UIT et ne distingue pas certains territoires et leur attribue donc le même pourcentage que le pays de rattachement (si le taux de connectivité d'un territoire rattaché est de fait très inférieur à celui du pays de rattachement, ce qui pourrait être le cas par exemple de Mayotte, les langues spécifiques à ce territoire, dans cet exemple le kibushi, bénéficieront d'un biais favorable).

5.2 Biais de la sélection des sources

Il y a évidemment un « biais de sélection », qui n'est pas propre à la méthodologie mais appartient à l'application de la méthode, où la décision sur la sélection des sources privilégie implicitement des critères propres à l'origine culturelle de l'auteur et ignore inconsciemment les données de pays trop éloignés de son expérience. Cela peut s'appliquer à chacun des indicateurs et avoir un impact particulier sur l'indicateur *trafic* où la sélection de sites Web a une influence certaine, même si le nombre de sites Web se compte par centaines. L'utilisation de la moyenne tronquée à 20 % a été mise en œuvre pour réduire de tels biais, après avoir vérifié que 20 % était un intervalle important capable d'éliminer la grande majorité des résultats centrés sur des sites Web à forte localité linguistique.

5.3 Les biais des sources

Les biais résultant des sources sont discutés dans le tableau ci-dessous, en notant chaque indicateur avec une valeur de 0 (totalement biaisé) à 20 (absence totale de biais).

Tableau 11: Évaluation du biais par indicateur

INDICATEUR	ÉVALUATION	COMMENTAIRES
INTERNAUTES	19→16	Cet indicateur dérive d'un micro-indicateur unique. La principale source est l'UIT. En 2017, il s'agissait de la source la mieux notée avec un 19/20, mais dans cette

		version, la note tombe à 16 car l'UIT a cessé de fournir sa propre estimation lorsque le pays ne produit pas de données officielles. Les chiffres de l'UIT ont été complétés par ceux de la Banque mondiale et une projection linéaire des données des années précédentes a été établie pour les autres cas. Cet indicateur est essentiel dans la méthode car il sert à pondérer les résultats dans plusieurs situations, cependant l'analyse factorielle a montré que l'impact d'une faible variation est modéré. À titre d'exemple, si le taux de personnes connectées pour le Brésil était fixé à 80 % au lieu de la valeur réelle de 74 %, la *puissance* du portugais passerait de 3,26 % à 3,39 %.
INDEX	15→18	Cet indicateur dérive d'un mélange de 25 micro-indicateurs évaluant différents paramètres de pays caractérisant la société de l'information. Les sources sont des organisations internationales, des ONGs ou des universités. Les biais, s'ils existent, sont marginaux. Le biais de sélection est ici extrêmement faible car on est proche de l'exhaustivité pour l'ensemble des micro-indicateurs.
CONTENUS	5→8	Il n'y a que 13 micro-indicateurs pour construire cet indicateur et 11 d'entre eux proviennent de Wikimedia. La répartition des *contenus* du Web par langue est un continent caché de l'Internet et les sources existantes sont, extrêmement rares et trop souvent biaisées. Malheureusement, le modèle n'échappe pas à cette situation dans son état actuel. Comme il s'appuie fortement sur les excellentes statistiques de Wikimedia, l'indicateur porte les biais de Wikimedia où la présence de langues asiatiques est bien inférieure à leur proportion dans la vie réelle. De toute évidence, le biais de sélection dans ce cas, qui dépend énormément des statistiques de Wikimedia, est extrêmement important. Un système de pondération a été mis en place pour réduire au maximum cette dépendance (ce qui de toute façon n'est certainement pas suffisant, c'est pourquoi la note est passée d'un très faible 5 à un insuffisant 8). Le biais propre à l'indicateur de *contenus* est de plus assez sensible (c'est à dire que les variations produisent de forts impacts dans les résultats) comme en témoigne l'expérimentation conduite en jouant avec le système de pondération. Quelques idées pour tenter de remédier à ce problème seront mises en œuvre lors de la prochaine édition. Pendant ce temps, les biais sont surmontés « à la main » en utilisant certaines techniques (voir Correction des biais).
TRAFIC	13→11	Cet indicateur est issu de la mesure du trafic par pays à l'aide d'Alexa.com sur une sélection de 338 sites du Web. En 2017, l'analyse des biais a montré que cette source était fortement biaisée en défaveur des pays asiatiques et du Brésil. En 2021, il apparaît que le biais contre les pays asiatiques a été corrigé (peut-être trop dans le cas de

		l'Inde!) et de nouveaux biais sont détectés défavorisant désormais les pays européens. Le biais de sélection est évident dans ce cas et la prochaine version augmentera sérieusement le nombre de sites mesurés. La possibilité de fusionner dans des proportions égales les résultats de Semrush et d'Alexa doit être explorée afin de contenir les biais existants.
INTERFACES	19	Ce sont des données objectives (présence ou non d'une langue dans l'interface d'une application ou comme cible d'un service de traduction en ligne). Le biais de sélection peut exister et il peut être nécessaire d'allonger la liste mais son impact est marginal. Intuitivement, on perçoit une augmentation, par rapport à 2017, du nombre de langues prises en charge dans les interfaces ou pour la traduction en ligne; cependant, cela reste un « indicateur radical » qui laisse de côté la grande majorité des langues du monde et se concentre vers un sous-ensemble très limité.
USAGE	12	Cet indicateur repose principalement sur des données d'abonnement aux réseaux sociaux par pays. Alors que les données collectées peuvent être considérées comme fiables, la méthode implique un biais défavorisant les pays non-occidentaux ayant des applications alternatives à Facebook, Twitter, Linkedin, etc. La prochaine campagne de mesure tentera d'identifier et intégrer les populations d'abonnés d'applications alternatives pour équilibrer les résultats et essayer de réduire le biais. Pendant ce temps, la correction du biais doit être effectuée à la main. Le biais de sélection n'existe pas vraiment car la sélection est dictée par l'étroitesse des options existantes. La prochaine version bénéficiera d'un petit budget pour la base de données payante qui devrait permettre d'étendre le nombre de micro-indicateurs.

Si la pondération présentée dans le tableau précédent est appliquée aux résultats dans la construction de la moyenne pondérée des macro-indicateurs de *puissance* (au lieu de la moyenne simple), de manière à tenir compte de la confiance relative envers les différents indicateurs du modèle, les ajustements suivants sont constatés sur les résultats(en couleur), à comparer avec les résultats précédents (sans couleur), pour en tenir compte au moment de la correction des biais.

Tableau 12 : Indicateurs macro pour les 15 premières langues après pondération des indicateurs

	PUISS.	**Capac.**	**Grad.**	**PUISS.**	**Capac.**	**Grad.**	**Effet**
Anglais	24,23%	1,86	1,58	26,48%	2,04	1,73	---
Chinois	15,77%	1,07	0,89	13,92%	0,95	0,79	+++
Espagnol	8,80%	1,68	1,26	8,73%	1,67	1,25	+
Hindi	3,63%	0,63	0,85	3,38%	0,58	0,79	+++
Français	3,62%	1,40	1,21	3,71%	1,44	1,24	-
Portugais	3,37%	1,36	1,10	3,35%	1,35	1,10	+

Arabe	3,28%	0,93	0,85	3,09%	0,88	0,80	++
Russe	3,24%	1,30	0,92	3,11%	1,25	0,88	++
Allemand	2,72%	2,08	1,30	2,86%	2,19	1,37	--
Japonais	2,51%	2,06	1,22	2,52%	2,07	1,22	
Malais	1,87%	0,79	0,85	1,76%	0,75	0,80	++
Turc	1,27%	1,49	1,05	1,24%	1,46	1,02	+
Italien	1,23%	1,88	1,36	1,37%	2,09	1,51	--
Coréen	0,97%	1,24	1,04	0,96%	1,22	1,03	
Bengali	0,91%	0,35	0,79	0,88%	0,34	0,78	+

5.3.1 Les biais de Wikimédia

Les statistiques de Wikipédia sont impeccables, cependant, il faut comprendre que bien qu'il s'agisse de l'une des applications de l'Internet les plus mondialisées, ses statistiques montrent des chiffres pour certaines langues asiatiques qui sont bien en deçà de leur présence relative dans l'Internet. Le tableau suivant compare les ratios entre le nombre d'articles Wikipédia et le nombre d'internautes; des écarts énormes avec des valeurs anormalement basses pour les langues asiatiques apparaissent (sauf exceptions remarquables).

Tableau 13: Trié par nombre d'articles Wikipédia

Langue	Articles	% ART TOTAL.	Pondération %	Art./L1+L2
Anglais	6332139	12,92%	0,28%	7
Cebuano	5853095	11,94 %	22,16%	851
Suédois	3050759	6,22%	14,11%	250
Allemand	2593827	5,29%	1,16%	22
Arabe	2433772	4,97 %	0,40%	11
Français	2342875	4,78%	0,53%	14
Néerlandais	2060512	4,20%	5,10%	92
Chinois	1752600	3,58%	0,07%	2
Russe	1736736	3,54%	0,41%	9
Italien	1703284	3,47%	1,51%	33
Espagnol	1698331	3,46%	0,19%	4
Serbo-croate	1514114	3,09%	4,57%	97
Polonais	1480982	3,02%	2,20%	45
Japonais	1277204	2,61 %	0,61%	11
Vietnamien	1266628	2,58%	1,00 %	24
Ukrainien	1100281	2,24%	2,00%	52
Portugais	1067241	2,18%	0,25%	6
Malais	936876	1,91 %	0,23%	8
Persan	816984	1,67%	0,59%	15
Coréen	543656	1,11%	0,40%	10
Finlandais	512026	1,04%	5,36%	99
Hongrois	489514	1,00 %	2,36%	49
Tchèque	484445	0,99%	2,18%	44
Roumain	421153	0,86%	1,06%	23
Arménien	420677	0,86%	6,60%	156
Azerbaïdjanais	420677	0,86%	1,06%	24
Turc	410954	0,84%	0,28%	6

Tatar	299494	0,61%	3,42%	73
Hébreu	298053	0,61%	1,92 %	37
Biélorusse	281379	0,57%	4,34%	87
Bulgare	273163	0,56%	2,00%	47
Danois	267641	0,55%	2,88%	49
Slovaque	237210	0,48%	1,98%	40
Kazakh	228493	0,47%	1,05%	23
Grec	195481	0,40%	0,89%	19
Ourdou	164062	0,33%	0,04%	3
Hindi	148545	0,30%	0,01%	1
Ouzbek	140894	0,29%	0,25%	9
Tamil	138490	0,28%	0,10%	4
Thaïlandais	137351	0,28%	0,14%	3
Bengali	109438	0,22%	0,02%	2

À noter, la présence du cebuano en deuxième position et la présence relative du chinois et des langues de l'Inde. Il est utile de vérifier un pourcentage pondéré en fonction du nombre de locuteurs L1+L2: l'anglais n'apparaît pas disproportionné et certaines langues semblent avoir une forte présence par rapport à leur population L1+L2, par ordre d'importance: cebuano, suédois, arménien, finnois, néerlandais, serbo-croate, biélorusse et tatar, pour les premiers.

Wikimedia est probablement à la fois l'espace virtuel avec la plus grande diversité linguistique et le seul qui fournit systématiquement des statistiques linguistiques fiables et claires sur toutes ses activités. Ajoutant l'importance centrale de sa fonction dans le Web, il s'agit sans aucun doute d'un indicateur incontournable lorsqu'il s'agit de contenus. Malheureusement, une analyse sérieuse montre que cet espace si particulier ne pourrait en aucun cas refléter une indication fidèle de la répartition des contenus par langue dans la Toile. L'importance des langues dans Wikimédia n'est pas toujours liée à leur importance réelle dans le cyberespace et certaines langues ont investi massivement cet espace, indépendamment de leur présence globale sur le Web. Ceci est clairement visible à travers les différents indicateurs Wikimedia présentés ci-après, montrant les premières positions.

Comme expliqué précédemment, le nombre d'articles n'est pas un excellent indicateur car, pour certaines langues, des bots ont été implémentés qui ont créé des articles à partir de traductions qui par la suite ne sont pas maintenus. Pour contrôler cela, il faut faire attention au nombre d'éditeurs actifs, au nombre d'éditions au cours d'une année donnée et à la profondeur, un indicateur créé pour refléter le degré d'actualisation des articles. Une formule a été élaborée pour intégrer ces facteurs et présentée précédemment. Les résultats triés par cette formule et présentés en pourcentage sont les suivants :

Tableau 14: Articles Wikipédia triés par formule

Anglais	53,96%
Hébreu	10,32%
Italien	6,99%
Serbo-croate	6,02%
Allemand	5,70%
Français	2,96%
Suédois	2,55%
Finlandais	2,39%
Néerlandais	1,54%
Japonais	0,97%
Polonais	0,87%
Arménien	0,84%
Hongrois	0,77%
Tchèque	0,63%
Ukrainien	0,54%
Danois	0,50%
Russe	0,48%
Persan	0,40%
Biélorusse	0,29%
Espagnol	0,29%
Portugais	0,18%
Arabe	0,16%
Roumain	0,13%
Bulgare	0,11%
Coréen	0,10%
Turc	0,10%
Grec	0,07%
Slovaque	0,04%
Cebuano	0,03%
Azerbaïdjanais	0,02%
Malais	0,02%
Thaïlandais	0,01%
Chinois	0,01%
Malayalam	0,00%
Kazakh	0,00%
Afrikaans	0,00%
Tatar	0,00%
Bengali	0,00%
Mongol	0,00%
Tagalog	0,00%

Il s'agit clairement d'une représentation plus juste de la réalité avec Wikipédia, en accordant une attention équilibrée au nombre d'éditeurs, d'éditions et profondeurs, et pondérées en fonction du nombre de locuteurs L1+L2. À noter que le cebuano est pénalisé avec cette formule pour sa politique d'utilisation de bots mais qu'une autre langue des Philippines arrive à se glisser

dans le tableau de tête : le tagalog ! La prédominance de l'anglais sur Wikimedia apparaît aussi plus clairement avec cette approche.

Wikimedia ne se résume pas à Wikipedia et des statistiques existent aussi pour chacun des autres indicateurs : WikiBooks, WikiQuote, WikiSource, Wikiversity, Wiktionnary, WikiNews et WikiVoyages, pour lesquels le nombre d'articles par langue est accessible. Pour ces éléments de Wikimedia, les sources sont présentées dans l'absolu, sans pondération en fonction du nombre de locuteurs, ne montrant que les premiers.

Tableau 15: Nombre de Wikibooks

Anglais	3851195	35,72 %
Allemand	961696	8,92%
Français	657991	6,10%
Portugais	473196	4,39%
Italien	411671	3,82%
Polonais	403336	3,74%
Hongrois	401256	3,72%
Espagnol	396546	3,68%
Néerlandais	349987	3,25%
Vietnamien	256386	2,38%
Russe	205469	1,91 %
Japonais	178783	1,66%
Arabe	174452	1,62%
Hébreu	164355	1,52%
Chinois	141302	1,31%
Finlandais	131314	1,22%
Persan	112964	1,05%
Malais	89019	0,83%
Hindi	73969	0,69%

Tableau 16: Nombre de citations (WikiQuote)

Anglais	33897	14,28%
Italien	30799	12,98%
Polonais	28960	12,20%
Russe	13148	5,54%
Tchèque	9263	3,90%
Persan	8495	3,58%
Allemand	7879	3,32%
Portugais	7443	3,14%
Espagnol	7116	3,00%
Serbo-croate	7022	2,96%
Français	5923	2,50%
Ukrainien	5798	2,44%
Slovaque	4547	1,92 %
Turc	4503	1,90 %
Bulgare	4389	1,85%
Hébreu	4202	1,77%

Tableau 17: Nombre de Wikisources

Français	2609546	25,3%
Anglais	2204231	21,3%
Chinois	778716	7,5%
Bengali	722295	7,0%
Polonais	669381	6,5%
Russe	642705	6,2%
Allemand	431714	4,2%
Italien	415032	4,0%
Tamil	411502	4,0%
Hébreu	214947	2,1%
Suédois	84882	0,8%
Arabe	80708	0,8%
Multilingue	78809	0,8%
Arménien	75487	0,7%
Portugais	73139	0,7%

Tableau 18: Nombre de Wikiversité

Allemand	49011	36,9%
Anglais	38612	29,0%
Français	17553	13,2%
Russe	5883	4,4%
Tchèque	5195	3,9%
Portugais	4692	3,5%
Italien	4472	3,4%
Espagnol	2662	2,0%
Finlandais	1914	1,4%
Slovène	1252	0,9%
Suédois	858	0,6%
Grec	644	0,5%
Japonais	207	0,2%

Tableau 19: Nombre d'entrées du Wiktionnaire

Anglais	5923218	19,2%
Malgache	5466228	17,7%
Français	3392407	11,0%
Chinois	1239843	4,0%
Serbo-croate	1177979	3,8%
Russe	1002462	3,2%
Espagnol	885649	2,9%
Allemand	737337	2,4%
Néerlandais	686499	2,2%
Suédois	674872	2,2%
Polonais	649612	2,1%
Kurde	635201	2,1%
Lituanien	616313	2,0%

Grec	462897	1,5%
Italien	434058	1,4%
Coréen	398737	1,3%
Finlandais	374056	1,2%

Il est important d'essayer de comprendre ce qui s'est passé avec le malgache et de se demander si son classement anormal en troisième position dans le macro-indicateur *gradient* invalide la méthode. Cette langue occupe la deuxième place dans Wiktionnaire et affiche un pourcentage du total d'entrées de 17 %, extrêmement disproportionné par rapport à sa population (18 millions de locuteurs) et bien plus encore par rapport à son très faible nombre de locuteurs connectés (1,8 million). Même si le poids de ce micro-indicateur a été fixé à 0,1 (le même que tous les services de Wikimedia, sauf la formule Wikipedia et les Wikibooks) la disproportion est si énorme qu'elle affecte une moyenne pondérée avec seulement 9 éléments et, en cascade, les macro-indicateurs *puissance* et *gradient*. Cette situation ne délégitime pas la définition de *gradient* mais elle est bien sûr un symptôme de la faiblesse de l'indicateur *contenu*.

Tableau 20: Nombre de Wikinews

Anglais	21687	14,9%
Français	20761	14,3%
Russe	17649	12,1%
Polonais	14357	9,9%
Espagnol	11312	7,8%
Chinois	8559	5,9%
Arabe	7578	5,2%
Serbo-croate	5650	3,9%
Tchèque	5608	3,9%
Catalan	4056	2,8%
Tamil	3363	2,3%
Suédois	3317	2,3%
Grec	3084	2,1%
Ukrainien	1738	1,2%
Roumain	1697	1,2%
Persan	1645	1,1%
Bulgare	1562	1,1%
Portugais	1474	1,0%
Allemand	1386	1,0%

Tableau 21: Nombre d'articles dans Wikivoyages

Anglais	28852	28,1%
Allemand	16545	16,1%
Persan	8674	8,5%
Italien	7619	7,4%
Français	7407	7,2%
Polonais	6946	6,8%
Russe	5438	5,3%
Néerlandais	3671	3,6%

Portugais	3624	3,5%
Chinois	2972	2,9%
Espagnol	2524	2,5%
Hébreu	2072	2,0%
Vietnamien	1624	1,6%
Suédois	1522	1,5%
Grec	1408	1,4%
Roumain	917	0,9%
Ukrainien	779	0,8%

La diversité des résultats selon chaque sujet empêche de tirer une conclusion systématique de l'analyse de ces chiffres, cependant quelques affirmations générales peuvent être faites :

- L'anglais est généralement, mais pas toujours, à la première place, quoique la proportion d'anglais soit moins prédominante que pour Wikipédia, et reste comprise entre 14 % et 36 %, avec une moyenne de 23,5 % (contre 29,4 % dans les indicateurs Wikipédia)[33].

- Le français et l'allemand obtiennent des scores élevés dans la plupart des indicateurs de Wikimedia.

- Le chinois, l'hindi, le bengali et le persan font leur chemin dans certains des indicateurs

- Certaines langues inattendues apparaissent en tête de liste pour certains indicateurs : le malgache et le tamil (outre le cebuano).

En conclusion, Wikimédia reste de loin l'endroit le plus diversifié linguistiquement de l'Internet avec quelques langues minoritaires réussissant à obtenir un score élevé, ce dont il faut se réjouir, mais cela ne reflète guère la réelle diversité des contenus sur le Web. L'anglais est largement prédominant, mais pas autant qu'auparavant. Dans tous les cas, la méthode doit en priorité améliorer la qualité de l'indicateur de *contenus* qui reste préoccupante. Une approche à explorer est celle d'identifier les applications analogues à Wikimedia qui ont réussi à occuper une niche dans d'autres espaces linguistiques (en particulier en Asie) et d'une certaine manière les introduire dans les statistiques.

5.3.2 Les biais d'Alexa

Le tableau suivant présente les différents tests et comparaisons réalisés entre Alexa et Semrush et, pour Alexa, entre les deux années d'utilisation (2017 et 2021). Pour Alexa, les chiffres de trafic capturés en 2017 ont été insérés dans le nouveau modèle 2021, afin de garantir une comparaison équitable. La comparaison ne se fait pas à partir des entrées (par pays) mais à partir des sorties du modèle (par langue) ; en d'autres termes, la comparaison est faite avec le produit du modèle insérant en entrée chacun des chiffres de trafic respectifs. Les comparaisons mettent en évidence (en rouge dans le tableau) de nombreuses anomalies.

[33] Ces pourcentages se réfèrent au nombre d'articles pour l'anglais divisé par le nombre total.

Tableau 22: Comparaisons de différentes mesures de trafic

	SEMRUSH 2021	ALEXA 2021	2021 (S-A)/S	ALEXA 2017	A21-A17/A21
Anglais	52,50%	35,83%	32%	45,40%	-27%
Chinois	1,88%	7,67%	-308%	4,94%	36%
Espagnol	14,45%	10,14%	30%	7,53%	26%
Français	4,48%	2,56%	43%	6,35%	-148%
Russe	1,88%	1,83%	3%	1,68%	8%
Allemand	2,61 %	1,33%	49%	2,94%	-122%
Portugais	2,18%	1,46 %	33%	1,63%	-12%
Arabe	1,02%	2,51%	-145%	2,54%	-1%
Hindi	1,26%	5,37%	-327%	1,60%	70%
Japonais	0,65%	1,94 %	-198%	1,90 %	2%
Malais	0,68%	0,98%	-44%	1,23%	-27%
Italien	0,89%	0,53%	41%	0,91%	-72%
Turc	0,60%	1,03%	-74%		
Polonais	0,47%	0,31%	34%	0,63%	-100%
Coréen	0,50%	0,90%	-78%	0,72%	20%
RESTE	13,95%	25,34%	-82%	18,99%	25%
TOTAL	100,00 %	100,00 %	0%	100,00 %	0%

1) Il est clair que Semrush et Alexa ne reflètent pas la même répartition du trafic par pays pour le même ensemble de sites Web. Les écarts sont importants dans de trop nombreux cas.

2) Alexa a corrigé le tir par rapport à son biais négatifs avec les pays asiatiques, en revanche, cette fois c'est Semrush qui semble ignorer les pays asiatiques et arabes.

3) En comparant les résultats d'Alexa de 2017 à 2021, on s'attendrait à des changements évolutifs donc relativement mineurs. Ce n'est pas le cas pour les langues suivantes : français, allemand, italiens et polonais dont les chiffres ont chuté de manière suspecte, une confirmation de la sensation, éprouvée lors des mesures, que les pays européens étaient sous-estimés dans les chiffres d'Alexa2021.

Enfin, ces comparaisons tendent à confirmer des situations qui seront pris en compte au moment de la correction des biais :

- L'anglais, l'espagnol, l'hindi paraissent surestimés
- Le français, l'allemand, l'italien et le polonais semblent sérieusement sous-estimés
- Le portugais et le malais semblent sous-estimés

Pour la prochaine édition, une forte attention doit être accordée à cet indicateur pour essayer de surmonter la situation, peut-être qu'une fusion des données des services existants pourrait être une alternative pour compenser les biais ?

5.4 Correction des biais

À ce stade, il n'est pas question d'appliquer la correction de biais à toutes les langues de l'étude mais de se concentrer uniquement sur les 15 premières langues en termes de *puissance*. Dans le futur, il serait intéressant d'intégrer la correction des biais dans le modèle (une première marche a été expérimentée avec la pondération des indicateurs en fonction de la confiance accordée).

Il existe une méthode qui a été utilisée en 2017 pour produire une estimation du pourcentage de contenus qui est basée sur la cohérence du facteur de productivité (ratio contenus sur population connectée) pour chaque langue considérée et, très important, pour le reste des langues. Cette méthode appliquée en 2021 conduit à l'estimation approximative suivante :

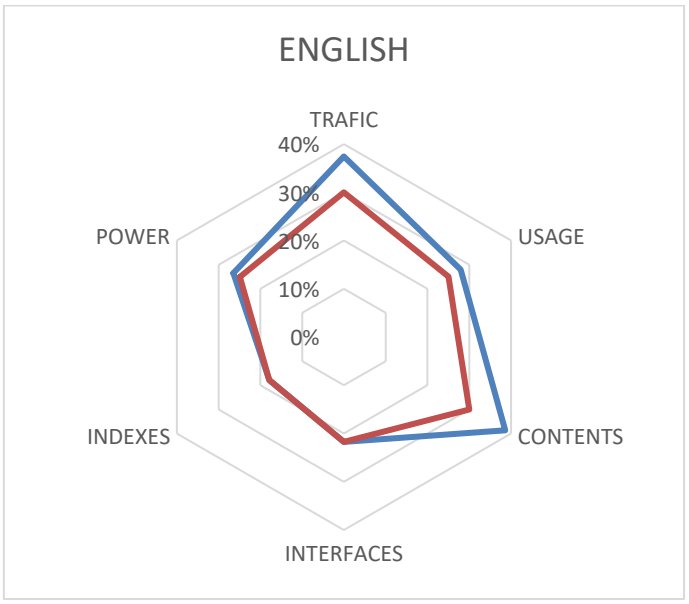
Tableau 23: Première méthode de correction du biais

LANG.	CONTENU	PRODUCTIVITÉ
Anglais	**25,0%**	1,92
Chinois	**15,0%**	1,02
Espagnol	**7,0%**	1,34
Français	**4,0%**	1,55
Hindi	**4,0%**	0,69
Portugais	**3,5%**	1,41
Russe	**3,5%**	1,41
Arabe	**2,5%**	0,71
Allemand	**2,5%**	1,92
Japonais	**2,5%**	2,05
Malais	**1,8%**	0,76
Italien	**1,4%**	2,14
Turc	**1,2%**	1,41
Coréen	**1,2%**	1,53
Bengali	**1,2%**	0,46
Vietnamien	**0,70%**	0,94
RESTE	**23,00%**	**0,58**

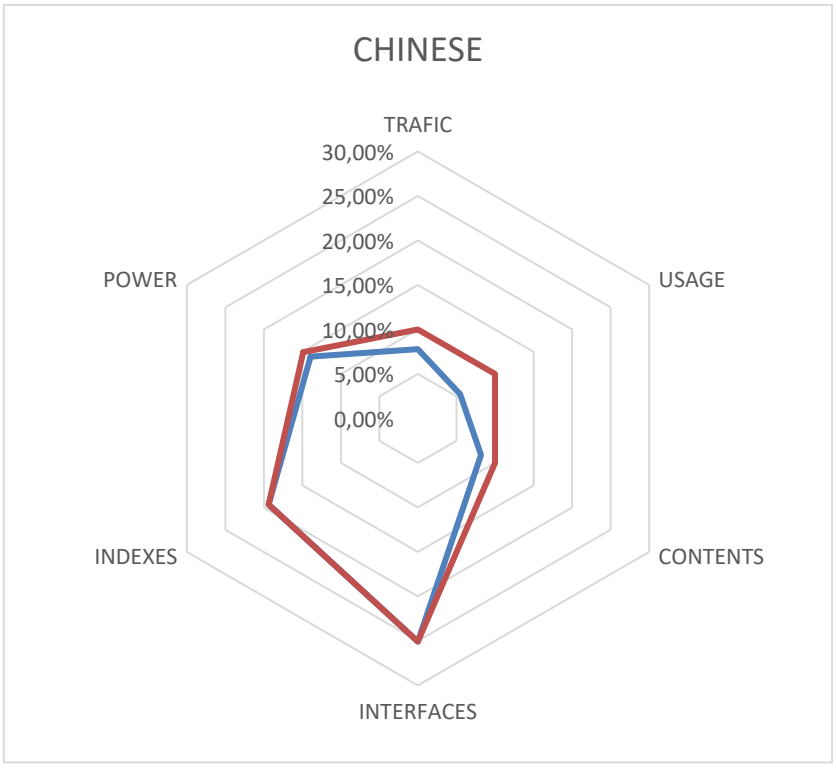
Cette fois, une nouvelle approche de la correction des biais a été ajoutée, travaillant spécifiquement et directement sur les biais respectifs de chaque indicateur, tels qu'ils ont été commentés dans les chapitres précédents. Le schéma du résultat sur la langue est examiné, indicateur par indicateur, à la lumière de ce que l'on sait des biais, et un nouveau chiffre possible est consigné. À partir de là, un nouveau chiffre de « puissance » est calculé avec des valeurs arrondies. En bleu, les données produites par le modèle, en rouge les corrections établies.

Tableau 24: Correction des biais 2ème méthode

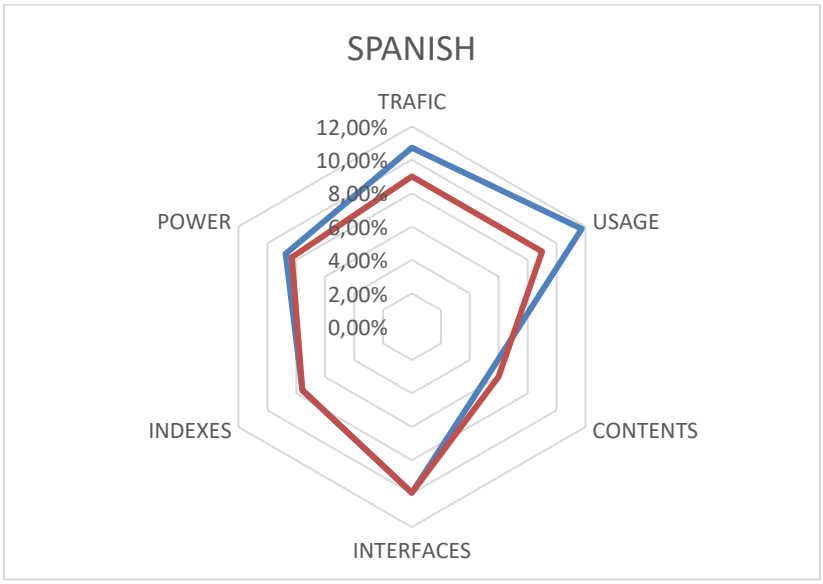
Anglais	TRAFIC	USAGE	CONTENU	INTERFACES	INDEXS	PUISSANCE
MODELE	37,44%	27,92%	38,61%	,21,73%	17,87%	**26,48%**
CORRECTION DE BIAIS	30%	25%	30%	22%	18%	25%



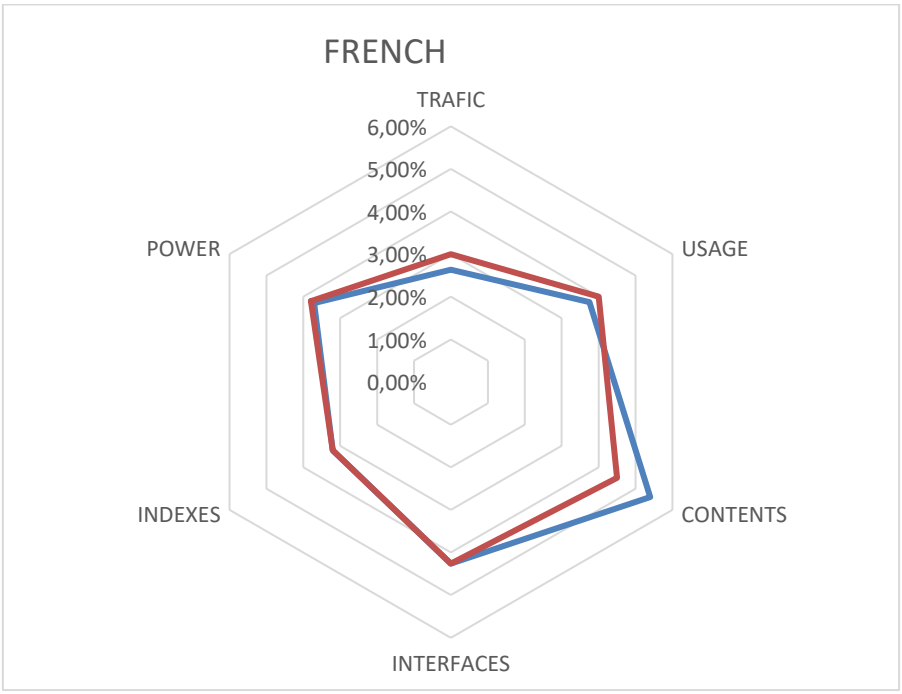
chinois	TRAFIC	USAGE	CONTENU	INTERFACES	INDEXS	PUISSANCE
MODELE	7,79 %	5,47%	8,18%	25,07%	19,38%	**13,92%**
CORRECTION DE BIAIS	10%	10%	10%	25%	19%	15%



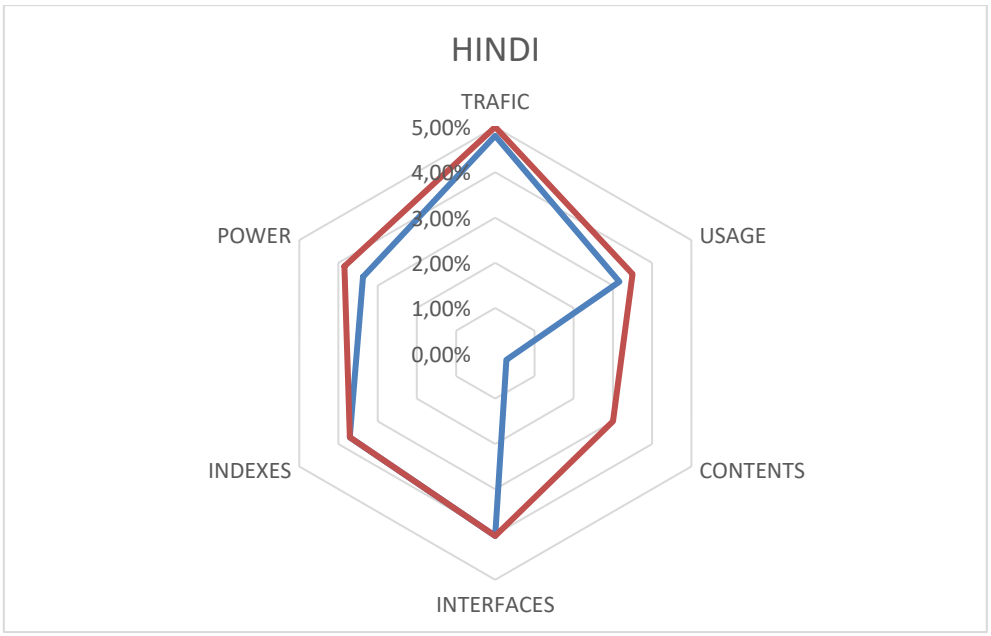
Espagnol	TRAFIC	USAGE	CONTENU	INTERFACES	INDEXS	PUISSANCE
MODELE	10,72%	11,74%	5,42%	9,94 %	7,59%	**8,73 %**
CORRECTION DE BIAIS	9%	9%	6%	dix%	8%	8%



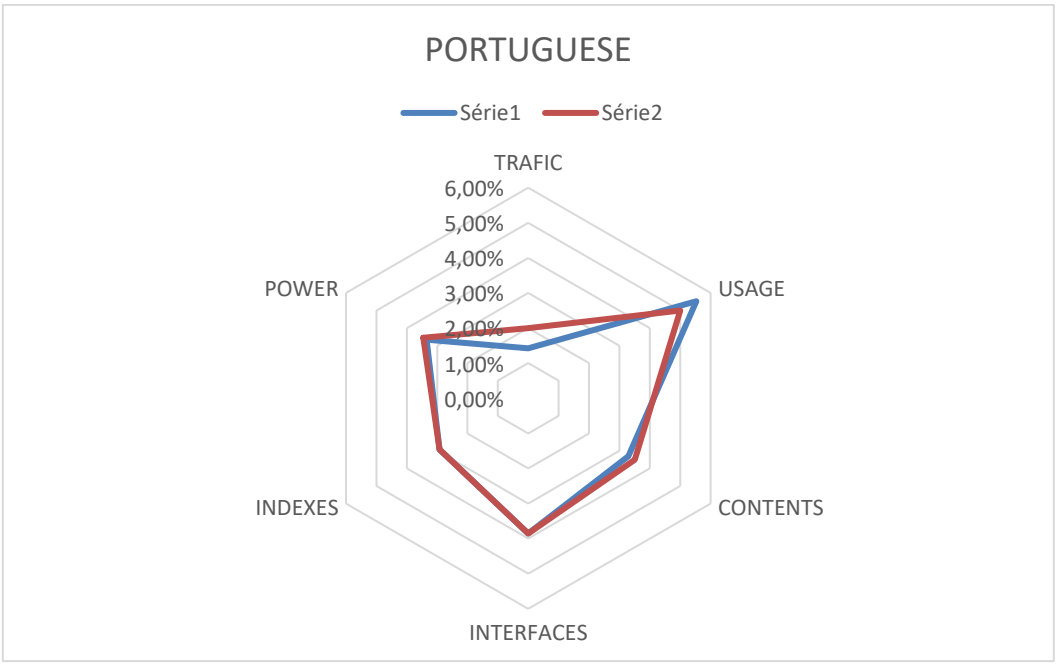
Français	TRAFIC	USAGE	CONTENU	INTERFACES	INDEXS	PUISSANCE
MODELE	2,64 %	3,75%	5,40%	4,26%	3,21%	**3,71%**
CORRECTION DE BIAIS	3,0%	4,0%	4,5%	4,3%	3,2%	3,8%



Hindi	TRAFIC	USAGE	CONTENU	INTERFACES	INDEXS	PUISSANCE
MODELE	4,81%	3,16%	0,28%	4,03%	3,71%	3,38%
CORRECTION DE BIAIS	5,0%	3,5%	3,0%	4,0%	3,7%	3,8%



Portugais	TRAFIC	USAGE	CONTENU	INTERFACES	INDEXS	PUISSANCE
MODELE	1,42%	5,53%	3,30%	3,85%	2,92%	3,35%
CORRECTION DE BIAIS	2,0%	5,5%	3%	3,9%	2,9%	3,5%



Le résultat de cet exercice de correction de biais est présenté ci-après et comparé aux résultats de la première méthode de correction :

Tableau 25: Résultats de la correction du biais

	SECONDE PUISSANCE	MÉTHODE CONTENU	*PUISSANCE PREMIÈRE MÉTHODE*
Anglais	25%	30,0%	*25%*
Chinois	15%	10%	*15%*
Espagnol	8%	6%	*7%*
Français	3,8%	4,5%	*4%*
Hindi	3,8%	3,0%	*4%*
Portugais	3,5%	2,8%	*3,5%*

Fait intéressant, les résultats des deux méthodes différentes sont assez proches.

6. CONCLUSIONS ET PERSPECTIVES

Cette seconde version de la méthode de production d'indicateurs de présence des langues dans l'Internet montre des améliorations importantes, notamment avec des données démolinguistiques plus fiables et dans la gestion des secondes langues et du multilinguisme. Elle progresse également avec une approche cohérente d'établissement du pourcentage mondial par rapport au nombre total de locuteurs L1+L2. Elle présente désormais un indicateur *index* plus complet. La méthode a amélioré l'analyse des biais produits en approfondissant celle des statistiques Wikimédia et présentant deux voies complémentaires pour compenser en partie ces biais.

La méthode rencontre cependant de nouveaux défis importants:

- ✓ avec le comportement divergents des outils de mesure du trafic ;
- ✓ avec un indicateur de *contenus* trop dépendant des chiffres de Wikimedia lesquels ne sont pas géographiquement homogène, ne reflètent pas vraiment la réalité des contenus et dont la forte sensibilité donne une influence disproportionnée, en particulier sur le macro-indicateur *gradient* ;
- ✓ avec un indicateur *usages* trop marqué par les applications occidentales des réseaux sociaux ;
- ✓ et avec le fait que l'UIT ne fournit plus d'estimations pour le pourcentage de personnes connectées à l'Internet par pays (et un problème particulier sur le pourcentage exact pour l'Inde).

Il est prévu une nouvelle version avant la fin de 2021 qui tentera de relever ces défis et d'essayer d'élargir le nombre de langues traitées, repoussant la limite aux langues comptant plus d'un million de locuteurs L1.

L'objectif de la future version sera également d'étendre le nombre de sites Web mesurés en termes de trafic afin de fournir des résultats différenciés par thèmes plus précis et plus fiables pour certaines langues.

Du côté des résultats, la tendance à la réduction relative de la dominance de l'anglais se poursuit, avec désormais une présence estimée autour de 25 % (contre 30 % en 2017), la croissance du chinois et l'apparition de l'hindi comme probable quatrième langue d'Internet, avec le français aujourd'hui, et probablement au-dessus du français dans les années à venir.

RÉFÉRENCES

[1] D. Pimienta, « Une approche alternative pour produire des indicateurs de langues dans l'Internet », 2017
http://funredes.org/lc2017/Alernative%20Languages%20Internet.docx

[2] - MAAYA, « NET.LANG : Vers un cyberespace multilingue », Éditions C&F, 2012 -http://net-lang.net/lang_fr

[3] - D. Pimienta, D. Prado, A. Blanco, « Douze ans de mesure de la diversité linguistique dans l'Internet: bilan et perspectives », UNESCO, 2009 - http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

[4] - J. Paolillo, D. Pimienta, D. Prado, et al., « Mesurer la diversité linguistique sur Internet », UNESCO,/2005- http://unesdoc.unesco.org/images/0014/001421/142186f.pdf

[5] – D. Pimienta, "La présence de la langue française dans le cyberespace (synthèse)", Rapport 2019 "La langue française dans le monde", pp. 337-341, OIF, Gallimard, 2019 - https://www.francophonie.org/sites/default/files/2021-04/LFDM-20Edition-2019-La-langue-française-dans-le-monde.pdf
Étude complète accessible à http://observatoire.francophonie.org/2018/Place-francais-sur-Internet-D-Pimienta.pdf

[5] – D. Pimienta, "Produire diffuser et protéger les biens communs numériques, Section 2: Contribuer à la production et à la promotion de contenus francophones (en français et dans les langues nationales) et de nouveaux modes d'expression numérique.", in Rapport 2018 État des lieux de la Francophonie numérique, OIF, 2018 - https://www.francophonie.org/rapport-numerique-2018.html

[6] – D. Pimienta, D. Prado - " Un milliard de Latins… dans l'Internet ?" in Hermès, La Revue, 2016/2 (n° 75) Langues romanes : un milliard de locuteurs
http://www.cairn.info/revue-hermes-la-revue-2016-2.htm

[7] – D. Pimienta, D. Prado "Le français dans l'Internet", Rapport 2014 "La langue française dans le monde", pp. 501-541, OIF, Nathan, 2014 - http://www.francophonie.org/Rapports-Publications.html

[8] – D. Pimienta, D. Prado, "Étude sur la place des langues de France dans l'Internet", Ministère de la culture de France, 2014.
http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et-la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet

ANNEXE 1. LISTE DES MICRO INDICATEURS ET SOURCES

MICRO-INDICATEUR	TYPE	THÈME	URL DE LA SOURCE
Amazon US - nombre de livres 2017	CONTENU	Book	Reprise de 2017
Valeur de la profondeur Wikipedia	CONTENU	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Nombre d'utilisateurs actifs de Wikipédia	CONTENU	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Nombre de modifications Wikimedia	CONTENU	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Nombre de livres Wiki par langue	CONTENU	Book	https://meta.wikimedia.org/wiki/Wikibooks/Table
Nombre d'articles Wikipédia par langue	CONTENU	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Nombre d'articles WikiQuote par langue	CONTENU	Book	https://stats.wikimedia.org/wikiquote/FR/Sitemap.htm
Nombre d'articles WikiSource par langue	CONTENU	Book	Nombre d'articles WikiSource par langue
Nombre d'articles Wikiversité par langue	CONTENU	S/T	https://stats.wikimedia.org/wikiversity/EN/Sitemap.htm
Nombre d'articles Wiktionnaire par langue	CONTENU	Dict	https://stats.wikimedia.org/wiktionary/EN/Sitemap.htm
Nombre d'articles WikiNews par langue	CONTENU	News	https://stats.wikimedia.org/wikinews/EN/Sitemap.htm
Nombre d'articles WikiVoyages par langue	CONTENU	Tur	https://stats.wikimedia.org/wikivoyage/EN/Sitemap.htm
T-Index pour le commerce électronique Projection 2021	CONTENU	e.com	https://translated.com/les-langues-qui-comptent
Index de l'administration en ligne	INDEX	S/T	https://publicadministration.un.org/egovkb/Data-Center*
Index de participation électronique	INDEX	S/T	https://publicadministration.un.org/egovkb/Data-Center
Index des services en ligne	INDEX	Infra	https://publicadministration.un.org/egovkb/Data-Center
Index de capital humain	INDEX	ICT	https://publicadministration.un.org/egovkb/Data-Center
Index des infrastructures de télécommunications	INDEX	Gov	https://publicadministration.un.org/egovkb/Data-Center
Index mondial de préparation au numérique de Cisco 2019	INDEX	S/T	https://www.cisco.com/c/dam/en_us/about/csr/reports/global-digital-readiness-index.pdf
Index de préparation à l'IA du gouvernement 2020	INDEX	ICT	https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf
Scores de liberté d'Internet	INDEX	Book	https://freedomhouse.org/countries/freedom-net/scores
Index de connectivité mondiale	INDEX	Gov	https://www.huawei.com/minisite/gci/en/country-rankings.html
Index mondial de cybersécurité 2018	INDEX	Gov	https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf
Index CNUCED du commerce électronique B2C, 2020	INDEX	Gov	https://unctad.org/system/files/official-document/tn_unctad_ict4d17_en.pdf
L'index mondial des données ouvertes	INDEX	Infra	https://index.okfn.org/place/
Classement mondial de la compétitivité numérique 2020	INDEX	Secu	https://www.imd.org/globalassets/wcc/docs/release-2020/digital/digital_2020.pdf
Index de préparation pour les technologies frontalières	INDEX	Econ	https://unctad.org/system/files/official-document/tir2020_en.pdf
Index mondial de l'innovation	INDEX	AI	https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2020.pdf
Accès aux connaissances de base	INDEX	Econ	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à l'information et aux communications	INDEX	Gov	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à l'enseignement supérieur	INDEX	Gov	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à l'électricité (% de la pop.)	INDEX	Infra	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à une éducation de qualité (0=inégal ; 4=égal)	INDEX	S/T	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Accès à la gouvernance en ligne (0=faible ; 1=élevé)	INDEX	Econ	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Censure des médias (0=fréquent ; 4=rare)	INDEX	Infra	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Liberté d'expression (0=pas de liberté ; 1=pleine liberté)	INDEX	Gov	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx

Universités pondérées par la qualité (points)	INDEX	e.com	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Documents citables	INDEX	Gov	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Femmes ayant fait des études supérieures	INDEX	Econ	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Années d'études supérieures	INDEX	S/T	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Langues de traduction de Bing Translator	INTERFACE	Tra	https://www.bing.com/translator/
Langues prises en charge par Amazon Kindle Direct Publishing	INTERFACE	Inter	https://kdp.amazon.com/en_US/help/topic/G200673300
Langues prises en charge par Cortana	INTERFACE	Tra	https://en.wikipedia.org/wiki/Cortana
Langues de Word référence prises en charge	INTERFACE	Inter	https://www.wordreference.com
Langues de traduction WordLingo	INTERFACE	Inter	http://www.worldlingo.com/en/languages/
Langues prises en charge par Facebook	INTERFACE	Tra	https://www.facebook.com/langue.php
Langues des publicités Facebook In-Stream prises en charge	INTERFACE	Tra	https://www.facebook.com/business/help/267128784014981
Langues du traducteur gratuit prises en charge	INTERFACE	Tra	http://www.free-translator.com
Langues prises en charge par la console Google Play	INTERFACE	Tra	https://support.google.com/googleplay/android-developer/table/4419860?hl=fr
Langues prises en charge par Google Cloud	INTERFACE	Inter	https://cloud.google.com/translate/docs/languages?hl=fr
Langues prises en charge par Google Traduction	INTERFACE	Inter	https://en.wikipedia.org/wiki/Google_Translate
Langues prises en charge par Google Scholar pour la recherche	INTERFACE	Inter	https://scholar.google.com/scholar_settings?sciifh=1&hl=fr&as_sdt=0,5#1
Langue prise en charge par Paralink Translator	INTERFACE	Inter	http://paralink.com
Langues du traducteur en ligne prises en charge	INTERFACE	Tra	https://www.online-translator.com/traduction
Langues du traducteur Reverso prises en charge	INTERFACE	Tra	https://www.reverso.net/text_translation.aspx?lang=FR
Langues prises en charge par la traduction gratuite	INTERFACE	Tra	https://www.freetranslations.org
Langues prises en charge par Skype	INTERFACE	Inter	https://support.skype.com/en/faq/FA34781/what-languages-are-supported-in-skype
Langues prises en charge par Systran	INTERFACE	Tra	https://support.systran.net/systranlinks/faq/
163.com	TRAFIC	GAM	https://www.alexa.com/siteinfo
17ok.com	TRAFIC	?	https://www.alexa.com/siteinfo
1and1.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
360.cn	TRAFIC	Secu	https://www.alexa.com/siteinfo
4shared.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
500px.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
6.cn	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
A2hosting.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Abilogic.com	TRAFIC	DIR	https://www.alexa.com/siteinfo
À propos de moi	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Academia.edu	TRAFIC	S/T	https://www.alexa.com/siteinfo
Adam4Adam.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Adictinggames.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
adobe.com	TRAFIC	ICT	https://www.alexa.com/siteinfo
Adultfriendfinder.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Aim.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Alexa.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Aliexpress.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Alipay.com	TRAFIC	Econ	https://www.alexa.com/siteinfo
Alivedirectory.com	TRAFIC	DIR	https://www.alexa.com/siteinfo
Amazon.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Amazonaws.com	TRAFIC	Host	https://www.alexa.com/siteinfo
Anastasiadate.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Android	TRAFIC	ICT	https://www.alexa.com/siteinfo
Angel.co	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Anobii.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Answers.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Aparat.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Pomme	TRAFIC	ICT	https://www.alexa.com/siteinfo
Musique d'Apple	TRAFIC	SN-Mu	https://www.alexa.com/siteinfo
Apple.com/Safari	TRAFIC	ICT	https://www.alexa.com/siteinfo

Archives.org	TRAFIC	Book	https://www.alexa.com/siteinfo
Archives-ouvertes.fr	TRAFIC	S/T	https://www.alexa.com/siteinfo
Armorgames.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Arvixe.com	TRAFIC	Host	https://www.alexa.com/siteinfo
Arxiv.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Ashleymadison.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Ask.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Ask.fm	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Atom.io	TRAFIC	App	https://www.alexa.com/siteinfo
Avvo.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Babytree.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Badoo.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Baidu.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Bandcamp.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Bartleby.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Base-search.net	TRAFIC	S/T	https://www.alexa.com/siteinfo
Bet365.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Beyond.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
bilibili.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Bing.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Bit.ly	TRAFIC	Tool	https://www.alexa.com/siteinfo
Bitbucket.org	TRAFIC	App	https://www.alexa.com/siteinfo
Bitcoin.com	TRAFIC	Econ	https://www.alexa.com/siteinfo
Bitshare.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Bl.uk	TRAFIC	Book	https://www.alexa.com/siteinfo
Blackle.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Blog.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Blogadda.com/	TRAFIC	Blog	https://www.alexa.com/siteinfo
Blogcatalog.com/	TRAFIC	Blog	https://www.alexa.com/siteinfo
Blogueur.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Blogspot.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Bluehost.com	TRAFIC	Host	https://www.alexa.com/siteinfo
Blurtit.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Bnf.fr	TRAFIC	Book	https://www.alexa.com/siteinfo
Bongacams.com	TRAFIC	Porn	https://www.alexa.com/siteinfo
booking.com	TRAFIC	Tur	https://www.alexa.com/siteinfo
Livres.google.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Box.com	TRAFIC	App	https://www.alexa.com/siteinfo
Supports.io	TRAFIC	App	https://www.alexa.com/siteinfo
Entreprise.com	TRAFIC	DIR	https://www.alexa.com/siteinfo
Busuu.com	TRAFIC	EDU	https://www.alexa.com/siteinfo
C9.io	TRAFIC	Cloud	https://www.alexa.com/siteinfo
Cafemom.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Cairn.info	TRAFIC	S/T	https://www.alexa.com/siteinfo
Canva.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Care2.com	TRAFIC	Advo	https://www.alexa.com/siteinfo
Caringbridge.org	TRAFIC	Health	https://www.alexa.com/siteinfo
Chacha.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Chaturbate.com	TRAFIC	Porn	https://www.alexa.com/siteinfo
Chrome.com	TRAFIC	ICT	https://www.alexa.com/siteinfo
Classmates.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Codeanywhere.com	TRAFIC	Cloud	https://www.alexa.com/siteinfo
Codepen.io	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Commonsensemedia.org	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Contentful.com	TRAFIC	APP	https://www.alexa.com/siteinfo
Couchsurfing.com	TRAFIC	Tur	https://www.alexa.com/siteinfo
Coursera	TRAFIC	MOOC	https://www.alexa.com/siteinfo
Creativecommons.org	TRAFIC	SEng	https://www.alexa.com/siteinfo
Crunchyroll.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Csdn.net	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Cyworld.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Dailymotion.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Dart-europe.eu	TRAFIC	S/T	https://www.alexa.com/siteinfo
Daum.net	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Deezer.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Délicieux	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Dépôtfiles.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Deviantart.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Discordapp.com	TRAFIC	App	https://www.alexa.com/siteinfo
disneyplus.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Dmoz.org	TRAFIC	DIR	https://www.alexa.com/siteinfo

Doaj.org	TRAFIC	DIR	https://www.alexa.com/siteinfo
Douban.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
doubleclick.net	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Draugiem.lv	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Dreamhost.com	TRAFIC	Host	https://www.alexa.com/siteinfo
Dreamwidth.org	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Dropbox.com	TRAFIC	App	https://www.alexa.com/siteinfo
Drupal.org	TRAFIC	CMS	https://www.alexa.com/siteinfo
Duckduckgo.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
DXY.cn	TRAFIC	Health	https://www.alexa.com/siteinfo
ebay.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Eclipse.org	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Edx.org	TRAFIC	MOOC	https://www.alexa.com/siteinfo
egnyte.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Eharmony.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Etoro.com	TRAFIC	Econ	https://www.alexa.com/siteinfo
Etsy.com	TRAFIC	Econ	https://www.alexa.com/siteinfo
Europeana.eu	TRAFIC	Book	https://www.alexa.com/siteinfo
Exalead.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Excitez.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Experienceproject.com	TRAFIC	Dead	https://www.alexa.com/siteinfo
Fandom.com	TRAFIC	VC	https://www.alexa.com/siteinfo
Fetlife.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Filefactory.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Fileserve.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Filmaffinity.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Filmow.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Flickr.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Flipboard.fr	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Flixster.com	TRAFIC	Film	https://www.alexa.com/siteinfo
FNAC.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Force.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Fotki.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Fotolog.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Foursquare.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Fun-mooc.fr	TRAFIC	MOOC	https://www.alexa.com/siteinfo
Funnyordie.com	TRAFIC	Hum	https://www.alexa.com/siteinfo
Futurelearn.com	TRAFIC	MOOC	https://www.alexa.com/siteinfo
G2a.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Gaiaonline.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Gameblog.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Gamefaqs.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Geni.com	TRAFIC	Gen	https://www.alexa.com/siteinfo
Gfycat.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Ghost.org	TRAFIC	Blog	https://www.alexa.com/siteinfo
Gigablast.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Gigasize.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Girlsaskguys.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Github.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Gmx.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Gmx.net	TRAFIC	Mail	https://www.alexa.com/siteinfo
Godaddy.com	TRAFIC	Host	https://www.alexa.com/siteinfo
GOG.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Goodreads.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Google.fr	TRAFIC	SEng	https://www.alexa.com/siteinfo
Gotinder.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Gravatar.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Grindr.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Gutenberg.org	TRAFIC	Book	https://www.alexa.com/siteinfo
Haosou.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Hathitrust.org	TRAFIC	Book	https://www.alexa.com/siteinfo
Salut5.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Hightail.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Hostgator.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Hotmail.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Huanqiu.com	TRAFIC	News	https://www.alexa.com/siteinfo
Hubpages.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Hulu.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Hushmail.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Ibiblio.org	TRAFIC	Book	https://www.alexa.com/siteinfo
Icloud.com	TRAFIC	Mail	https://www.alexa.com/siteinfo

Icq.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
imdb.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Imgur.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Indiblogger.in	TRAFIC	Blog	https://www.alexa.com/siteinfo
Influenster.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
HébergementInmotion.com	TRAFIC	Host	https://www.alexa.com/siteinfo
Instagram.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Iqiyi.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Isbn.org	TRAFIC	Book	https://www.alexa.com/siteinfo
Italki.com	TRAFIC	EDU	https://www.alexa.com/siteinfo
Itch.io	TRAFIC	Gam	https://www.alexa.com/siteinfo
Jasminedirectory.com	TRAFIC	DIR	https://www.alexa.com/siteinfo
jd.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Jekyllrb.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Jetbrains.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
joinclubhouse.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Joomla.com	TRAFIC	CMS	https://www.alexa.com/siteinfo
Journalseek.net	TRAFIC	S/T	https://www.alexa.com/siteinfo
Jstor.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Jurn.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Justanswer.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Kaixin001.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Kakao.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Kompas.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Kongregate.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Last.fm	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Bibliothèque.harvard.edu	TRAFIC	Book	https://www.alexa.com/siteinfo
Librarything.com	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Line.me	TRAFIC	MSG	https://www.alexa.com/siteinfo
Linkedin.com	TRAFIC	SN-pr	https://www.alexa.com/siteinfo
Linux.org	TRAFIC	ICT	https://www.alexa.com/siteinfo
Liquidweb.com	TRAFIC	Host	https://www.alexa.com/siteinfo
Live.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Livejasmin.com	TRAFIC	Porn	https://www.alexa.com/siteinfo
Livejournal.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Liveleak.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Logoslibrary.eu	TRAFIC	Book	https://www.alexa.com/siteinfo
Lycos.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Mail.aol.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Mail.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Mail.google.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Mail.ru	TRAFIC	Mail	https://www.alexa.com/siteinfo
Mail.yandex.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Mamba.ru	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Match.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Mediafire.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Medium.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Meetic.fr	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Meetup.com	TRAFIC	SN-pr	https://www.alexa.com/siteinfo
Mega.io	TRAFIC	Cloud	https://www.alexa.com/siteinfo
Mendeley.com	TRAFIC	S/T	https://www.alexa.com/siteinfo
Messenger.yahoo.com/	TRAFIC	MSG	https://www.alexa.com/siteinfo
Metacafe.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Metafilter.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Microsoft.com	TRAFIC	ICT	https://www.alexa.com/siteinfo
Métropoles.com	TRAFIC	News	https://www.alexa.com/siteinfo
Microsoftonline.com	TRAFIC	ICT	https://www.alexa.com/siteinfo
Miniclip.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Mixi.jp	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Mocospace.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Moodle.org	TRAFIC	CMS	https://www.alexa.com/siteinfo
Mouthshut.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Mozilla.org	TRAFIC	ICT	https://www.alexa.com/siteinfo
Msn.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Mubi.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
monheritage.com	TRAFIC	Gen	https://www.alexa.com/siteinfo
Mavie.com	TRAFIC	Dead	https://www.alexa.com/siteinfo
Myshopify.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Myspace.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Napster.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Naver.com	TRAFIC	SEng	https://www.alexa.com/siteinfo

Netcraft.com	TRAFIC	Secu	https://www.alexa.com/siteinfo
Netflix.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Newgrounds.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Nicovideo.jp	TRAFIC	Vid	https://www.alexa.com/siteinfo
Ning.com	TRAFIC	SN-pr	https://www.alexa.com/siteinfo
Bloc-notes-plus-plus.org	TRAFIC	Tool	https://www.alexa.com/siteinfo
Novoed.com	TRAFIC	MOOC	https://www.alexa.com/siteinfo
Oatd.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Odnoklassniki.ru	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Office.com	TRAFIC	ICT	https://www.alexa.com/siteinfo
Ok.ru	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Okcupid.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Okezone.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Oovoo.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Openclassrooms.com	TRAFIC	MOOC	https://www.alexa.com/siteinfo
Opengrey.eu	TRAFIC	S/T	https://www.alexa.com/siteinfo
Openlibrary.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Openoffice.org	TRAFIC	ICT	https://www.alexa.com/siteinfo
Openthesis.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Opera.com	TRAFIC	ICT	https://www.alexa.com/siteinfo
Origine.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Outlook.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Panda.tv	TRAFIC	Vid	https://www.alexa.com/siteinfo
Paypal.com	TRAFIC	Econ	https://www.alexa.com/siteinfo
Pen.io	TRAFIC	Blog	https://www.alexa.com/siteinfo
Periscope.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Periscope.tv	TRAFIC	Vid	https://www.alexa.com/siteinfo
Photobucket.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Pikiran-rakyat.com	TRAFIC	News	https://www.alexa.com/siteinfo
Pinterest.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Playstation.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Playstore.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Plurk.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Pornhub.com	TRAFIC	Porn	https://www.alexa.com/siteinfo
Primevideo.com	TRAFIC	Film	https://www.alexa.com/siteinfo
Protonmail.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Qq.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Question.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Quora.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Qwant.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Rapidshare.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Ravelry.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Reddit.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Rediff.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Rediffmail.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Renren.com	TRAFIC	SN-Fr	https://www.alexa.com/siteinfo
Researchgate.net	TRAFIC	S/T	https://www.alexa.com/siteinfo
Réverbnation.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Roblox.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Rumble.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Rutube.ru	TRAFIC	Vid	https://www.alexa.com/siteinfo
Salesforce.com	TRAFIC	App	https://www.alexa.com/siteinfo
Sapo.pt	TRAFIC	SEng	https://www.alexa.com/siteinfo
Enregistrerde.net	TRAFIC	Tool	https://www.alexa.com/siteinfo
Scielo.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Scienceopen.com	TRAFIC	S/T	https://www.alexa.com/siteinfo
Search.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Secondlife.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Semanticscholar.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Sharecare.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Similarweb.com	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Sina.com.cn	TRAFIC	Port	https://www.alexa.com/siteinfo
Sitebuilder.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Skype.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Skyrock.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Slack.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Slideshare.net	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Smugmug.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Snapchat.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
so.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Socolar.com	TRAFIC	S/T	https://www.alexa.com/siteinfo

Sogou.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
sohu.com	TRAFIC	Port	https://www.alexa.com/siteinfo
Somech.com	TRAFIC	DIR	https://www.alexa.com/siteinfo
Sony.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Soso.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Soundcloud.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Espaces.ru	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Spip.net	TRAFIC	CMS	https://www.alexa.com/siteinfo
Spotify.com	TRAFIC	SN-mu	https://www.alexa.com/siteinfo
Squarespace.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Stackexchange.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Stackoverflow.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Startpage.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Steam.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Steampowered.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Straightdope.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Stumbleupon.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Sublimetext.com	TRAFIC	App	https://www.alexa.com/siteinfo
Svbtle.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Tagged.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Taobao.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Taringa.net	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Teamspeak.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Teamviewer.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Technorati.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Télégramme - interface	TRAFIC	MSG	https://www.alexa.com/siteinfo
Telegram.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Telegram.org	TRAFIC	MSG	https://www.alexa.com/siteinfo
Theblogchatter.com/	TRAFIC	Blog	https://www.alexa.com/siteinfo
Thèses.fr	TRAFIC	S/T	https://www.alexa.com/siteinfo
Tianya.cn	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Tiktok.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Tinyurl.com	TRAFIC	Tool	https://www.alexa.com/siteinfo
Tmall.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Trombi.com	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Tudou.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Tuenti.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Tumblr.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Twitch.tv	TRAFIC	Gam	https://www.alexa.com/siteinfo
Twoo.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
Typepad.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Udacity.com	TRAFIC	MOOC	https://www.alexa.com/siteinfo
Udemy.com	TRAFIC	MOOC	https://www.alexa.com/siteinfo
Uploaded.net	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Uploading.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Veoh.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Viadeo.com	TRAFIC	SN-pr	https://www.alexa.com/siteinfo
Viber.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Vimeo.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Vk.com	TRAFIC	SN-Mu	https://www.alexa.com/siteinfo
Wattpad.com	TRAFIC	SN-fr	https://www.alexa.com/siteinfo
Wayn.com	TRAFIC	Tur	https://www.alexa.com/siteinfo
Wdl.org	TRAFIC	Book	https://www.alexa.com/siteinfo
Webcrawler.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Webometrics.info	TRAFIC	Mktg	https://www.alexa.com/siteinfo
Wechat.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Weebly.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Weheartit.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Weibo.com	TRAFIC	Blog	https://www.alexa.com/siteinfo
Wetransfer.com	TRAFIC	FiSh	https://www.alexa.com/siteinfo
Whatsapp.com	TRAFIC	MSG	https://www.alexa.com/siteinfo
Wistia.com	TRAFIC	SN-Im	https://www.alexa.com/siteinfo
Wix.com	TRAFIC	App	https://www.alexa.com/siteinfo
Wolframalpha.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Wordpress.com	TRAFIC	CMS	https://www.alexa.com/siteinfo
Worldcat.com	TRAFIC	Book	https://www.alexa.com/siteinfo
Worldwidescience.org	TRAFIC	S/T	https://www.alexa.com/siteinfo
Xbox.com	TRAFIC	Gam	https://www.alexa.com/siteinfo
Xhamster.com	TRAFIC	Porn	https://www.alexa.com/siteinfo
Xing.com	TRAFIC	SN-pr	https://www.alexa.com/siteinfo
Xinhuanet.com	TRAFIC	News	https://www.alexa.com/siteinfo

Xvideos.com	TRAFIC	Porn	https://www.alexa.com/siteinfo
yahoo.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Yammer.com	TRAFIC	SN-pr	https://www.alexa.com/siteinfo
Yandex.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Yelp.com	TRAFIC	SEng	https://www.alexa.com/siteinfo
Youku.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Youtube	TRAFIC	Vid	https://www.alexa.com/siteinfo
Yy.com	TRAFIC	Vid	https://www.alexa.com/siteinfo
Zhanqi.tv	TRAFIC	Vid	https://www.alexa.com/siteinfo
Zhihu.com	TRAFIC	Q/A	https://www.alexa.com/siteinfo
Zillow.com	TRAFIC	e.com	https://www.alexa.com/siteinfo
Zoho.com	TRAFIC	Mail	https://www.alexa.com/siteinfo
Zoom.us	TRAFIC	MSG	https://www.alexa.com/siteinfo
Zoosk.com	TRAFIC	SN-Da	https://www.alexa.com/siteinfo
FACEBOOK %utilisateurs par pays (NapoleonCat 2021)	USAGES		https://napoleoncat.com/stats/
INSTAGRAM %utilisateurs par pays (NapoleonCat 2021)	USAGES		https://napoleoncat.com/stats/
MESSENGER %utilisateurs par pays (NapoleonCat 2021)	USAGES		https://napoleoncat.com/stats/
LINKEDIN %utilisateurs par pays (NapoleonCat 2021)	USAGES		https://napoleoncat.com/stats/
Linkedin %utilisateur par pays (ApolloTech 2021)	USAGES		https://www.apollotechnical.com/linkedin-users-by-country/
Twitter %utilisateurs par pays (Statista 2021)	USAGES		https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
FACEBOOK % utilisateurs à partir de l'IWS 2021	USAGES		https://www.internetworldstats.com/stats1.htm + stats2.htm+…stats6.htm
% d'audience Facebook (Statista 2021)	USAGES		https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/
YouTube % de personnes connectées dans le pays (Statista 2021)	USAGES		https://www.statista.com/statistics/1219589/youtube-penetration-worldwide-by-country/
Netflix % d'abonnés par pays (CompariTech 2020)	USAGES		https://www.comparitech.com/tv-streaming/netflix-subscribers/
% d'audience Pinterest (Statista 2021)	USAGES		https://www.statista.com/statistics/328106/pinterest-penetration-markets/
REDDIT % d'utilisateurs par pays (Statista 2021)	USAGES		https://backlinko.com/reddit-users
2012/21 % de téléchargements OpenOffice cumulés par pays	USAGES		http://www.openoffice.org/stats/countries.html
# Serveurs Internet sécurisés	USAGES		https://data.worldbank.org/indicator/IT.NET.SECR
% Abonnement haut débit fixe dans le pays (BM 2021)	USAGES		https://data.worldbank.org/indicator/IT.NET.BBND.P2
% Abonnement Tél. Fixe+ Mobile dans le pays (BM 2021)	USAGES		https://data.worldbank.org/indicator/IT.MLT.MAIN.P2 + https://data.worldbank.org/indicator/IT.CEL.SETS.P2

TYPOLOGIE	QTÉ	THÈME
?	1	
Advo	1	Plaidoyer
App	10	Applications
Blog	20	
Book	18	Livres
Cloud	3	Nuage (informatique)
CMS	5	Système de gestion de contenu
DIR	7	Annuaire
e.com	9	Commerce électronique
Econ	5	Économie
EDU	2	Cours
FiSh	11	Partage de fichiers
Film	8	Films à la demande
Gam	20	Jeux

Gen	2	Généalogie
Health	2	Santé
Host	7	Hébergement Web
Hum	1	Humour
ICT	13	TIC
Mail	17	Courrier électronique
Mktg	10	Commercialisation
MOOC	8	CLOM
MSG	23	Messagerie
News	4	Journaux
Porn	6	Pornographie
Port	8	Portail
Q/A	13	Question Réponse
S/T	22	Science et technologie (recherche)
Secu	2	Sécurité
SEng	26	Moteur de recherche
SN-Da	20	Rencontres Réseaux Sociaux
SN-Fr	28	Réseaux sociaux d'amitié
SN-Im	24	Images Réseaux Sociaux
SN-Mu	10	Réseaux sociaux de musique
SN-pr	6	Réseaux sociaux professionnels
Tool	14	Outils informatiques
Tur	3	Tourisme
VC	1	Communauté virtuelle
Vid	13	Vidéo

ANNEXE 2 : MACROLANGUES

CODE ISO	MACRO LANGUES	NOMBRE DE LANGUES FUSIONNÉES
ara	***Arabe***	***29***
aym	***Aymara***	***2***
aze	***Azerbaïdjanais***	***3***
bal	***Baloutchi***	***3***
bik	***Bikol***	***8***
bnc	***Bontok***	***5***
bua	***Bouriate***	***3***
chm	***Mari***	***2***
cre	***Cri***	***6***
del	***Delaware***	***2***
den	***Boniche***	***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	***Kongo***	***3***
kpe	**Kpelle**	**2**
kur	***Kurde***	***3***
lah	***Lahnda***	***7***
lav	***Letton***	***2***
luy	***Luyia***	***14***
man	***Mandingue***	***6***
mlg	***Malgache***	***11***
mon	***Mongol***	***3***
msa	***Malais***	***36***
mwr	***Marwari***	***6***
nep	***Népalais***	***2***
oji	***Ojibwé***	***7***
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	***Tamasheq***	***4***
uzb	***Ouzbek***	***2***
yid	***Yiddish***	***2***
zap	***Zapotèque***	***57***
zha	***Zhuang***	***16***
zho	***Chinois***	***15***
zza	***Dimli***	***2***

ANNEXE 3 : LISTE DES PAYS OU TERRITOIRES OU L'UIT NE PROPOSE PAS DE DONNÉES

Code ISO	NOM DU PAYS	POPULATION
AX	Île Åland	27 652
AS	Samoa américaines	55 990
IO	Territoire britannique de l'océan Indien	4 000
BQ	Pays-Bas caribéens	18 740
CX	L'île Christmas	1 170
CC	Îles Cocos (Keeling)	630
CK	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
MQ	Martinique	377 100
NF	Île de Norfolk	1 500
KP	*Corée du Nord*	*25 579 000*
MP	Îles Mariannes du Nord	53 280
PW	Palaos	17 550
PN	Pitcairn	36
RE	Réunion	751 580
BL	Saint-Barthélemy	7 850
MF	Saint Martin	28 500
PM	Saint-Pierre-et-Miquelon	6 340
SX	Saint-Martin	33 470
TC	Îles Turques-et-Caïques	30 170
VA	*État du Vatican*	*330*
EH	*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 incluses dans un pays donné
2) Il n'y a pas de source ni d'estimation du pourcentage de personnes connectées à l'Internet (en italique dans le tableau).

ANNEXE 4 : RÉSULTATS POUR TOUTES LES LANGUES

Rang			.Connect.M.	W.Pop.	TRAFIC	Connec.L.	USAGE	CONT.	INTER.	INDEX	PUISS.	CAPAC.	GRAD..
	ISO	Total ou moyenne ---->	100%	100%	100%	54,70%	100%	100%	100%	100%	100%	0,75	0,74
		Reste	10,13 %	12,66 %	7,90%	43,76 %	8,59 %	2,88%	0,02%	6,91 %	6,07 %	0,48	0,60
54	afr	Afrikaans	0,19%	0,17%	0,08 %	59,75%	0,11%	0,15%	0,10%	0,17%	0,13%	0,79	0,73
102	aka	Akan	0,06 %	0,09 %	0,02%	38,80 %	0,05%	0,00%	0,01%	0,05%	0,03%	0,35	0,49
60	amh	Amharique	0,21%	0,55%	0,09 %	20,57 %	0,11%	0,01%	0,12%	0,11%	0,11%	0,19	0,51
8	ara	*Arabe*	3,89%	3,53%	2,30 %	60,14%	3,02%	2,05 %	4,29 %	3,01%	3,09 %	0,88	0,80
74	asm	Assamais	0,11%	0,15%	0,12%	40,03 %	0,08 %	0,00%	0,03%	0,09 %	0,07 %	0,49	0,66
119	awa	Awadhi	0,03%	0,04 %	0,03%	39,25%	0,02%	0,00%	0,00%	0,03%	0,02%	0,43	0,60
42	aze	*Azerbaïdjanais*	0,31%	0,23%	0,26%	74,76%	0,16%	0,11%	0,17%	0,27%	0,22%	0,94	0,69
106	bal	*Baloutchi*	0,05%	0,09 %	0,06 %	30,72 %	0,04 %	0,00%	0,00%	0,03%	0,03%	0,36	0,63
127	bam	Bamanankan	0,03%	0,14%	0,01%	12,94%	0,02%	0,00%	0,00%	0,01%	0,01%	0,10	0,42
53	bar	Bavarois	0,22%	0,14%	0,10%	87,68 %	0,17%	0,00%	0,00%	0,33%	0,14%	0,97	0,61
94	bel	Biélorusse	0,06 %	0,04 %	0,02%	82,27%	0,03%	0,03%	0,03%	0,06 %	0,04 %	1,00	0,66
15	ben	Bengali	1,14 %	2,58%	1,22%	24,15%	1,13 %	0,26%	0,72%	0,84%	0,88%	0,34	0,78
112	bew	Betawi	0,04 %	0,05%	0,01%	47,69 %	0,05%	0,00%	0,00%	0,04 %	0,02%	0,50	0,57
34	bho	Bhojpuri	0,37%	0,51%	0,40%	39,85%	0,27%	0,00%	0,03%	0,32%	0,23%	0,46	0,63
118	bik	*Bikol*	0,03%	0,04 %	0,01%	43,03 %	0,04 %	0,00%	0,00%	0,03%	0,02%	0,51	0,65
109	bjj	*Kanauji*	0,04 %	0,06 %	0,05%	40.00%	0,03%	0,00%	0,00%	0,04 %	0,03%	0,45	0,62
116	bug	*Bugis*	0,04 %	0,04 %	0,01%	47,94 %	0,04 %	0,00%	0,00%	0,03%	0,02%	0,50	0,57
63	bul	*Bulgare*	0,10%	0,08 %	0,05%	70,34%	0,08 %	0,13%	0,08 %	0,12%	0,09 %	1.18	0,92
69	ceb	Cebuano	0,12%	0,15%	0,06 %	43,15%	0,19%	0,00%	0,02%	0,11%	0,08 %	0,54	0,69
38	ces	Tchèque	0,19%	0,13%	0,07 %	81,37 %	0,13%	0,50%	0,18%	0,25%	0,22%	1,70	1.14
55	dan	Danois	0,10%	0,05%	0,04 %	97,82 %	0,08 %	0,26%	0,08 %	0,16%	0,12%	2.19	1.22
9	deu	Allemand	2,09 %	1,30%	1,32%	87,65%	1,95 %	5,84 %	2,97%	2,98%	2,86%	2.19	1,37
123	doi	*Dogri*	0,03%	0,04 %	0,03%	40.00%	0,02%	0,00%	0,00%	0,02%	0,02%	0,46	0,63
107	dyu	Jula	0,07 %	0,12%	0,02%	30,85%	0,04 %	0,00%	0,00%	0,04 %	0,03%	0,24	0,43
37	ell	Grec	0,18%	0,13%	0,21%	77,71 %	0,17%	0,37%	0,19%	0,24%	0,22%	1,75	1.23
1	eng	Anglais	15,30 %	13.01 %	37,4 %	64,33 %	27,9%	38,61%	21,73 %	17,87 %	26,48 %	2.04	1,73
125	ewe	Éwé	0,03%	0,05%	0,01%	31,78 %	0,02%	0,00%	0,00%	0,02%	0,01%	0,26	0,45
19	fas	Persan	0,95%	0,81%	0,55%	64,58%	0,39%	0,74%	0,75%	0,81%	0,70%	0,87	0,73
44	fin	Finlandais	0,09 %	0,06 %	0,04 %	89,67%	0,06 %	0,74%	0,08 %	0,14%	0,19%	3.42	2.09
4	fra	Français	3,00%	2,58%	2,64 %	63,67 %	3,75%	5,40 %	4,26 %	3.21%	3,71%	1,44	1.24
70	ful	*Fulfulde*	0,19%	0,31%	0,07 %	33,16 %	0,09 %	0,00%	0,00%	0,12%	0,08 %	0,25	0,42
89	grn	*Guarani*	0,08 %	0,06 %	0,03%	68,83 %	0,06 %	0,00%	0,01%	0,07 %	0,04 %	0,64	0,51
73	gsw	Suisse allemand	0,10%	0,06 %	0,08 %	91,56 %	0,09 %	0,00%	0,01%	0,17%	0,08 %	1.21	0,72
28	guj	Gujarati	0,44%	0,60%	0,53%	40,49 %	0,35%	0,05%	0,24%	0,39%	0,34%	0,56	0,76
91	hat	Créole haïtien	0,05%	0,08 %	0,06 %	38,59 %	0,06 %	0,00%	0,03%	0,03%	0,04 %	0,50	0,70
45	hau	Haoussa	0,43%	0,72%	0,16%	32,61 %	0,16%	0,00%	0,10%	0,28%	0,19%	0,26	0,44
20	hbs	*Serbo-croate*	0,27%	0,19%	0,14%	77,78 %	0,21%	2,49%	0,22%	0,31%	0,61%	3.14	2.21
26	heb	Hébreu	0,14%	0,09 %	0,08 %	85,46%	0,11%	2,20 %	0,13%	0,19%	0,47%	5.24	3,35
103	hil	Hiligaynon	0,05%	0,06 %	0,02%	43,08 %	0,07 %	0,00%	0,00%	0,04 %	0,03%	0,51	0,65
5	hin	Hindi	4,26 %	5,80%	4,81%	40,18%	3,16 %	0,28%	4,03%	3,71%	3,38 %	0,58	0,79
82	hmn	*Hmong*	0,09 %	0,07 %	0,06 %	64,80%	0,05%	0,00%	0,03%	0,09 %	0,05%	0,72	0,61
75	hne	Chhattisgarhi	0,12%	0,16%	0,13%	40.00%	0,08 %	0,00%	0,00%	0,10%	0,07 %	0,45	0,62
41	hun	Hongrois	0,18%	0,12%	0,08 %	79,92 %	0,15%	0,57%	0,13%	0,20%	0,22%	1,79	1.22

83	hye	**Arménien**	0,05%	0,04 %	0,02%	69,86%	0,03%	0,14%	0,02%	0,05%	0,05%	1,41	1.11
101	ibb	**Ibibio**	0,08 %	0,10%	0,03%	41,98 %	0,03%	0,00%	0,00%	0,06 %	0,03%	0,31	0,41
62	ibo	**Ibo**	0,22%	0,28%	0,08 %	42,02 %	0,08 %	0,00%	0,05%	0,16%	0,10%	0,35	0,45
97	ilo	**Ilocano**	0,05%	0,06 %	0,03%	43,82 %	0,08 %	0,00%	0,00%	0,05%	0,03%	0,56	0,69
12	ita	**Italien**	0,91%	0,66%	0,51%	75,65 %	0,97%	3,39 %	1,22%	1,20%	1,37%	2.09	1,51
27	jav	**Javanais**	0,58%	0,66%	0,20%	47,74 %	0,69%	0,00%	0,14%	0,51%	0,35%	0,53	0,61
dix	jpn	**Japonais**	2,07 %	1,22%	1,98 %	92,62 %	1,76%	3,55%	2,77%	3,01%	2,52%	2.07	1.22
93	kab	**Amazigh**	0,07 %	0,07 %	0,04 %	62,12 %	0,05%	0,00%	0,00%	0,06 %	0,04 %	0,58	0,51
30	kan	**Kannada**	0,42%	0,57%	0,47%	40,12 %	0,31%	0,08 %	0,23%	0,36%	0,31%	0,55	0,75
104	kas	**Cachemire**	0,05%	0,07 %	0,06 %	38,84 %	0,04 %	0,00%	0,00%	0,04 %	0,03%	0,45	0,63
110	kau	***Kanuri***	0,06 %	0,09 %	0,02%	39,21%	0,02%	0,00%	0,00%	0,04 %	0,02%	0,29	0,40
56	kaz	**Kazakh**	0,18%	0,13%	0,07 %	76,98 %	0,10%	0,07 %	0,10%	0,17%	0,11%	0,90	0,64
64	khm	**Khmer**	0,14%	0,17%	0,07 %	43,40%	0,16%	0,02%	0,08 %	0,09 %	0,09 %	0,53	0,66
121	kik	**Gikuyu**	0,03%	0,08 %	0,01%	22,57%	0,03%	0,00%	0,01%	0,03%	0,02%	0,22	0,53
111	kin	**Kinyarwanda**	0,06 %	0,13%	0,02%	24,69%	0,02%	0,00%	0,01%	0,04 %	0,02%	0,19	0,42
132	kln	***Kalenjin***	0,02%	0,04 %	0,01%	22,62 %	0,02%	0,00%	0,00%	0,01%	0,01%	0,21	0,50
137	kmb	**Kimbundu**	0,00%	0,02%	0,00%	16,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,14	0,48
108	kok	***Konkani***	0,04 %	0,06 %	0,05%	39,76 %	0,03%	0,00%	0,00%	0,04 %	0,03%	0,46	0,63
130	kon	***Kongo***	0,02%	0,12%	0,01%	11,62 %	0,02%	0,00%	0,00%	0,01%	0,01%	0,09	0,44
14	kor	**Coréen**	0,93%	0,79%	0,93%	64,73 %	0,99%	0,85%	1,10 %	0,95%	0,96%	1.22	1.03
136	ktu	**Kituba**	0,01%	0,05%	0,00%	10.00%	0,01%	0,00%	0,00%	0,00%	0,00%	0,07	0,39
40	kur	***Kurde***	0,32%	0,24%	0,20%	73,02 %	0,28%	0,04 %	0,15%	0,29%	0,22%	0,89	0,67
39	lah	***Lahnda***	0,31%	0,96%	0,41%	17,43 %	0,26%	0,01%	0,15%	0,18%	0,22%	0,23	0,71
134	lua	**Luba-kasaï**	0,01%	0,07 %	0,00%	10,05%	0,01%	0,00%	0,00%	0,00%	0,01%	0,07	0,40
117	lug	**Ganda**	0,05%	0,11%	0,01%	25,01%	0,02%	0,00%	0,00%	0,03%	0,02%	0,18	0,39
133	luy	***Luyia***	0,01%	0,03%	0,00%	22,98%	0,01%	0,00%	0,00%	0,01%	0,01%	0,20	0,48
95	mad	**Madura**	0,07 %	0,08 %	0,02%	47,70%	0,08 %	0,00%	0,00%	0,06 %	0,04 %	0,50	0,57
65	mag	**Magahi**	0,15%	0,20%	0,16%	39,99%	0,11%	0,00%	0,00%	0,13%	0,09 %	0,45	0,62
51	mai	**Maithili**	0,24%	0,33%	0,25%	39,28 %	0,18%	0,00%	0,02%	0,20%	0,15%	0,44	0,62
35	mal	**Malayalam**	0,28%	0,37%	0,35%	42,54 %	0,26%	0,04 %	0,18%	0,25%	0,23%	0,62	0,80
120	man	***Mandingue***	0,04 %	0,08 %	0,01%	26,96 %	0,03%	0,00%	0,00%	0,02%	0,02%	0,20	0,42
23	mar	**Marathi**	0,70%	0,96%	0,79%	40,06 %	0,52%	0,06 %	0,44%	0,61%	0,52%	0,54	0,74
99	mey	**Hassaniyya**	0,07 %	0,09 %	0,03%	43,68 %	0,05%	0,00%	0,00%	0,05%	0,03%	0,35	0,44
77	mlg	***Malgache***	0,03%	0,18%	0,01%	9,79%	0,03%	0,32%	0,01%	0,01%	0,07 %	0,40	2.21
92	mon	***Mongol***	0,06 %	0,06 %	0,03%	58,99%	0,04 %	0,01%	0,02%	0,06 %	0,04 %	0,65	0,61
126	mos	**Moré**	0,03%	0,08 %	0,01%	23,19 %	0,02%	0,00%	0,00%	0,02%	0,01%	0,18	0,42
11	msa	***Malais***	2,20 %	2,36 %	0,89%	51.00%	2,79%	0,79%	1,91%	1,99 %	1,76%	0,75	0,80
67	mwr	***Marwari***	0,14%	0,20%	0,16%	39,81%	0,11%	0,00%	0,00%	0,13%	0,09 %	0,45	0,62
52	mya	**Birman**	0,24%	0,41%	0,08 %	31,85%	0,25%	0,03%	0,11%	0,14%	0,14%	0,35	0,60
86	nap	**Napolitain-calab.**	0,07 %	0,06 %	0,03%	74,39 %	0,08 %	0,00%	0,00%	0,10%	0,05%	0,84	0,62
58	nep	***Népalais***	0,16%	0,25%	0,09 %	35,70 %	0,14%	0,03%	0,14%	0,11%	0,11%	0,45	0,69
22	nld	**Néerlandais**	0,40%	0,24%	0,19%	92,02 %	0,42%	1,13 %	0,47%	0,60%	0,53%	2.26	1,34
90	nod	**Thaïlandais nord**	0,07 %	0,06 %	0,03%	66,47%	0,08 %	0,00%	0,00%	0,07 %	0,04 %	0,70	0,57
122	nya	**Chichewa**	0,04 %	0,14%	0,01%	15,87%	0,02%	0,00%	0,01%	0,02%	0,02%	0,12	0,42
43	ori	***Oriya***	0.30%	0,41%	0,33%	39,96 %	0,22%	0,01%	0,14%	0,26%	0,21%	0,51	0,70
84	orm	***Oromo***	0,13%	0,36%	0,04 %	20,07%	0,06 %	0,00%	0,01%	0,07 %	0,05%	0,14	0,39
36	pan	**Pendjabi Est**	0,33%	0,50%	0,44%	35,80%	0.30%	0,00%	0,03%	0,27%	0,23%	0,45	0,69
17	pol	**Polonais**	0,58%	0,39%	0,31%	81,17%	0,53%	1,57%	0,69%	0,73%	0,74%	1,88	1,26

6	por	**Portugais**	3,05%	2,49%	1,42%	67,16 %	5,53%	3,30%	3,85%	2,92%	3,35 %	1,35	1.10
57	pus	***Pachtou***	0,16%	0,51%	0,20%	17,49 %	0,16%	0,00%	0,06 %	0,09 %	0,11%	0,22	0,69
85	que	***Quechua***	0,07 %	0,07 %	0,04 %	56,82 %	0,09 %	0,00%	0,01%	0,07 %	0,05%	0,66	0,64
78	raj	**Rajasthan**	0,11%	0,16%	0,13%	38,99%	0,08 %	0,00%	0,00%	0,10%	0,07 %	0,44	0,62
32	ron	**Roumain**	0,32%	0,23%	0,15%	75,66 %	0,26%	0,25%	0.30%	0,35%	0,27%	1.18	0,86
135	run	**Rundi**	0,01%	0,11%	0,00%	4,67%	0,01%	0,00%	0,00%	0,00%	0,00%	0,04	0,42
7	rus	**Russe**	3,51%	2,49%	1,81%	77,20 %	2,28 %	3,38 %	3,88%	3,78%	3,11%	1,25	0,88
100	sat	**Santhali**	0,05%	0,07 %	0,06 %	39,17%	0,04 %	0,00%	0,00%	0,05%	0,03%	0,44	0,62
68	sin	**Cinghalais**	0,12%	0,17%	0,06 %	39,46%	0,11%	0,09 %	0,05%	0,11%	0,09 %	0,53	0,73
66	slk	**Slovaque**	0,11%	0,07 %	0,04 %	82,47%	0,07 %	0,12%	0,08 %	0,13%	0,09 %	1.30	0,86
114	sna	**Shona**	0,05%	0,09 %	0,02%	30,31%	0,03%	0,00%	0,02%	0,03%	0,02%	0,26	0,46
72	snd	**Sindhi**	0,11%	0,32%	0,15%	18,73 %	0,10%	0,01%	0,03%	0,06 %	0,08 %	0,24	0,70
98	som	**Somali**	0,06 %	0,21%	0,04 %	15,24 %	0,06 %	0,00%	0,02%	0,03%	0,03%	0,16	0,57
79	sot	**Soto. Du sud**	0,13%	0,13%	0,06 %	56,47%	0,08 %	0,00%	0,01%	0,12%	0,07 %	0,51	0,49
105	sou	**Thaïlandais Sud**	0,05%	0,04 %	0,02%	66,68 %	0,06 %	0,00%	0,00%	0,05%	0,03%	0,70	0,57
3	spa	**Espagnol**	7,00 %	5,24 %	10,7 %	73,08 %	11,7%	5,42 %	9,94 %	7,59%	8,73 %	1,67	1,25
80	sqi	***Albanais***	0,08 %	0,06 %	0,05%	75,48 %	0,08 %	0,06 %	0,03%	0,08 %	0,06 %	1.12	0,81
124	suk	**Sukuma**	0,04 %	0,08 %	0,01%	25,00%	0,02%	0,00%	0,00%	0,02%	0,01%	0,18	0,40
47	sun	**Sonde**	0,27%	0,31%	0,09 %	47,69 %	0,33%	0,01%	0,06 %	0,24%	0,17%	0,54	0,62
46	swa	***Swahili***	0,32%	0,78%	0,12%	22,84%	0,21%	0,01%	0,20%	0,20%	0,18%	0,23	0,55
29	swe	**Suédois**	0,22%	0,13%	0,09 %	93,49 %	0,23%	0,87%	0,24%	0,34%	0,33%	2,61	1,53
25	tam	**Tamil**	0,62%	0,82%	0,71%	41,35%	0,51%	0,19%	0,39%	0,55%	0,50%	0,60	0,80
87	tat	**Tatar**	0,07 %	0,05%	0,03%	78,05 %	0,04 %	0,01%	0,03%	0,08 %	0,04 %	0,87	0,61
24	tel	**Telugu**	0,69%	0,92%	0,80%	40,71 %	0,53%	0,07 %	0,38%	0,60%	0,51%	0,55	0,74
113	tgk	**Tadjik**	0,05%	0,08 %	0,02%	32,22%	0,03%	0,00%	0,01%	0,03%	0,02%	0,29	0,49
33	tgl	**Tagalog**	0,24%	0,25%	0,33%	53,60 %	0,43%	0,06 %	0,15%	0,24%	0,24%	0,98	1,00
21	tha	**Thaïlandais**	0,72%	0,59%	0,29%	66,85 %	0,82%	0,33%	0,62%	0,67%	0,57%	0,98	0,80
129	tir	**Tigrigna**	0,03%	0,10%	0,01%	15,68 %	0,02%	0,00%	0,00%	0,01%	0,01%	0,12	0,41
76	tsn	**Setswana**	0,14%	0,13%	0,06 %	58,16 %	0,09 %	0,00%	0,01%	0,13%	0,07 %	0,53	0,50
96	tso	**Tsonga**	0,08 %	0,10%	0,03%	43,30%	0,04 %	0,00%	0,01%	0,06 %	0,04 %	0,38	0,48
61	tts	**Thaïlandais n.est**	0,18%	0,14%	0,07 %	66,65 %	0,20%	0,00%	0,00%	0,17%	0,10%	0,70	0,57
115	tuk	**Turkmène**	0,04 %	0,07 %	0,02%	31,48 %	0,02%	0,02%	0,01%	0,02%	0,02%	0,32	0,55
13	tur	**Turc**	1,21%	0,85%	1,03%	77,98 %	1,59 %	0,94%	1,43%	1,22%	1,24%	1,46	1.02
81	uig	**Ouïghour**	0,12%	0,10%	0,04 %	64,75%	0,03%	0,00%	0,03%	0,13%	0,06 %	0,58	0,49
31	ukr	**Ukrainien**	0,37%	0,32%	0,17%	63,96 %	0,25%	0,26%	0,33%	0,40%	0.30%	0,92	0,79
131	umb	**Umbundu**	0,02%	0,07 %	0,01%	16,00%	0,01%	0,00%	0,00%	0,01%	0,01%	0,14	0,48
18	urd	**Ourdou**	0,98%	2,22 %	1,33%	24,12 %	0,82%	0,03%	0,54%	0,65%	0,72%	0,33	0,74
49	uzb	***Ouzbek***	0,27%	0,32%	0,10%	45,90%	0,13%	0,06 %	0,13%	0,20%	0,15%	0,46	0,54
16	vie	**Vietnamien**	0,94%	0,74%	0,58%	69,04 %	1,15 %	0,46%	0,81%	0,83%	0,79%	1.07	0,85
128	vls	**Flamand occ.**	0,02%	0,01%	0,01%	90,43 %	0,02%	0,00%	0,00%	0,03%	0,01%	1.12	0,68
88	wol	**Wolof**	0,10%	0,12%	0,03%	46,09 %	0,05%	0,00%	0,00%	0,07 %	0,04 %	0,36	0,43
59	xho	**Xhosa**	0,20%	0,19%	0,09 %	59,96 %	0,12%	0,02%	0,05%	0,19%	0,11%	0,59	0,54
50	yor	**Yoruba**	0,32%	0,42%	0,11%	41,74 %	0,12%	0,00%	0,10%	0,23%	0,15%	0,36	0,47
71	zha	***Zhuang***	0,17%	0,14%	0,06 %	64,67%	0,04 %	0,01%	0,00%	0,18%	0,08 %	0,54	0,45
2	zho	***Chinois***	17,65 %	14,72%	7,79 %	65,59 %	5,47 %	8,18 %	25,07%	19,38%	13,92%	0,95	0,79
48	zul	**Zoulou**	0,29%	0,27%	0,13%	59,57 %	0,17%	0,03%	0,09 %	0,27%	0,16%	0,60	0,55