Internet et diversité linguistique : cyber-géographie des langues ayant le plus grand nombre de locuteurs.

Daniel Pimienta
Observatoire de la Diversité linguistique et culturelle dans l'Internet, Mai 2022

Traduit au français à partir de - "Internet and linguistic diversity: the cyber-geography of languages with the largest number of speakers" in *LinguaPax Review 2021, Language Technologies and Language Diversity*, page 9. - https://www.linguapax.org/wp-content/uploads/2022/02/LinguapaxReview9-2021-low.pdf

Note : L'ensemble du document est actualisé avec la dernière mesure (version 3 du modèle) à la date de mars 2022.

L'Observatoire de la diversité linguistique et culturelle dans l'Internet[1] vient de publier les résultats de sa dernière étude[2], mettant à jour et améliorant leurs travaux antérieurs de 2017 visant à produire des indicateurs de présence des langues de plus de 5 millions de locuteurs L1[3]. Nous proposons dans cet article d'analyser les données obtenues et d'en interpréter le sens, en termes de cyber-géographie des langues.

D'abord, une explication concernant les langues retenues : ce n'est pas une décision politique assumée de se focaliser sur les langues ayant le plus grand nombre de locuteurs et de laisser de côté les autres, et notamment les langues classées comme indigènes, à cette période que L'UNESCO a marqué, en 2019, l'année des langues indigènes (https://es.iyil2019.org/) puis la décennie des langues indigènes 2022-2032[4]. La sélection des langues avec le plus grand nombre de locuteurs est simplement une restriction résultant de la méthodologie adoptée par laquelle l'analyse des biais conduit à la conclusion que ceux-ci seraient trop élevés pour les langues avec un nombre de locuteurs inférieur à cinq millions.

Ceci nous amène, avant de discuter des résultats, à présenter brièvement la méthodologie, les principales sources et les biais pouvant affecter les données produites pour les langues considérées.

La source démolinguistique de cette nouvelle édition est le "Global Dataset #24" d'Ethnologue, de mars 2021, sans doute la source la plus complète en termes de langues, ainsi que la plus à jour et la plus fiable, bien qu'il doit être clair que la perfection n'existe pas dans ce domaine et que les professionnels du domaine peuvent questionner la validité de certains chiffres. Nous avons également adopté le regroupement de certaines langues en macro-langues[5]. Pour la version 1 de 2017, 130 langues avec L1>5 M ont été traitée et la liste peut être consultée dans Pimienta (2017). La version 2 en 2021 a permis d'étendre la couverture à L1 >1M, ce qui représente 329 langues[6], un chiffre plus proche du nombre estimé de 500 langues présentes dans l'Internet. La version 3, arrivée à son terme en mars 2022, a permis

[1]Nous préférons utiliser la forme d'expression *l'Internet* plutôt que la forme recommandée *Internet* car il nous semble que confondre le protocole de communication (*Internet*) avec le réseau mondial des personnes et de l'information (*l'Internet*) projette une histoire simplifiée de l'Internet et qui cache de nombreux apports précieux des réseaux qui existaient avant la convergence des protocoles (comme Bitnet ou Usenet, par exemple).
[2]Voir http://funedes.org/lc2021. Cette étude s'est concentrée en particulier sur le portugais et a été possible grâce au soutien du Département de la culture et de l'éducation du Ministère des affaires étrangères du Brésil, dans le cadre de l'Institut international de la langue portugaise et sous la coordination de la Chaire UNESCO sur les politiques linguistiques pour le multilinguisme. La dernière version (V3), avec le soutien de l'OIF, a permis à la méthode, en mars 2022, d'atteindre sa maturité et de contrôler tous les biais (voir http://funredes.org/lc2022).
[3]Nous utilisons la terminologie L1 pour désigner la langue maternelle et L2 pour la ou les deuxième(s) langue(s).
[4] https://en.unesco.org/news/unesco-launches-global-task-force-making-decade-action-indigenous-languages
[5]Les macro-langues sont indiquées en *italique*.
[6] http://funredes.org/lc2021/Results1M.xlsx

d'éliminer ou à réduire les biais restants et dans cette traduction les données sont actualisées avec cette dernière version[7].

La méthodologie détaillée, les sources et les biais associés, ainsi que les résultats des deux premières versions sont entièrement documentés dans Pimienta (2019 et 2021). Il s'agit d'une approche indirecte de mesure de la présence des langues dans l'Internet, basée sur un grand nombre de sources de données sur les langues ou les pays dans l'Internet. Les données par pays ont été transformées en données par langue, par une technique de pondération avec les données démolinguistiques, ceci étant une des plus grandes originalités de la méthode[8].

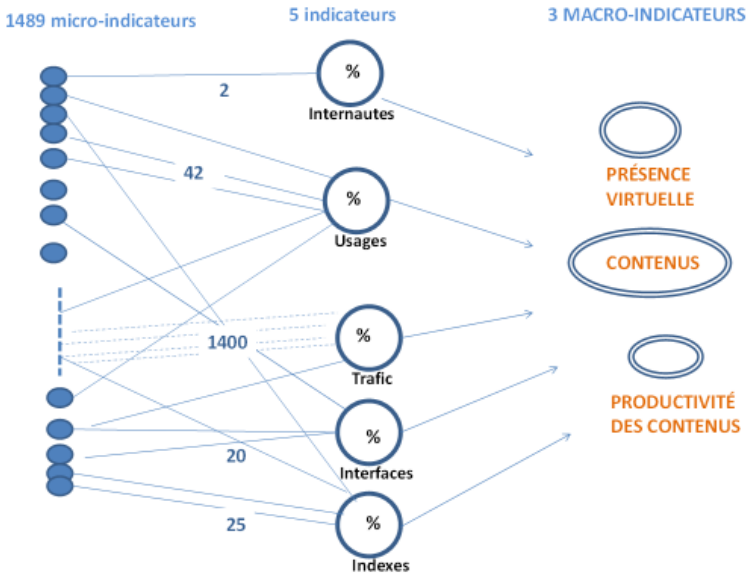
La présence est mesurée par des calculs statistiques effectués à partir de sources primaires, en termes de pourcentage global de la population L1+L2, selon 5 indicateurs[9]. A partir de ces 5 indicateurs, 3 macro-indicateurs sont calculés :

Contenus: la moyenne des 5 indicateurs (poids absolu de la langue dans l'Internet, ce qui favorise évidemment les langues avec le plus de locuteurs)
Présence virtuelle: pourcentage de contenus divisé par pourcentage de locuteurs (un poids relatif qui permet de mesurer la force des langues indépendamment de leur nombre de locuteurs)
Productivité des contenus: Pourcentages de contenus divisé par pourcentages de locuteurs connectés.

Toutes les données sont traitées en pourcentage du nombre mondial de locuteurs L1+L2, une valeur évidemment supérieure à la population mondiale[10]. Selon la dernière version d'Ethnologue, les données mondiales sont les suivantes :

- ✔ Population mondiale (total mondial des locuteurs de L1) : 7 231 699 136
- ✔ Total mondial des locuteurs L1 + L2 : 10 361 716 756
- ✔ Le "taux mondial de multilinguisme" est donc de 10 361 716 756 / 7 231 699 136 = 1,4328 (autrement dit, 43% de la population mondiale est au moins bilingue).

[7] Voir http://funredes.org/lc2022
[8]Crédits à Daniel Prado qui est à l'origine de cette idée en 2012.
[9]Les 5 indicateurs sont :
Internautes: pourcentage mondial de locuteurs connectés
Trafic: pourcentage de trafic dans la langue considérée
Usages: pourcentage de participation aux plateformes ou aux ressources de connectivité
Interfaces: présence de la langue dans les interfaces applicatives ou comme langue de traduction en ligne
Index: transformation en termes de langues, de divers classements des pays selon des paramètres liés à la société de l'information.
[10]Bien entendu, dans ce décompte, une même personne compte une fois pour sa langue maternelle et autant de fois pour chaque langue seconde.



Aucune méthode statistique n'est exempte de biais et il est très important d'identifier et d'analyser ces biais et leurs effets sur les résultats. L'indication de la répartition des locuteurs de L2 par pays, que propose désormais Ethnologue, a permis d'éliminer le biais le plus important de la méthode, qui consistait à extrapoler les résultats en termes de L1 à L2 (biais qui favorisait des langues comme l'anglais et français à forte population de L2 dans les pays à faible taux de connectivité). Il subsiste un biais notable dans la méthode, celui de considérer qu'au sein d'un pays le pourcentage de personnes connectées à l'Internet est identique pour toutes les langues présentes. Ainsi, le pourcentage de locuteurs du catalan connectés à l'Internet en Espagne est calculé de manière identique au pourcentage de locuteurs d'espagnol ou d'arabe, que ce soit en L1 ou en L2, bien que la réalité soit probablement différente, avec certaines langues avec des taux supérieurs à la moyenne nationale et d'autres avec des taux inférieurs. Ce biais a été jugé acceptable si le modèle n'a pas vocation à comparer les langues au sein d'un même pays et si les langues à faible nombre de locuteurs ne sont pas traitées ; dans tous les cas, il est important de le savoir lors de l'interprétation des résultats.

Les autres biais résultent de la sélection des sources pour les calculs des indicateurs et sont résumés dans le tableau suivant, en notant de 0 (biais inacceptable) à 20 (totalement exempt de biais) chaque indicateur et en montrant les évolutions entre 2017 et 2022.

INDICATEUR/ HYPOTHESE	ÉVAL. V1 2017	ÉVAL. V2 2021	ÉVAL. V3 2022	COMMENTAIRES
METHODE : L1	17	17	17	L'hypothèse selon laquelle tous les locuteurs se voient attribués le même taux de connexion national est valide à condition de ne pas comparer les langues à l'intérieur du même pays et à condition de ne pas traiter des langues avec faible nombre de locuteurs. Cette hypothèse sur laquelle repose l'ensemble des calculs comporte un biais qui défavorise les langues européennes dans les pays en développement (qui ont probablement un taux de connexion supérieur à la moyenne) et favorise les langues du Sud d'immigration dans les pays développés (qui ont probablement un taux inférieur à la moyenne nationale).
METHODE : L2	13	19	19	La méthode initiale par extrapolation proportionnelle des résultats L1 entrainait un biais assez fort en faveur des langues européennes

				dans les pays en développement (spécialement anglais et français).Ce biais a disparu à partir de V2 en utilisant les données fournies par Ethnologue pour L2 par pays. Reste les biais d'Ethnologue qu'il est impossible d'estimer et qui, comme ceux de l'UIT pour les taux de connexion, peuvent être considérés comme marginaux.
INTERNAUTES	19	16	19	La principale source est l'UIT. En 2017, il s'agissait de la source la mieux notée mais dans la version 2, 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. La Banque Mondiale couvre plus large dans sa dernière actualisation et semble assumer des mises à jour encore plus fréquente que l'UIT.
INDEX	15	18	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 ONG 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 à partir de V2.
CONTENUS	5	8		Il n'y avait que 13 micro-indicateurs pour construire cet indicateur et 11 d'entre eux provenaient de Wikimedia où la présence de langues asiatiques est bien inférieure à leur proportion dans la vie réelle. Un système de pondération a été mis en place pour réduire au maximum cette dépendance et une formule lise au point pour Wikipedia dans la V2. Pour la V3, après une tentative lourde mais infructueuse de rajouter l'ensemble des encyclopédies en ligne, la décision a été prise de supprimer cet indicateur et de reformuler la méthode.
TRAFIC	13	11	17	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. 3 autres outils ont été testés sans succès. Le biais de sélection est corrigé en V3 en offrant une sélection systématisée et adéquatement pondérée de sites.
INTERFACES	19	19	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. Idéalement, il faudrait évaluer l'ensemble du support technologique associé à chaque langue et le pondérer mais il n'existe pas de sources disponibles.
USAGE	12	12	16	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 V3 identifie et intègre les populations d'abonnés d'applications alternatives, en particulier en Chine, Inde ou Russie pour équilibrer les résultats et réduire les biais. Il reste un espace d'amélioration.

La dernière version était centrée sur le français, avec le soutien de l'OIF[11], et visait à réduire les biais susmentionnés, en essayant par des moyens alternatifs de refléter les applications des pays asiatiques qui offrent des services similaires à ceux de Wikimedia ou des réseaux sociaux les plus populaires.

Le lecteur peut trouver cette lecture introductive quelque peu ennuyeuse, présentant, avant les résultats, la méthodologie, les sources et les biais. Cependant, nous considérons comme un devoir éthique de donner les éléments d'appréciation des biais avant d'offrir des résultats pour éviter la pratique courante et désastreuse d'utiliser des données trouvées dans l'Internet comme une réalité, sans la précaution que devrait impliquer l'analyse de sa méthodologie de production et ses biais.

Que nous disent les résultats de l'étude de 2022 par rapport à ceux de 2021, 2017 et antérieur[12]?

Regardons d'abord les langues ayant le plus de contenus sur la Toile.

	LOCUTEURS CONNECTÉS	CONTENUS
Chinois	**18,46%**	**21,60%**
Anglais	**14,83%**	**19,60%**
Espagnol	**6,79%**	**7,85%**
Hindi	**4,19%**	**3,76%**
Russe	**3,51%**	**3,76%**
Français	**2,98%**	**3,33%**
Portugais	**2,99%**	**3,13%**
Arabe	**3,97%**	**3,09%**

Étant donné l'intervalle de confiance estimé à +-20%, il faut considérer que le chinois et l'anglais sont ex-aequo en première place avec entre 16 et 24% des contenus chacun. L'hindi, le russe le français et le portugais sont en quatrième position avec entre 3 et 4% des contenus chacun. En neuvième position se trouve le japonais et l'allemand, suivi par le malais, puis le turc, l'italien et le coréen.

Les 3 facteurs qui détermineront l'évolution future sont, par ordre d'importance : la démographie, le dépassement de la fracture numérique et la capacité à créer des contenus. La démographie favorise l'hindi, qui dépassera probablement rapidement le français et pourrait, à moyen terme, dépasser l'espagnol. La démographie finira par favoriser l'arabe par rapport aux autres langues voisines, dont le portugais, qui pourrait renforcer sa position face au russe.

Nous pouvons clairement voir que le centre de gravité de l'Internet s'éloigne rapidement du monde occidental, où il est né et a prospéré à ses débuts, vers les langues asiatiques et l'arabe. La démographie devrait à terme favoriser les langues du continent africain, et aussi les langues européennes de la

[11] http://francophonie.org

[12]L'Observatoire a mené des campagnes de mesures, avec une autre méthodologie, entre 1998 et 2007, qu'il a été obligé d'interrompre car l'évolution des moteurs de recherche (perte de fiabilité en tant qu'outil scientifique) a rendu obsolète la méthode. Les résultats sont toujours consultables sur http://funredes.org/lc.

colonisation africaine, mais la fracture numérique africaine est encore très marquée et plus lente à se réduire par rapport à la croissance mondiale de l'Internet. Ce tableau, réalisé avec les derniers résultats de 330 langues et actualisé en mars 2022, présente clairement cette situation :

	Langues africaines	Langues américaines	L'arabe comme macro-langue	Langues asiatiques	langues européennes	Langues non inclues	TOTAL
Internautes %	**29,8%**	**56,7%**	**64,0%**	**49,3%**	**82,6%**	**47,06%**	**56,91%**
Contenus	**2,89%**	**0,22%**	**3,09%**	**44,77%**	**45,39%**	**3,64%**	**100%**
Présence virtuelle	**0,31**	**0,71**	**0,88**	**0,93**	**1,47**	**0,47**	**1**
Productivité des contenus	**0,56**	**0,69**	**0,79**	**1,00**	**1,15**	**0,57**	**1**
Locuteurs L1+L2	**9,21%**	**0,31%**	**3,53%**	**48,24%**	**30,91%**	**7,81%**	**100%**
Population connectée	**5,21%**	**0,32%**	**3,89%**	**44,63%**	**39,51%**	**6,36%**	**100%**
Nombre de langues avec L1>1M	**138**	**8**	**1**	**135**	**47**		**329**

Le tableau se lit comme suit : il y a 138 langues d'Afrique couvertes par l'étude ; en moyenne, 29,8% de leurs locuteurs sont connectés à l'Internet, ensemble, elles représentent 9,21% du total mondial des locuteurs L1+L2, cependant, elles ne représentent que 2,89% du poids total de la Toile et 5,21% de la population L1+L2 connectée.

Les langues d'origine européenne continuent de dominer l'Internet mais l'essor récent des langues asiatiques est visible par leur première place en termes de personnes connectées, et la proportion des contenus associés est très proche de celui des langues européennes et continue de croître rapidement. Dans tous les cas, à mesure que le taux de connectivité des pays asiatiques augmentera et approchera le taux moyen exceptionnel des langues européennes, de plus de 80%, elles prendront également la première place en termes de contenus.

Le pourcentage de contenus favorise par définition les langues avec un plus grand nombre de locuteurs. Observons maintenant les langues de tête dans les macro-indicateurs de **présence virtuelle** et de **productivité de contenus**, ainsi que les **langues les plus connectées** pour avoir une indication *indépendante* du nombre de locuteurs.

	PRÉSENCE VIRTUELLE
Japonais	2,18
Norvégien	1,88
Allemand	1,82
Suédois	1,82
Danois	1,78
Hollandais	1,73
Finlandais	1,69
Catalan	1,68
Suisse allemand	1,63
Polonais	1,59
Italien	1,53
Estonien	1,51
Russe	1,51
Anglais	1,51

Les places de tête appartiennent clairement aux langues nationales des pays et régions reconnus pour leur leadership dans le domaine de la société de l'information.

Les langues les plus connectées sont les suivantes :

	% LOCUTEURS CONNECTÉS
Norvégien	96,89%
Danois	96,42%
Suédois	93,94%
Catalan	92,88%
Japonais	92,63%
Finlandais	92,07%
Suisse allemand	91,55%
Limbourgeois	91,42%
Flamand occidental	91,30%
Néerlandais	91,14%
Galicien	91,07%
Haut-saxon	89,81%
Estonien	89,26%
Allemand	89,17%
Letton	89,04%

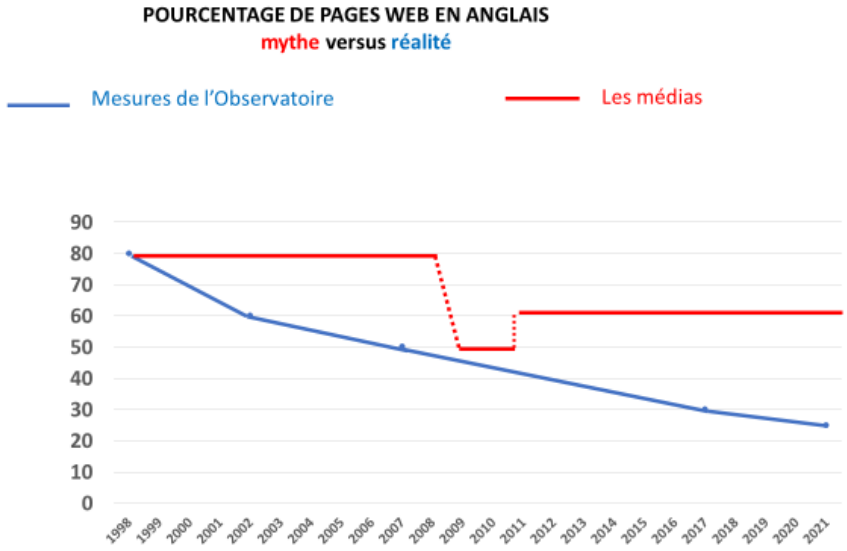
Et, enfin, les langues avec la plus haute productivité des contenus :

	Productivité des contenus
Japonais	1,34
Anglais	1,32
Chinois	1,17
allemand	1,16
Espagnol	1,16
Italien	1,14
Français	1,12
Norvégien	1,10
Suédois	1,10
Coréen	1,09
Hollandais	1,08
Russe	1,07
Grec	1,07
Capverdien	1,05
Danois	1,05
Portugais	1,05

Nous nous concentrons désormais sur les premières langues de l'Internet, à commencer par l'anglais. La source de données la plus utilisée, et longtemps la seule, sur la présence des langues sur le Web est W3Techs[13].

Comment est-il possible que cette source indique un pourcentage de contenus en anglais de 62,2% en 2022 alors que l'Observatoire propose un chiffre entre 18 et 24% ? La mesure directe des pages Web avec un algorithme de reconnaissance de la langue ne devrait pas se tromper. Mais oui, pourtant, et c'est à cause de la non prise en compte du multilinguisme! Il existe de fait une situation historique de mésinformation sur l'espace de l'anglais dans l'Internet, illustrée par ces deux courbes :

[13]https://w3techs.com/technologies/overview/content_language



Jusqu'en 2009, les médias publiaient des rapports sur la présence de l'anglais sur le Web le positionnant, inchangé depuis une décennie, à 80%, alors que nos mesures indiquaient une baisse progressive vers 50%. Les médias se sont basés sur 3 publications qui proposaient, avec la même méthodologie, les mêmes résultats, en 1997, 1999 et 2002. La méthodologie n'était pas vraiment biaisée mais statistiquement invalide[14], voir (Pimienta, 2009) pour plus de détails. Après les publications de l'UNESCO sur le sujet (Pimienta 2006, 2009), les médias[15] ont progressivement adopté cette nouvelle valeur de 50 %. Puis W3Techs est apparu comme la seule source, dont les résultats se sont maintenus à une valeur comprise entre 50 et 60% depuis 2011[16].

Comment fonctionne W3Techs et quel est le problème qui se pose par rapport au multilinguisme ?

W3Techs sélectionne les 10 millions de sites les plus visités sur le Web, tel qu'indiqué par l'application de marketing numérique Alexa (http://alexa.com). Mettons de côté les biais en faveur de l'anglais des algorithmes de reconnaissance de la langue et le fait que la sélection des 10 millions de sites les plus visités (sur près de 1,2 milliard de sites Web existants[17], soit moins de 1% d'entre eux) privilégient les sites en anglais. De nos jours, analyser l'ensemble du Web semble être une tâche inaccessible et notre intention n'est pas de sous-estimer le travail louable de W3Techs, qui fournit de nombreuses données très utiles. Concentrons-nous sur sa gestion du multilinguisme. W3Techs applique quotidiennement son algorithme sur **la page d'accueil** de ces 10 millions de sites.

La décision de se limiter à la page d'accueil fait partie du problème. La comptabilisation des langues sur le Web doit se faire au niveau des pages et non au niveau des sites, puisqu'un site web limité à une page et un autre avec des milliers de pages ne peuvent pas être comptabilisés de la même manière. À cela

[14]Un algorithme de reconnaissance de la langue était appliqué à la page d'accueil de 3 000 sites choisis au hasard, à partir des numéros IP, et les pourcentages étaient calculés. La méthode statistique adaptée à cette approche serait de répéter cette mesure un grand nombre de fois avec d'autres échantillons, puis d'analyser la distribution de la variable aléatoire constituée par les résultats obtenus, en moyen des techniques statistiques (moyenne, variance, etc.). Un seul tir à l'arc sur une cible n'indique généralement pas la capacité du tireur !

[15]Et aussi malheureusement Wikipédia (https://en.wikipedia.org/wiki/Languages_used_on_the_Internet) dont on attend plus de prudence.

[16]Voir https://w3techs.com/technologies/history_overview/content_language/ms/y).

[17]Source:https://news.netcraft.com/archives/category/web-server-survey/

s'ajoute la possibilité que ce site aux milliers de pages ait des pages dans différentes langues, même si la page d'accueil est principalement en anglais, augmentant ainsi la taille de l'erreur. Aujourd'hui, de nombreux sites parmi les plus visités (comme Facebook par exemple) proposent des dizaines de versions linguistiques dans la page d'accueil ; comptabiliser la page d'accueil en anglais, c'est omettre toutes ces versions. Enfin, il est très fréquent que la page d'accueil d'un site dans une langue autre que l'anglais comporte certains mots en anglais (par exemple, des boutons de navigation ou le mot copyright) ; la compter comme une page en anglais, comme c'est probablement le cas avec l'algorithme W3Techs, c'est ignorer les pages dans d'autres versions linguistiques.

Il n'est pas nécessaire d'être statisticien pour comprendre que la méthode, en ne prenant pas en compte la réalité du multilinguisme, peut se tromper dans des proportions gigantesques... Que pourrait faire l'algorithme de W3Techs pour améliorer ses produits, sans abandonner une approche pragmatique, c'est-à-dire sans se lancer le défi d'analyser toutes les pages de tous les sites ?

- ✔ Analyser les options linguistiques offertes dans la page d'accueil et tenir compte de chaque option, autant que de la version anglaise.
- ✔ Trouver une méthode pour obtenir une estimation, même approximative, du nombre de pages du site et multiplier chaque version linguistique par cette valeur pour avoir une comptabilité en pages et non pas en sites ;
- ✔ Lorsque l'algorithme signale plus d'une langue sur la page d'entrée, par principe, ne pas la comptabiliser en anglais.

D'autres facteurs devraient nous alerter sur l'invraisemblance des données de W3Techs et attirer l'attention sur certaines anomalies statistiques symptomatiques d'une erreur grossière :

- cela n'a aucun sens que la proportion de contenus en anglais soit restée stable au cours des 10 dernières années alors que dans la même période les pays asiatiques et arabes ont envahi le Web et qu'un ensemble de langues non européennes[18] représente maintenant environ le tiers des utilisateurs;
- la présence des internautes anglophones (L1+L2) est passée de 32% en 2007 à 13% aujourd'hui;
- montrer le chinois avec seulement 1,5% des contenus et l'hindi avec 0,1% lorsque ces deux langues représentent respectivement 18,5% et 4,2% des personnes connectées.

Pour clore ce chapitre, le fait que la proportion de pages Web en anglais diminue ne signifie nullement que la présence en termes absolus de l'anglais diminue, ni qu'elle ait fini de croître ; cela signifie simplement que de nouvelles langues prennent de plus en plus de place, réduisant la proportion des autres langues présentes, dont l'anglais. Bien sûr, l'anglais reste une langue prépondérante dans l'Internet, dont la **productivité de contenus** est la plus haute, juste après le japonais, et avec une forte avance sur les suivantes.

Nous avons discuté dans (Pimienta, 2017) les biais dans différents projets et comment le manque de considération du multilinguisme peut conduire à des erreurs flagrantes. Le cas fréquent le plus typique est le calcul d'éléments basés sur L1+L2, en divisant par la population mondiale, ce qui provoque des erreurs de magnitude, cachées dans les valeurs du reste des langues. Le nombre de locuteurs L1+L2 est bien supérieur à la population mondiale, nous avions estimé la proportion de personnes multilingues à 25% en 2017, dans cette nouvelle version, Ethnologue nous propose un chiffre plus précis de 43%.

[18]Chinois, hindi, arabe, turc, bengali, vietnamien, ourdou, persan et marathi.

Voyons maintenant les langues qui suivent l'anglais. Le chinois est passé de la deuxième position en 2017 à la première ex-aequo, en termes de contenus, mais il occupe déjà la première place en termes de personnes connectées dans le monde et, contrairement aux pays occidentaux, où beaucoup sont au-dessus des 90% de personnes connectées, il a de la marge pour progresser. Le tableau suivant présente des données avec un biais minimal, obtenues à partir des données de l'UIT et de la Banque Mondiale sur les pourcentages de personnes connectées à l'Internet par pays, pondérés avec les données démolinguistiques d'Ethnologue[19]. La présence des langues asiatiques et de l'arabe est notable.

	% MONDIAL INTERNAUTES	% MONDIAL LOCUTEURS	% LOCUTEURS CONNECTÉS
Chinois	**18,46%**	14,72%	71,38%
Anglais	**14,83%**	13,01%	64,86%
Espagnol	**6,79%**	5,24%	73,72%
Hindi	**4,19%**	5,80%	41,16%
Arabe	**3,97%**	3,53%	63,99%
Russe	**3,51%**	2,49%	80,32%
Portugais	**2,99%**	2,49%	68,43%
Français	**2,98%**	2,58%	65,80%
Malais	**2,36%**	2,36%	56,93%
Allemand	**2,04%**	1,30%	89,17%
Japonais	**1,99%**	1,22%	92,63%
Turc	**1,17%**	0,85%	78,05%
Italien	**0,87%**	0,66%	75,83%

Pour conclure, une analyse historique des langues dans l'Internet.

PÉRIODE	FONCTIONNALITÉS
1970-1990	L'Internet est né dans le monde occidental, très marqué par la langue anglaise dans sa phase historique initiale, tant pour des raisons technologiques (la langue des professionnels du réseau[20]) que par la nature de ses premiers utilisateurs (le monde de la recherche), dont une forte proportion utilise l'anglais comme L2 même si ce n'est pas sa langue maternelle. L'anglais dominait le réseau pendant cette période.
1990-2010	Cette période correspond à la naissance du Web (1992) : les langues européennes investissent l'Internet, qui devient un espace privilégié pour ces langues, avec une dominance de l'anglais qui passe de 80 % à 50 % sous l'effet de la poussée des autres langues européennes.
2010-2020	L'Internet devient à la fois le moteur et le sujet de la mondialisation, et la proportion d'internautes ayant l'anglais comme L1 ou L2 diminue rapidement pour se rapprocher du pourcentage mondial réel (moins de 20 %). Ainsi, sa proportion sur le Web est logiquement proche de sa proportion dans le monde réel, tout en gardant un avantage historique. Cependant, le continent africain est encore à la traîne et la fracture numérique y est encore énorme[21].
2020-2030	Nous entrons dans une nouvelle phase de mondialisation où le poids démographique commence à être le facteur dominant, du moins au sein du monde arabe et asiatique. Dans cette période, le centre de gravité linguistique de l'Internet va se déplacer vers les langues asiatiques et l'arabe et si le continent africain parvient à surmonter sa fracture numérique, sa démographie pourrait réserver des surprises...

[19]Et, comme le reste, sur la base L1+L2.

[20] Pendant cette période initiale, l'adoption comme standard du code ASCII à 7 bits empêchaient l'utilisation d'accents et d'autres signes diacritiques tels que le tilde, et cela a pris plusieurs années avant d'être surmonté.

[21] En 2021, sur les 56 pays ayant un taux de connexion à l'Internet inférieur à 30%, 34 sont africains, 7 asiatiques et 6 sud-américains. Sur ces 34 pays africains, 14 comptent moins de 10 % de personnes connectées.

En conclusion, nous pensons que la croyance selon laquelle la lingua franca d'Internet est l'anglais est une illusion : ce qui caractérise de plus en plus l'Internet, c'est le multilinguisme[22] et l'économie numérique est clairement de plus en plus dirigée par le facteur du multilinguisme,

La lutte contre la désinformation est devenue un enjeu majeur en cette période de crise sanitaire où elle peut entraîner la mort. Selonnotre vision (Pimienta, Rodriguez, 2020), la nécessité de développer des programmes de **littératie numérique** vastes et complets est une urgence aussi aiguë que celle du réchauffement climatique. Clairement, ces programmes doivent inclure l'éducation des citoyens à traiter les données proposées sur le Web, avec un esprit critique, et une exigence ferme de transparence méthodologique et algorithmique, ainsi qu'une présentation honnête des biais inhérents à toute approche de données construites, que ce soit avec des méthodes statistiques ou autres. Il est évident que les progrès de l'intelligence artificielle, basée sur une utilisation intensive des données, rendent ce besoin d'autant plus critique.

RÉFÉRENCES

- Pimienta D. (2021). « Version nouvelle et améliorée d'une approche alternative pour la production d'indicateurs linguistiques dans l'Internet. », *Observatoire de la diversité linguistique et culturelle dans l'Internet*. http://funredes.org/lc2021/ALI%20V2-ES.pdf

- Pimienta D., Rodriguez LG. (2020) "Rock the Internet Blues: Une vision critique de l'évolution de l'Internet depuis la société civile", publié en espagnol dans *Revista Ibero-Americana de Ciência da Informação*, V13 N3, Pp. 979-1000 - Traduction au français : http://funredes.org/RokInternetBlues

- Pimienta D. (2019) Une approche alternative pour produire des indicateurs de présence des langues dans l'Internet. *Observatoire de la diversité linguistique et culturelle dans l'Internet*. http://funredes.org/lc2019/Alternativa%20Lengua%20Internet.docx

- Pimienta D., Prado D., Blanco A. (2009). « Douze années de mesure de la diversité linguistique sur l'Internet: bilan et perspectives », UNESCO, CI-2009/WS/1. https://unesdoc.unesco.org/ark:/48223/pf0000187016_fre

- Pimienta D. (2005) « Diversité linguistique dans le cyberespace : modèles de développement et de mesure », dans *Mesurer la diversité linguistique dans l'Internet*, UNESCO, CI.2005/WS/06. https://unesdoc.unesco.org/ark:/48223/pf0000142186_fre

[22] L'Internet est probablement l'espace humain où le multilinguisme s'exprime le mieux et le plus, compte tenu de ses caractéristiques sans frontières, son degré de multilinguisme pourrait bien être supérieur à celui des humains, que ce soit en termes de contenus, de trafic, d'usages ou d'interfaces...