Ressource : Indicateurs de présence des langues sur Internet

Daniel Pimenta
Observatoire de la diversité linguistique et culturelle dans l'Internet
http://funredes.org/lc
Lien de la ressource : http://funredes.org/lc2022

Résumé

Des indicateurs fiables et maintenus de l'espace des langues sur Internet sont nécessaires pour soutenir des politiques publiques appropriées et des études linguistiques bien informées. Les sources actuelles sont rares et souvent fortement biaisées. Le modèle de production d'indicateurs sur la présence des langues dans l'Internet, lancé par l'Observatoire en 2017, a atteint un niveau de maturité raisonnable et ses produits de données sont partagés sous licence CC-BY-SA 4.0. Il traite désormais 329 langues (locuteurs L1 > un million) et tous les biais associés au modèle ont été contrôlés à un seuil acceptable, permettant de compter sur un intervalle de confiance estimé à +-20%. Certains des indicateurs (principalement le pourcentage de locuteurs L1+L2 connectés à l'Internet par langue et dérivés) reposent sur Ethnologue Global Dataset #24 pour les données démolinguistiques et l'UIT, complété par la Banque mondiale, pour le pourcentage de personnes connectées à l'Internet par pays. Le reste des indicateurs s'appuie sur les sources précédentes plus une grande combinaison de centaines de sources différentes pour les données liées aux contenus Web par langue. Cet article porte sur la description des nouvelles ressources linguistiques créées. Les considérations méthodologiques ne sont exposées que brièvement et seront développées dans un autre article.

Mots clés: Ressource linguistique, Langues, Internet, Indicateurs, Multilinguisme

1. Introduction

L'Observatoire de la Diversité Linguistique et Culturelle dans l'Internet[1] travaille depuis 1996 sur des méthodes alternatives de mesure des indicateurs de présence des langues dans l'Internet. La méthode standard pour calculer le pourcentage de contenus Web par langue consiste logiquement à appliquer un algorithme de reconnaissance de la langue à toutes les pages Web existantes et à compter. L'énorme extension du Web rend cette approche peu pratique, sauf pour cibler des sous-ensembles plus petits, comme cela a été fait efficacement par le Language Observatory Project, avant que le projet ne disparaisse (Mikami, 2005).

Les tentatives d'utiliser cette approche en l'appliquant à une cible avec un nombre limité de pages Web, censées représenter fidèlement toute la Toile, sont sujettes à d'énormes biais, comme le montre la méthode définie par Alis Technologies en 1997[2] et réutilisée en 1999 (Lavoie, 1999) et 2003 (O'Neil, 2003) par OCLC. Huit mille sites Web ont été sélectionnés au hasard en fonction des numéros IP et les conclusions ont été tirées à partir d'une mesure unique, au lieu d'une série répétitive traitée statistiquement comme une variable aléatoire.

Depuis 2011, W3Techs[3], certainement un excellent fournisseur de statistiques fiables pour les technologies Web, fournit des résultats mis à jour quotidiennement pour les contenus Web par langue, en appliquant un algorithme de reconnaissance de la langue à la page d'accueil des 10 millions de sites Web classés comme les plus visités par Alexa.com[4]. La méthode est analogue à celle utilisée pour les 25 autres technologies Web étudiées par cette entreprise, fournissant des résultats extrêmement intéressants. Cependant, les langages sont un type de technologie Web assez différent des bibliothèques de scripts Java ou des logiciels de serveurs Web et traiter les langues des pages Web de la même manière peut entraîner d'énormes erreurs.

Le problème commence par le fait de se concentrer uniquement sur les **pages d'accueil** de la sélection de sites Web : si vous envisagez de calculer les pourcentages de contenus sur la Toile, par langue, vous devez viser les pages Web afin d'éviter de donner le même poids à un site Web de dix pages par rapport à un site Web de dix mille pages. De plus, les pages d'accueil des sites non anglophones comportent assez souvent des mots en anglais (soit par volonté de présenter le site en anglais, soit parce que quelques mots anglais tels que *copyright*, *abstract* ou pour les boutons de navigation sont présents). C'est une cause d'erreur pour l'algorithme. Le gros de l'erreur est de toute façon ailleurs : il est causé par le **manque de considération pour le multilinguisme** qui fait que l'algorithme compte comme des sites anglais uniquement, des sites qui offrent également des dizaines d'options linguistiques dans leurs interfaces. Très souvent, le site Web définit automatiquement l'option de langue, selon les préférences

[1] http://funredes.org/lc
[2] https://web.archive.org/web/20010730164601/http://alis.isoc.org/palmares.en.html
[3] http://W3Techs.com
[4] Un site de collecte et d'analyse du trafic Web appartenant à la société Amazon, sur le point d'être retiré du marché.

de l'utilisateur, une pratique de plus en plus courante, en particulier pour les sites les plus visités au niveau mondial (Facebook.com n'est qu'un exemple) et l'algorithme ne compte qu'une seule langue pour la page d'accueil, l'anglais. Pas étonnant alors que, depuis 2011, le pourcentage d'anglais sur le Web est maintenu stable et même en croissance par W3Techs, malgré le fait qu'il est évident que l'Internet a radicalement changé au cours de la dernière décennie, le chinois devenant la première langue en termes d'utilisateurs, et la plupart des langues asiatiques et l'arabe étant en plein essor. Le Web est aujourd'hui probablement plus multilingue que l'humanité.

Selon les dernières données d'Ethnologue, le rapport des locuteurs L1+L2 sur les locuteurs L1 est de 10 361 716 756 / 7 231 699 136 = 1,43. Personne ne sera alors surpris que plus de 50% des sites Web affichent des pages dans plus d'une seule langue. Ne pas donner l'attention méritée au multilinguisme devient un biais inacceptable pour de telles études.

W3Techs pourrait, sans changer sa sélection actuelle de sites Web et son programme de base, corriger ses biais géants avec quelques remaniements tels que :

- Analyser les options linguistiques proposées sur la page d'accueil et compter chaque option, pas seulement la version anglaise.
- Trouver une méthode pour obtenir une estimation approximative du nombre de pages du site Web et multiplier chaque version linguistique par ce nombre afin de compter les pages Web au lieu des sites Web.
- Lorsque l'algorithme signale plus d'une langue sur la page d'accueil, par précaution, ne pas compter ce site Web comme anglais, mais plutôt comme la deuxième langue trouvée.

Les nouveaux résultats seraient alors radicalement différents…

Le problème inquiétant est qu'en raison de l'unicité de la source, de la qualité éprouvée du reste de ses enquêtes, de son histoire à long terme et de son marketing efficace, un grand pourcentage de la communauté de la recherche linguistique (et des décideurs publics) utilise les données de W3Techs sans questionnement. Malheureusement, de bonnes théories alimentées par des chiffres erronés peuvent difficilement fournir des résultats corrects.

L'exemple le plus symptomatique de la situation est donné par l'agrégateur de statistiques Statista[5]qui intitule son annonce 2022 sur les langues dans l'Internet[6] avec une déclaration qui sonne comme un fait indéniable : *l'anglais est la langue universelle d'Internet*, étayée par les données de W3techs, où les contenus Web en anglais représentent 63,7 % du total tandis que ceux en chinois n'en représentent que 1,3 %.

Dans le même temps, l'Observatoire de la Diversité Linguistique et Culturelle dans l'Internet annonce l'anglais et le chinois, ensemble, au même pourcentage, autour de 20%, tandis que l'hindi, avec ses 224 millions d'internautes, atteint 3,8% (contre seulement 0,1% mesuré par W3Techs) et conclut sa dernière annonce par cette phrase : *La transition de l'Internet entre la domination des langues européennes, anglais en tête, vers les langues asiatiques et l'arabe, chinois en tête, est bien avancée et le gagnant est le multilinguisme, mais les langues africaines tardent à prendre leur place.*

L'une, au moins, des deux sources doit être extrêmement erronée et les chercheurs devraient faire preuve de prudence et vérifier les biais d'une méthode avant de tirer des conclusions à partir des données qu'elle produit…

2. Les méthodes alternatives

Pendant la période passée 1998-2007, la méthode alternative de l'Observatoire, qui a fourni des séries cohérentes pendant une décennie, se limitait à l'anglais, l'allemand et les 5 langues latines (français, italien, espagnol, portugais et roumain). Elle utilisait les moteurs de recherche pour compter un vocabulaire comparable[7]pour chaque langue (Pimienta, 2009). Après 2007, «l'évolution marketing» des moteurs de recherche a rendu la méthode obsolète car leurs retours du nombre d'occurrence d'un mot recherché perdaient totalement en fiabilité.

Aujourd'hui, 329 langues sont considérées, celles comptant plus d'un million de locuteurs L1, selon Ethnologue, limitation adoptée pour éviter des biais trop forts conséquence de l'hypothèse de travail de l'approche : *toutes les langues pratiquées dans un même pays sont comptabilisés avec le même pourcentage de locuteurs connectés à*

---

[55] http://statista.com En cours de route, je ne perdrais pas l'occasion de questionner l'éthique de deux phénomènes émergents qui pourraient bien être corrélés. 1) Trop de chercheurs paresseux citent Statista comme source de données au lieu de la véritable source. 2) Statista propose certaines données en libre accès mais l'identification de la source de ces données n'est accessible qu'aux clients payants. Faisons donc plus simple et citons Google comme la mère de toutes les sources ou, encore plus simple, citons Internet comme la matrice de toutes les sources ! 😊

[6] https://www.statista.com/chart/26884/languages-on-the-internet/

[7] Un ensemble de mots pour chaque langue, sélectionné avec beaucoup de précautions linguistiques, dont les occurrences ont été rapportés par les moteurs de recherche et ont permis, par comptage, les résultats.

l'Internet, chiffre national fourni par l'UIT ou la Banque mondiale. Cette hypothèse interdit de comparer les langues au sein d'un même pays, elle est difficilement applicable aux langues à faible nombre de locuteurs, et tend à donner un biais positif pour les langues d'immigration dans les pays en développement (qui peuvent être moins connectées que la moyenne) et, à l'inverse, un biais négatif pour les langues européennes dans les pays en développement (qui ont tendance à être mieux connectées que la moyenne).

La méthode actuelle est une **approximation indirecte des contenus**, basée sur l'observation expérimentale que *le rapport entre le pourcentage mondial de contenus et le pourcentage mondial de locuteurs connectés est toujours resté compris entre 0,5 et 1,5* (pour les langues à existence numérique complète).

Cela suggère l'existence d'une sorte de loi économique naturelle, qui lierait, pour chaque langue, **l'offre** (contenus et applications web) **à la demande** (locuteurs connectés à l'Internet). Lorsque le nombre de personnes connectées augmente, le nombre de pages web augmente logiquement en même temps, et à peu près dans la même proportion. Cela se produit parce que les gouvernements, les entreprises, les institutions éducatives, etc., et certains particuliers créent des contenus pour répondre à cette demande.

En outre, des enquêtes et des études ont constamment rapporté que les internautes préfèrent utiliser leur langue maternelle dans l'Internet et profitent également de l'occasion pour utiliser, comme deuxième option, leur(s) deuxième(s) langue(s)[8].

Ainsi, en fonction de chaque langue, il y a une sorte de **modulation** du rapport mentionné (contenus / connectés), pour le rendre supérieur ou inférieur à un. Cela signifierait que certaines langues ont plus de production de contenus que d'autres, en fonction d'un ensemble de facteurs concernant les langues dans leur contexte national, tels que :

- Évidemment, le nombre correspondant de **locuteurs L2**, puisque certaines personnes produisent, par exemple pour des raisons économiques, des contenus dans une langue différente de leur langue maternelle.

Mais aussi:

- La proportion du **trafic** Internet en fonction du contexte tarifaire, culturel ou éducatif du pays.
- Le nombre **d'abonnements** aux réseaux sociaux et autres applications Internet.
- Le **support technologique numérique** de la langue et sa présence dans les interfaces d'application et les programmes de traduction, qui faciliteraient ou non la production de contenus.
- Le niveau de submersion du pays où vit le locuteur en termes de manifestation de la **société de l'information** (commerce électronique, applications du gouvernement pour payer les impôts, etc.).

Ainsi, s'il était possible de collecter différents indicateurs sur chacune des caractéristiques évoquées, on approcherait la fluctuation de la modulation des contenus web autour de la valeur un et en déduirait en quelque sorte la proportion des contenus. C'est le **cœur de la méthode** et il est synthétisé dans le schéma suivant qui montre tous les indicateurs qui sont traités pour chaque langue et la quantité correspondante de sources que le modèle utilise. La première et la deuxième version de la méthodologie sont entièrement documentées (y inclus l'analyse de tous les bais), voir comme point de départ la synthèse dans (Pimienta, 2019). La description détaillée de la version 3 est en route.

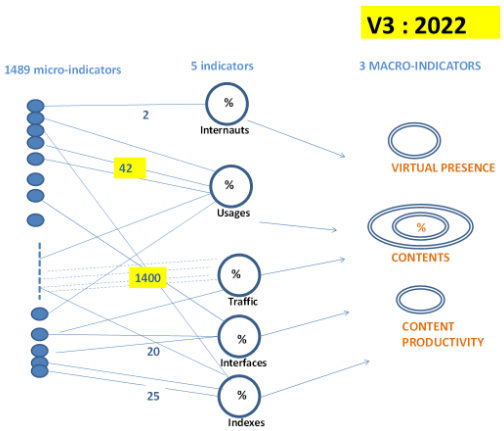

Figure 1: Schéma des indicateurs

[8] Voir par exemple le rapport d'enquête de l'Union européenne https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556 ou, pour le cas difficile de l'Inde, ce rapport : https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

Ce schéma a évolué, de la version 1 à la version 3, selon les progrès de la tâche difficile, mais indispensable, de traquer les biais, tant en nombre de sources que d'indicateurs. Les calculs du modèle assez complexe établi reposent principalement sur une variété **d'opérations de pondération**, avec, la plupart du temps, **le vecteur de pourcentage de personnes connectées par pays**, qui est le centre de gravité mathématique du processus.
Les sources d'indicateurs par langue disponibles sont rares ; la majorité des indicateurs sont obtenus par pays et, la plupart d'entre eux ne couvrent qu'un sous-ensemble de pays. La source de données est donc extrapolée à tous les pays, en pondérant avec les données de connectivité mentionnées, et la transformation des données par pays en données par langue est obtenue par pondération avec les données démolinguistiques (quantité de locuteurs L1+L2 de chaque langue dans chaque pays).

3. Indicateurs produits par le modèle

Pour chacune des 329 langues traitées, le modèle produit les indicateurs suivants, pour chaque langue (notez que tous les pourcentages mondiaux sont basés sur les chiffres L1+L2 et représentent la part correspondante pour chaque langue).

Indicateurs intermédiaires :
*Internautes :*locuteurs connectés à l'Internet
Usages
Trafic
Interfaces et programmes de traduction : en termes de pourcentage mondial du nombre correspondant d'interfaces et de programmes de traduction qui incluent la langue
Index: en termes de pourcentage mondial de la notation des pays dans les paramètres de la société de l'information

Sorties du modèle (également appelées macro-indicateurs) :
Locuteurs connectés: pourcentage du total mondial des locuteurs L1+L2 connectés à l'Internet
Contenus: pourcentage de contenus Web (calculé comme la moyenne des 5 indicateurs intermédiaires)
Productivité du contenu: ratio contenus/Internautes
Coefficient de présence virtuelle: ratio contenus/ locuteurs

Indicateurs plus avancés
Cyber-géographie des langues: une répartition des sorties du modèle cumulées par familles de langues (européennes, asiatiques, arabes, américaines ou africaines)

Indicateur de cyber-mondialisation

$$CGI(L) = (L1+L2)/L1(L) \times S(L) \times C(L)$$

Où:

L1+L2/L1(L) est le rapport du multilinguisme de la langue L
S(L) est le pourcentage de pays du monde qui détiennent des locuteurs de la langue L
C(L) est le % de locuteurs de la langue L connectés à l'Internet.

C'est un indicateur des **atouts stratégiques** d'une langue dans le cyberespace.

De plus, pour certaines langues, il a été affiché la liste des pays qui détiennent les pourcentages les plus importants de locuteurs connectés.

Les fichiers Excel contenant les résultats finaux peuvent être téléchargés à l'adresse http://funredes.org/lc2022.

Une base de données d'accès aux résultats, avec possibilité d'interrogation par nom de langue ou code iso, est en projet.

4. Exemples d'indicateurs produits

Ci-après sont présentés quelques exemples de données, limités aux meilleurs résultats, pour la majorité des cas. Les mêmes données sont disponibles pour chacune des 329 langues traitées.

La pyramide inversée ci-dessous doit être lue comme l'expression de l'intervalle de confiance : le pourcentage de chinois dans les contenus Web est compris entre 16 % et 24 %, toutes les langues restantes représentent ensemble entre 18 % et 26 % du total des contenus.

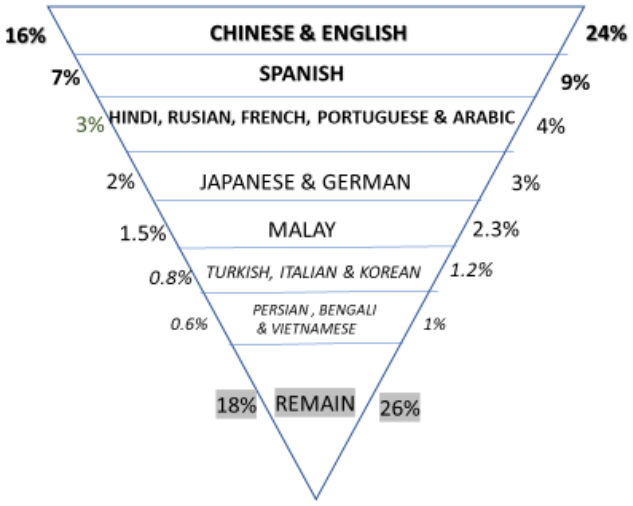

Figure 2: Fenêtre de pourcentages de contenus pour les principales langues

LANGUE	LOCUTEURS CONNECTÉS
Norvégien	**96,89%**
Danois	**96,42 %**
Suédois	**93,94 %**
Catalan	**92,88%**
Japonais	**92,63 %**
Finlandais	**92,07 %**
Allemand, Suisse	**91,55 %**
Limbourgeois	**91,42 %**
Flamand occidental	**91,30%**
Néerlandais	**91,14 %**
Galicien	**91,07 %**
Saxon, Supérieur	**89,81 %**
Estonien	**89,26%**
Allemand, Standard	**89,17%**
Letton	**89,04 %**
Bavarois	**88,24%**

Table 1: Principales langues en termes de locuteurs connectés

Rang Contenus			Internautes	Population Mondiale	Locuteurs connectés	Contenus	Présence Virtuelle	Productivité Contenus
L1+L2	ISO	LANGUES	L1+L2	L1+L2	L1+L2	L1+L2	L1+L2	L1+L2
1	**zho**	***chinois***	18,46%	14,72%	71,38%	21,60%	1,47	1,17
2	**eng**	**anglais**	14,83%	13,01%	64,86%	19,60%	1,51	1,32
3	**spa**	**espagnol**	6,79%	5,24%	73,72%	7,85%	1,50	1,16
4	**hin**	**hindi**	4,19%	5,80%	41,16%	3,76%	0,65	0,90
5	**rus**	**russe**	3,51%	2,49%	80,32%	3,76%	1,51	1,07
6	**fra**	**français**	2,98%	2,58%	65,80%	3,33%	1,29	1,12
7	**por**	**portugais**	2,99%	2,49%	68,43%	3,13%	1,26	1,05
8	**ara**	***arabe***	3,97%	3,53%	63,99%	3,09%	0,87	0,78
9	**jpn**	**japonais**	1,99%	1,22%	92,63%	2,66%	2,18	1,34
10	**deu**	**allemand**	2,04%	1,30%	89,17%	2,37%	1,82	1,16
11	**msa**	***malais***	2,36%	2,36%	56,93%	1,96%	0,83	0,83
12	**tur**	**turc**	1,17%	0,85%	78,05%	1,14%	1,35	0,98
13	**ita**	**italien**	0,87%	0,66%	75,83%	1,00%	1,53	1,14
14	**kor**	**coréen**	0,90%	0,79%	65,16%	0,98%	1,24	1,09
15	**fas**	***persan***	1,08%	0,81%	75,91%	0,88%	1,09	0,82
16	**ben**	**bengali**	1,11%	2,58%	24,55%	0,88%	0,34	0,79
17	**vie**	**vietnamien**	0,92%	0,74%	70,96%	0,85%	1,15	0,92
18	**urd**	**ourdou**	0,95%	2,22%	24,38%	0,66%	0,30	0,70
19	**tha**	**thaïlandais**	0,80%	0,59%	77,95%	0,65%	1,12	0,82
20	**pol**	**polonais**	0,60%	0,39%	87,09%	0,63%	1,59	1,04
21	**mar**	**marathe**	0,69%	0,96%	41,06%	0,58%	0,60	0,83
22	**tel**	**télougou**	0,68%	0,92%	41,69%	0,56%	0,60	0,82
23	**tam**	**tamil**	0,61%	0,82%	42,15%	0,51%	0,62	0,83
24	**jav**	**javanais**	0,62%	0,66%	53,76%	0,44%	0,66	0,70
25	**nld**	**néerlandais**	0,38%	0,24%	91,14%	0,41%	1,73	1,08
26	**guj**	**gujarati**	0,44%	0,60%	41,47%	0,36%	0,61	0,83
27	**ukr**	**ukrainien**	0,40%	0,32%	71,02%	0,35%	1,09	0,88
28	**kan**	**kannada**	0,41%	0,57%	41,11%	0,33%	0,59	0,82
29	**ron**	**roumain**	0,32%	0,23%	79,57%	0,30%	1,29	0,93
30	**aze**	***azerbaïdjanais***	0,33%	0,23%	81,54%	0,28%	1,21	0,85
		RESTE	**22,60%**	**30,10%**		**15,13%**		
		TOTAL	**100,00%**	**100,00%**		**100,00%**		

Table 2: Principaux indicateurs pour les 30 principales langues en pourcentage de contenus

Le tableau doit être lu ainsi : l'anglais représente 13% de la population mondiale L1+L2 et 14,8% de la population connectée à l'Internet ; 64,7 % des anglophones L1+L2 sont connectés à l'Internet ; 19,6 % des contenus Web est en anglais; le coefficient de présence virtuelle de l'anglais est de 1,5, ce qui signifie que les contenus en anglais sont surreprésentés dans un facteur supérieur à 50 % ; la productivité de contenus en anglais est de 1,32, la plus élevée après le japonais.

Les macro langues sont mentionnées en cursive.

LANGUE	PRÉSENCE VIRTUELLE
Japonais	**2.18**
Norvégien	**1,88**
Allemand, Standard	**1,82**
Suédois	**1,82**
Danois	**1,78**
Néerlandais	**1,73**
Finlandais	**1,69**
Catalan	**1,68**
Allemand, suisse	**1,63**
Polonais	**1,59**
Italien	**1,53**
Estonien	**1.51**
Russe	**1.51**
Anglais	**1.51**
Hébreu	**1,50**
Grec	**1,50**
Espagnol	**1,50**
Chinois	**1,47**
Letton	**1.46**
Galicien	**1.46**

Table 1: Langues principales en présence virtuelle

LANGUE	Productivité Contenus
Japonais	**1.34**
Anglais	**1.32**
Chinois	**1.17**
Allemand, standard	**1.16**
Espagnol	**1.16**
Italien	**1.14**
Français	**1.12**
Norvégien	**1.10**
Suédois	**1.10**
Coréen	**1.09**
Néerlandais	**1.08**
Russe	**1.07**
Grec	**1.07**
Capverdien	**1.05**
Danois	**1.05**
Portugais	**1.05**
Finlandais	**1.04**
Polonais	**1.04**
Catalan	**1.03**
Allemand, suisse	**1.02**
Hébreu	**1,00**

Table 2: Langues principales en productivité des contenus

LANGUES DE (*)	AFRIQUE	AMÉRIQUES	MONDE ARABE	ASIE	EUROPE	PACIFIQUE (**)
Internautes	**29,8 %**	**56,7 %**	**64,0 %**	**49,3 %**	**82,6 %**	
Contenus	**2,89 %**	**0,22 %**	**3,09 %**	**44,77%**	**45,39%**	
Présence Virtuelle	**0,28**	**0,68**	**0,87**	**0,65**	**1.39**	
Productivité de contenus	**0,51**	**0,68**	**0,78**	**0,72**	**0,95**	
Population mondiale	**9,15 %**	**0,*31 %**	**3,53 %**	**48,21%**	**30,91 %**	
Population connectée en % mondial	**5,18 %**	**0,32 %**	**3,89 %**	**44,60%**	**39,51 %**	
Nombre de langues	**138**	**8**	**1**	**135**	**47**	**0**

Table 3: Cyber-géographie des langues

(*) Doit être compris comme langues indigènes. Par exemple, les 8 langues indigènes des Amériques avec plus d'un million de locuteurs de L1 inclus dans le modèle sont : l'aymara, le guarani, le créole haïtien, le hunsrik, le créole jamaïcain, le q'eqchi', le kiche et le quechua.
(**) Aucune langue du Pacifique n'est incluse car aucune ne compte plus d'un million de locuteurs.

LANGUE	CGI	CGI %
Anglais	**1,61**	**14,24 %**
Français	**1,09**	**9,66 %**
Allemand	**0,42**	**3,75 %**
Russe	**0,31**	**2,76 %**
Espagnol	**0,27**	**2,40 %**
Arabe	**0,18**	**1,56 %**
Malais	**0,17**	**1,51 %**
Italien	**0,17**	**1,50 %**
Chinois	**0,16**	**1,46 %**
Portugais	**0,15**	**1,37 %**
Thaïlandais	**0,15**	**1,37 %**
Romani	**0,15**	**1,35 %**
Turc	**0,15**	**1,34 %**
Grec	**0,15**	**1,31 %**
Ukrainien	**0,15**	**1,31 %**
Polonais	**0,13**	**1,15 %**
Persan	**0,12**	**1,10 %**
Roumain	**0,12**	**1,06 %**
Hindi	**0,12**	**1,04 %**

Table 4: Indicateur de cyber-mondialisation

La deuxième colonne est calculée en divisant la valeur CGI par le total des CGI pour toutes les langues traitées. Il est évoqué comme un moyen de mesurer, par exemple, le poids relatif des deux premières positions, proche de 25% du total.

CHINOIS	L1+L2	%Connectés	Connectés	% total connectés
TOTAL	1 525 335 340	71,38%	1 088 735 519	100%
Chine	1 448 870 000	70,64%	1 023 512 815	94,01 %
Chine–Taïwan	37 320 000	88,82 %	33 148 541	3,04 %
Chine–Hong Kong	10 942 800	92,41 %	10 112 585	0,93 %
Malaisie	7 838 700	89,56 %	7 019 949	0,64 %
Singapour	4 026 000	75,88%	3 054 766	0,28 %
États-Unis	2 894 390	88,50%	2 561 503	0,24 %
Vietnam	2 500 000	70,64%	1 766 054	0,16 %
Indonésie	2 054 000	53,73%	1 103 542	0,10 %
Thaïlande	1 729 000	77,84%	1 345 918	0,12 %
Canada	1 212 600	97.00%	1 176 222	0,11 %
Philippines	1 010 280	43,03%	434 689	0,04 %
RESTE	4 937 570	71,04 %	3 507 738	0,32 %

Table 5: Répartition des locuteurs du chinois connectés par principaux pays

HINDI	L1+L2	%Connectés	Connectés	% total connectés
TOTAL	600 800 970	41,15%	247 258 401	100%
Inde	596 000 000	41,00%	244 360 000	98,87 %
Koweït	700 000	98,60 %	690 200	0,28 %
États-Unis	643 000	88,50%	569 048	0,23 %
Népal	1 307 600	25,00 %	326 900	0,13 %
Afrique du Sud	463 000	68,00 %	314 840	0,13 %
Arabie Saoudite	171 000	97,86 %	167 345	0,07 %
Australie	160 000	86,54%	138 472	0,06 %
Canada	111 000	97.00%	107 670	0,04 %
Yémen	316 000	30,00 %	94 800	0,04 %
DU REPOS	929 370	52,63%	489 127	0,20 %

Table 6: Répartition des locuteurs de l'hindi connectés par principaux pays

5. Références bibliographiques

Ethnologue Global Dataset (2022). https://www.ethnologue.com/product/ethnologue-global-dataset-0

Lavoie B.F., O'Neill E. T. (1999). How "World Wide" is the Web? *Annual review of OCLC Research*, https://web.archive.org/web/20031006155123/http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003496

Mikami Y., et al. (2005). The Language Observatory Project (LOP), In *Poster Proceedings of the Fourteenth International World Wide Web Conference*, pp. 990-991, May 2005, Japan

O'Neill E.T., Lavoie B.F., Bennett R. (2003). Trend in the Evolution of the Public Web: 1998 – 2002. *D-Lib Magazine*, 9.4 http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

OIF (2022). Le français dans le monde, Gallimard, ISBN : 9782072976865. Synthèse en ligne: https://francophonie.org/sites/default/files/2022-03/Synthèse_La_langue_française_dans_le_monde_2022.pdf

Pimienta, D., Prado D., Blanco A. (2009). Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, in *UNESCO Publications for the World Summit on the Information Society*, CI.2009/WS/1 http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

Pimienta D. (2019). Indicators of Languages in the Internet, in Proceedings of International Conference Language Technologies for All (LT4All), 4-6 December 2019, UNESCO, Paris; PP 315-319
https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.79.pdf

6. Remerciements

Les études de la version 3 ont été financées par l'Organisation Internationale de la Francophonie et les résultats ont alimenté le Chapitre Internet de l'ouvrage « Le français dans le monde » (OIF, 2022).

L'idée d'utiliser différentes sources de données par pays et de les transformer en données par langue a été imaginée pour la première fois par Daniel Prado en 2012.