Ceci est la révision par l’auteur de la traduction en français réalisée par l’outil gratuit et bien pratique <https://onlinedoctranslator.com>, lequel fait usage de GoogleTranslate en conservant le format du document source, de l’article paru dans **Forum of Linguistic Studies**| Volume 06 | Numéro 05 | Novembre 2024

https://journals.bilpubgroup.com/index.php/fls/index

**ARTICLE**

Est-il vrai que plus de la moitié du contenu Web est en anglais ? Pas si l’on prend en compte le multilinguisme !

*Daniel Pimienta* [*O*](https://orcid.org/0000-0001-7314-6471)

Observatoire de la Diversité Linguistique et Culturelle, 06000 Nice, France

INFO ARTICLE : Reçu : 24 août 2024 ; révisé : 4 septembre 2024 ; accepté : 14 septembre 2024 ; publié en ligne : 6 novembre 2024

DOI: <https://doi.org/10.30564/fls.v6i5.7144>

CITATION: Pimienta, D., 2024. Est-il vrai que plus de la moitié des contenus Web sont en anglais ? Pas si l'on accorde une attention particulière au multilinguisme ! Forum for Linguistic Studies. 6(5) : 201–212. DOI :<https://doi.org/10.30564/fls.v6i5.7144>

DROITS D'AUTEUR : Copyright © 2024 par l'auteur(e). Publié par Bilingual Publishing Co. Il s'agit d'un article en accès libre sous la licence Creative Commons Attribution - Pas d'Utilisation Commerciale 4.0 International (CC BY-NC 4.0) (https://creativecommons.org/licenses/by-nc/4.0/).

**Résumé**

La croyance selon laquelle l’anglais est et restera largement dominant comme première langue de l’Internet, en termes de contenu, et qu’il est la lingua franca naturelle dans le cyberespace joue contre la mobilisation des ressources humaines et financières pour intégrer les langues minoritaires. Nous soutenons que cette croyance repose sur des données biaisées et que le multilinguisme caractérise de plus en plus l'Internet et la traduction sa lingua franca. Nous contestons la validité d'une source largement utilisée, depuis 2011, pour affirmer que l'anglais représente un pourcentage stable de plus de 50 % des contenus Web. Cette source commerciale, W3Techs, est bien-réputée et considérée comme fiable pour ses enquêtes sur les technologies Web, explorant un large échantillon du Web. Cependant, les langues diffèrent des autres technologies Web, dans le fait que plusieurs langues peuvent être utilisées sur un site web. Ne pas tenir compte du caractère multilingue du Web est un biais important qui conduit à des erreurs majeures. L'étude du taux de multilinguisme de l'échantillon de sites Web utilisés par W3Techs permet de conclure que le pourcentage de contenus en anglais sur le Web se situe dans une fourchette de 20 à 30 %, une valeur cohérente avec les résultats des trois méthodes alternatives référencées. Nous prévoyons de créer pour 2025 un outil de mesure des langues et du taux de multilinguisme dans une série de sites Web, en veillant à identifier toutes les versions linguistiques des sites Web, une question complexe. Cet outil sera appliqué au même échantillonnage et devrait clore définitivement cette question.

***Mots clés:***Biais; Langues sur le Web; Multilinguisme; Anglais; Lingua franca

# Introduction

La présence archi-dominante des contenus en anglais dans l’Internet est tenue pour un fait avéré par la plupart des médias et de nombreux chercheurs, y compris ceux qui travaillent dans le domaine des technologies du langage. Cette situation influence insidieusement les politiques et les décisions techniques autour du concept de l'anglais comme lingua franca du cyberespace, c'est-à-dire la langue unique qui est systématiquement utilisée pour rendre possible la communication entre des personnes qui ne partagent pas la même langue.

Il y a, cependant, une contradiction entre la volonté de connecter tout le monde à l’Internet (aujourd'hui, 2/3 de la population mondiale est déjà connectée selon l'UIT) et la réalité du pourcentage limité de personnes dans le monde qui sont, soit des locuteurs en langue maternelle de l’anglais (notés L1), soit ayant l’anglais comme langue seconde (notés L2) ; L1+L2 désignant l'ensemble des personnes utilisant l'anglais soit comme première soit comme seconde langue. Différentes sources proposent des valeurs autour de 1,5 milliard pour la valeur L1+L2 pour l’anglais et aucune source ne franchit la ligne des 2 milliards. Cela implique que moins de 19% de la population mondiale comprend l'anglais.

« *Oui mais, plus de 50% des contenus web sont en anglais !* » Cet article montre que cette affirmation courante, appuyée par une source souvent citée, est fausse. En réalité, le chiffre correct se situe entre 20% et 30%, et cet article corrige le biais de cette source et pointe vers des sources alternatives pour confirmer la démonstration.

Pourquoi est-ce important ? Pourquoi est-ce une préoccupation pour les chercheurs qui se consacrent à l’intégration des langues moins bien dotées, sous-financées, menacées, minorées ou minoritaires ?

Certainement pas par rancœur ou par lutte politique contre la langue anglaise ! C'est important, parce que cette conviction a longtemps joué et joue encore un rôle démotivant pour la localisation et la création de contenu dans d'autres langues, en particulier celles qui sont minoritaires ou qui ont moins de ressources.

Ce chiffre ne concerne-t-il que les chercheurs en langues minoritaires ? Pas du tout ! Il s’agit également d’une préoccupation pour les entreprises, car la part du commerce électronique dans le commerce total a franchi le seuil des 20 % en 2020 et continue de croître[[1]](#footnote-1). De nombreuses sources expertes[[2]](#footnote-2) s'accordent à dire qu'il est indispensable que les applications de commerce électronique parlent la langue maternelle de leurs clients. Par conséquent, obtenir des chiffres erronés sur la réalité de la langue utilisée dans l’Internet pourrait se révéler néfaste pour les affaires.

Enfin, cet indicateur surestimé sert d’écran pour masquer la réalité du multilinguisme dans l’Internet, d’ores et déjà le domaine humain ayant la plus grande diversité linguistique et ce n’est qu’un début. Chaque jour, les internautes font un usage plus large de la variété complète de leurs langues (L1 et L2), que ce soit pour communiquer ou pour rechercher des informations, et cet usage est exactement la définition du multilinguisme. Le nombre approximatif de langues existant dans la sphère numérique (on dit qu’elles sont « localisées ») est passé de moins de 100 avant 2000, à 500 avant 2020, et atteint aujourd’hui 750[[3]](#footnote-3). Cette croissance ne s’arrêtera pas, même si l’objectif de quelque 8 000 langues existantes indique qu’il reste encore un long chemin à parcourir.

En février 2022, Statista[[4]](#footnote-4) a affirmé que « *l’anglais est la langue universelle de l’Internet* »[[5]](#footnote-5). Statista a appuyé cette affirmation avec les données mesurées par W3Techs[[6]](#footnote-6) qui affirmait, à la même date, que 63,7% des sites web étaient en anglais[[7]](#footnote-7).

La navigation dans les données historiques annuelles de W3Techs[[8]](#footnote-8) montre des chiffres toujours supérieurs à 50% depuis 2011 et aucune des autres langues n'y atteint jamais 10%. Les analyses basées sur ces chiffres sont innombrables et il est probable que de nombreuses politiques publiques ont été conçues sur les mêmes bases.

Entre 1998 et 2007, d'autres alternatives ont existé concernant la langue des contenus web [1]. Cependant, depuis 2011 et jusqu'en 2017, W3Techs est devenue la source unique de ces données et est logiquement devenue une référence universelle pour les données linguistiques sur les contenus web, promue non seulement par Statista mais aussi par Wikipedia[[9]](#footnote-9). La renommée de cette source s’étend à la communauté des chercheurs, même si sa méthodologie n’a pas été publiée et encore moins évaluée par des pairs.

Un examen minutieux de la méthodologie utilisée par W3Techs[[10]](#footnote-10) révèle l’absence de considération de la nature multilingue du Web. Une question légitime se pose alors : quels seraient les chiffres produits si la propriété multilingue de nombreux sites Web recevait l'attention qu'elle mérite de la part de la méthode pilotant son algorithme ?

Cet article aborde ce point en analysant, d’une part, la méthode utilisée par W3Techs et le biais résultant de l’hypothèse selon laquelle tous les sites Web sont monolingues. D’autre part, en calculant le « *taux de multilinguisme* » de l’échantillon du Web utilisé par W3Techs pour produire ses chiffres. Le taux de multilinguisme d’un ensemble de sites Web est défini par le rapport de la somme de toutes les versions linguistiques des sites Web sur le nombre total de sites Web. Sur la base de cette analyse, nous avons tenté de dé-biaiser, à l’aide d’une équation simple, la valeur produite par W3Techs du pourcentage de contenus en anglais.

Le résultat de cette tentative de rectifier les chiffres du W3Techs pour l'anglais conduit à la conclusion que le pourcentage corrigé de contenus en anglais sur le Web pourrait se situer entre 20% et 30%, et non pas dépasser les 50%. C'est la même fourchette fournie par une autre source, l'Observatoire de la diversité linguistique et culturelle dans l’Internet (OBDILCI[[11]](#footnote-11)). OBDILCI produit des indicateurs de langues dans l’Internet, depuis 2017, en utilisant une approche différente[[12]](#footnote-12). Cette fourchette de 20 à 30 % est également celle déduite d'une étude récente portant sur les sites Web des domaines nationaux de l'Union européenne [2]. Une troisième et dernière source de données sur la présence des langues sur le Web[[13]](#footnote-13) propose des chiffres extrêmement proches de ceux d'OBDILCI, avec l'anglais à 26,3%. Il semble (et cela reste à confirmer) que la méthode de Netsweeper consiste à mesurer les pages web et non les sites web, ce qui implique une prise en compte complète du multilinguisme. La taille de l'échantillon de Netsweeper est également bien supérieure à celle de W3Techs avec 12 milliards de pages annoncées (contre 1 million de sites web de l'échantillon utilisé par W3Techs).

Bien que les études quantitatives sur la nature multilingue du Web soient rares et qu’aucune n’ait jusqu’à présent abordé spécifiquement le taux de multilinguisme du Web, des études qualitatives sur le sujet existent. Dès 2007, [3] s’est intéressée à « l’Internet multilingue ». Des études portant sur un segment particulier du Web ont suivi, comme [4], pour les sites Web universitaires, ou [2], pour les sites Web de domaine de premier niveau de l’Union européenne. L’idée transversale à travers ces études est cohérente avec la notion que l’Internet, est, a été et sera encore de plus en plus multilingue. Un auteur a même soutenu, en 2019, que le Web avait dépassé l’étape du multilinguisme et avait évolué vers *l’hyperlinguisme* [5]. Plus récemment, des auteurs [9] ont plaidé pour une meilleure compréhension de l’utilisation multilingue sur Internet car elle a des implications pour des questions de linguistique appliquée telles que l’étude de l’hétéroglossie, l’apprentissage des langues, l’éducation linguistique et la politique linguistique.

Le présent article a pour but d'évaluer l'ampleur du biais de W3Techs, qui consiste à ignorer le fait que le Web est par nature, et de plus en plus, un domaine multilingue. Cependant, d'autres biais existent qui augmentent par erreur le pourcentage d'anglais et ils ont été analysés[[14]](#footnote-14).

# Matériels et méthodes

**2.1. Première et deuxième langues (L1, L2)**

Le nombre total de locuteurs de L1 est généralement calculé comme la population mondiale, en supposant que tous les humains n'ont qu'une seule langue maternelle et en l'attribuant aux bébés dès leur naissance. D'après l'ensemble de données mondiales d'Ethnologue #26 de mars 2023[[15]](#footnote-15), il y a 7,404 millions de locuteurs de L1 dans le monde, dont 380 millions sont des locuteurs en anglais, soit 5,13 % de la population mondiale. Ethnologue est généralement considéré comme le catalogue de langues le plus complet au monde[[16]](#footnote-16). Cependant, les chiffres démolinguistiques sont extrêmement difficiles à recueillir, et aucune source n’est considérée comme exempte d’erreurs dans ce domaine.

Une proportion importante de la population humaine parle plus d'une langue. La majorité des personnes sont monolingues et leur nombre de langues L2 est nul. Cependant, de nombreuses personnes sont bilingues et d'autres parlent 3 ou 4 langues. Le nombre de personnes parlant plus de 4 langues est assez faible[[17]](#footnote-17). Le nombre de personnes ayant plus de langues L2 diminue rapidement avec le nombre de langues, en raison de limitations évidentes, bien que la littérature mentionne qu'il existe des locuteurs de plus de 60 langues et que des personnes possédant plus de 200 langues ont existé[[18]](#footnote-18).

Le nombre total de locuteurs L1+L2 calculé par Ethnologue en 2023 est de 10 599 millions, ce qui implique que le taux de multilinguisme de l'humanité serait de 10 599 / 7 404 = 1,432. Ainsi, environ 40 % de la population mondiale parle plus d'une langue. Il convient de noter que le chiffre de 10 599 millions inclut la même personne autant de fois que cette personne parle des langues différentes. Bien qu'il existe des différences entre les sources sur les chiffres de L1, les différences sont beaucoup plus importantes pour les chiffres de L2. D'une part, la définition du niveau de maîtrise des langues à comptabiliser comme L2 n'est pas précise. D'autre part, les calculs sont fastidieux car les locuteurs de L2 sont répartis dans un plus grand nombre de pays. Parmi les locuteurs de L2, 1 078 millions sont anglophones selon l'ensemble de données Ethnologue #26. Le total mondial des locuteurs anglophones de L1+L2 serait donc de 380 + 1 078 = 1 458 millions. D'autres sources avancent le chiffre de 1 180 millions[[19]](#footnote-19)ou 1 500 millions[[20]](#footnote-20)et Crystal a exprimé la possibilité que ce chiffre tende à long terme vers 2 000 millions [7].

Il est courant de voir réalisé le calcul du pourcentage de locuteurs L1+L2 d'une langue en divisant le nombre de locuteurs L1+L2 par la population mondiale. Procéder ainsi nécessiterait d'avertir que le total des pourcentages pour toutes les langues dépasserait alors 100 %. Lorsque le total est forcé à 100 %, cela provoque une erreur cachée dans les chiffres pour le reste des langues. Pour maintenir un pourcentage réel, la division devrait être effectuée sur le nombre total de locuteurs L1+L2. Avec cette règle, le pourcentage mondial de locuteurs de l'anglais L1+L2 serait, d'après la source Ethnologue : 1 456 / 10 599 = 13,74 %. Même en utilisant la limite de 2 milliards de locuteurs anglais fixée par [7], le pourcentage total de personnes comprenant l'anglais comme première ou deuxième langue ne pourrait pas franchir le seuil des 20 %. En d'autres termes, un peu plus de 4 personnes sur 5 dans le monde ne comprennent pas l'anglais.

Si la tendance est que presque tous les humains auront accès à Internet, ce chiffre de 20 % devrait représenter une frontière difficile à franchir. L’anglais ne peut pas être la lingua franca de l’Internet s’il assure la communication entre moins de 20 % des internautes. C’était le cas aux premiers stades d’Internet, lorsqu’un pourcentage élevé d’internautes, universitaires, chercheurs et hommes d’affaires, avaient l’anglais comme L1 ou L2. Aujourd’hui, le pont est fait par la traduction, fortement assistée par des applications, et demain, c’est l’intelligence artificielle qui fera le pont.

* 1. **Les langues des pages Web**

En théorie, le calcul de la répartition des langues dans la Toile se base sur les pages Web et non sur les sites Web. La définition formelle du pourcentage de présence d'une langue sur le Web est : le nombre total de pages Web dans cette langue divisé par le nombre total de pages Web. Certaines pages Web sont multilingues, et la somme des pourcentages pour toutes les langues sera alors supérieure à 100 %.

Selon Netcraft[[21]](#footnote-21), il existe aujourd'hui plus de 1,2 milliard de sites Web, dont 200 millions sont actifs. Une source[[22]](#footnote-22) évalue le nombre total de pages Web à environ 50 milliards, dont moins de 10 % seraient indexées par les moteurs de recherche.

Le calcul de la présence des langues sur le Web en comptant les sites Web plutôt que les pages Web est donc une simplification compréhensible, mais la prudence est de mise. Cette méthode implique un processus assez similaire à celui décrit précédemment pour les humains. Les sites Web ont une langue L1 et certains ont des langues L2. En faisant une analogie avec les pourcentages de locuteurs L1+L2 par langue, les calculs corrects du pourcentage de langues dans un échantillon de sites Web doivent être effectués sur le nombre total de versions linguistiques de l'échantillon (et non sur le total des sites Web).

L'expérience montre que les sites Internet souffrent de moins de limitations que les humains et peuvent franchir plus facilement la limite de 4 langues. Wikipedia.org est aujourd'hui disponible en 331 langues, Facebook.com en 112, YouTube.com en 85 et Google.com en 87. En fait, de nombreux sites Internet connus dépassent les 100 langues tandis que de nombreux autres sites moins connus, notamment dans le domaine du commerce électronique, proposent des dizaines de versions linguistiques. Google Traduction[[23]](#footnote-23) permet, depuis juin 2024, de traduire dynamiquement des pages web en 244 langues, et est utilisé par certains sites web pour ouvrir, largement et gratuitement, leur couverture linguistique.

* 1. **Le taux de multilinguisme du Web**

La question du taux de multilinguisme de l'ensemble du Web est ouverte et il n'est pas facile d'y répondre. L'intuition et l'expérience nous disent que cet indicateur pourrait être plus élevé pour le Web que pour l'humanité (il serait alors supérieur à 1,43) car le facteur de limitation est clairement beaucoup plus faible. De toute évidence, de nombreux humains parlant différentes langues peuvent contribuer au contenu d'un même site Web. Les incitations à s'adresser à un plus grand nombre de personnes, en utilisant différentes langues sur le site Web, existent, en particulier dans le commerce électronique. Le taux de multilinguisme est donc un indicateur clé pour le Web. Des études ciblées ont été menées au niveau de l'Union européenne [2] mais aucune donnée mondiale n'existe à ce jour, et l'automatisation de cette analyse pour l'ensemble du Web n'est pas aisée.

Cependant, pour atteindre l'objectif d'évaluer la méthodologie de W3Techs, il suffit de se concentrer, plutôt que sur l'ensemble du Web, sur le même échantillonnage que celui utilisé par W3Techs pour sa mesure.

W3Techs utilise depuis de nombreuses années la liste des 10 millions de sites Web les plus visités, proposée par alexa.com (un fournisseur commercial), mais ce service a pris fin en mai 2022. W3Techs est alors passé à l'utilisation de la liste d'un million de sites Web les plus visités proposée par Tranco[[24]](#footnote-24), une organisation à but non lucratif. Tranco se présente comme « un classement des meilleurs sites axé sur la recherche et endurci contre la manipulation ». Le passage d’Alexa à Tranco pourrait d’ailleurs expliquer pourquoi les résultats de W3Techs pour 2023 sont sensiblement différents de ceux de 2022, avec une baisse notable de l’anglais et une forte croissance de nombreux suiveurs proches de l’anglais dans le classement…

Ayant à l’esprit l’évaluation de W3Techs, nous avons donc développé une approximation non automatique des données fournies par Tranco.

Nous avons analysé manuellement (navigation et recherche des options de langue du site) 7 séries de 100 sites web extraits au hasard de la liste Tranco et avons calculé les résultats. Notez que la liste Tranco est triée du plus visité au moins visité.

Le comptage a été effectué en parcourant chaque site web, un par un, en recherchant les options linguistiques et en prenant note dans le même fichier Excel, en ayant un site web par ligne, afin de faire un comptage final. Au cours de ce processus, les situations suivantes ont été identifiées :

* Dans la majorité des cas, les options linguistiques se trouvent en haut de la page d’accueil et le nombre de versions linguistiques est facilement compté.
* Dans certains cas, ils apparaissent en bas de page, avec la même facilité de comptage.
* Dans les cas moins fréquents, les options linguistiques sont implicites comme options de pays et nous avons vérifié, pour chaque pays, si chaque langue présentée a été effectivement traduite avant de procéder au comptage. De plus, nous avons évité de compter deux fois si la même langue est utilisée dans différents pays.
* Dans certains cas, l'étendue des langues autorisées par les sites Web est explicitement indiquée après avoir spécifié le pays.
* Certains sites web très importants et célèbres (comme Facebook et Google) préfèrent déduire automatiquement la langue à partir de l'appareil de l'utilisateur. Changer cela n'est pas une chose évidente car cela implique de rechercher de la page de configuration. Dans ce cas, nous avons recherché la page de configuration pour compter le nombre de langues proposées.

Les calculs ont été réalisés en créant un fichier Excel pour chaque échantillonnage avec les sites web en lignes et les paramètres comptés en colonnes. Sept fichiers ont été définis et remplis, correspondant à :

* L'ensemble des sites Web dans les 100 premières positions de l'échantillon Tranco
* L'ensemble des sites Web dans les 100 dernières positions de l'échantillon Tranco
* Cinq séries de 100 sites Web obtenues à l'aide de la fonction aléatoire d'Excel appliquée à différentes plages de l'échantillonnage Tranco comme mentionné dans le tableau 1.

Les résultats sont résumés dans le tableau 1 ci-dessous, où :

* M.Taux : est le taux de multilinguisme, c'est-à-dire le nombre total de versions linguistiques découvertes sur le nombre total de sites Web analysés ;
* Invalide : est le pourcentage de sites Web jugés invalides au cours du processus (différentes situations sont décrites ci-après)
* Mono : est le pourcentage de sites Web avec une langue unique
* Bi : est le pourcentage de sites Web avec 2 versions linguistiques
* Tri: est le pourcentage de sites Web avec 3 versions linguistiques
* Multi: est le pourcentage de sites Web avec plus de 3 versions linguistiques
* M.Moy : est le nombre moyen de versions linguistiques des sites Multi.

Nous avons toujours adopté une approche conservatrice, préférant laisser le nombre de M.Taux à 1 ou ne pas sélectionner un chiffre plus élevé, en cas de doute. Dans certains cas, les versions linguistiques sont générées automatiquement à l'aide de la page dynamique de Google Translate. Lorsque cela se produit, dans la plupart des cas, les 132 options linguistiques potentielles sont proposées. Cette approche affecte fortement positivement la moyenne de M.Taux et pourrait être considérée comme légitime en termes de multilinguisme. Cependant, conformément à l'approche conservatrice, nous avons toujours compté ces sites Web comme monolingues.

**Tableau1.** Résultats de l'analyse de sites Web.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **PREMIER**  **100** | **1000–10000** | **100000–1000000** | **1–1000000** | **1–1000000** | **1–1000000** | **DERNIER**  **100** | **MOYENNE** |
| **M.Taux** | 44.10 | 3.07 | 2.09 | 1,94 | 2.31 | 2,97 | 1,81 | 2.23 |
| **Invalide** | 23,0% | 19,0% | 25,0% | 33,0% | 25,0% | 26,0% | 25,0% | 27% |
| **Mono** | 18,2% | 67,9% | 72,0% | 82,1% | 70,7% | 78,4% | 84,0% | 77% |
| **Bi** | 6,5% | 11,1% | 17,3% | 7,5% | 16,0% | 5,4% | 5,3% | 10% |
| **Tri** | 2,6% | 2,5% | 2,7% | 3,0% | 4,0% | 4,1% | 4,0% | 4% |
| **Multi** | 74,0% | 19,8% | 9,3% | 9,0% | 10,7% | 13,5% | 8,0% | 10% |
| **M.Moy** | 59 | 11 | 10 | 10 | 11 | 15 | 9 | 11 |

L'analyse montre que les deux premières positions, pour les sites les plus visités de l'échantillon, ont un taux de multilinguisme extrêmement élevé : plus de 44 pour les cent premiers, et plus de 3 pour l'échantillon aléatoire entre 1 000 et 10 000. Conformément à l'approche conservatrice, nous avons écarté les premières colonnes pour le calcul de la moyenne et conservé uniquement les 6 colonnes de droite. Nous avons inclus la mesure des 100 derniers sites les moins visités du classement, où logiquement le M.Taux est le plus bas.

Ces mesures ne sont évidemment pas suffisantes pour appliquer des lois statistiques sur la distribution de ce que l'on pourrait considérer comme des variables aléatoires (les 7 éléments mesurés et représentés dans la première colonne). Elles ne représentent qu'un premier niveau d'approximation des données, à prendre avec précaution et dans un large intervalle de confiance.

Le résultat le plus stable semble être le nombre élevé de sites Web invalides, environ 25 %, ce qui montre que les sites Web naissent, peuvent tomber malades pendant un certain temps et finir par mourir.

Le taux moyen de multilinguisme de cet échantillon de sites Web semble être supérieur à 2, ce qui le rend supérieur au taux humain, ce qui n'est pas une surprise. Sur la base des résultats du tableau 1, la répartition typique des sites Web valides semble proche de :

* 75% des sites Web sont monolingues
* 10% des sites Web sont bilingues
* 4% des sites Web sont trilingues
* 10 % des sites Web ont plus de 3 langues, avec une moyenne autour de 11.

Ces résultats, bien que très approximatifs, aideront néanmoins à calculer le biais des valeurs fournies par W3Techs pour l'anglais puisque les résultats de W3Techs sont basés sur le même échantillon de Tranco. La question de savoir ce qui se passe pour les nombreux sites Web qui ne figurent pas dans la liste du million de sites les plus visités reste ouverte et est liée au biais de travailler avec les sites Web les plus visités.

Il est intéressant de comparer ces résultats avec ceux de [2], qui sont basés sur un échantillon plus large de sites Web (plus de 100 000). Ces sites Web sont limités aux ccTLD[[25]](#footnote-25) de l'Union européenne (comme .fr ou .uk) et non corrélées au plus grand nombre de visites. Les auteurs ont autorisé l'accès public aux données résultant de leur collecte[[26]](#footnote-26). Il existe un fichier par pays, reprenant tous les sites analysés et les différents comptages. Un autre fichier résume toutes les données collectées par pays et les totaux.

Il est possible de calculer dans chaque fichier pays le nombre moyen de versions linguistiques pour les sites Web comportant plus de 2 langues et de reporter la valeur dans le fichier récapitulatif. Des calculs du taux de multilinguisme ainsi que du pourcentage de sites Web en anglais sont ensuite effectués. Les résultats sont présentés dans les tableaux 2 et 3, où les données grises sont le résultat de nos calculs tandis que le reste des données est tiré directement de la source [2].

* Population : nombre de résidents dans le pays
* Sites Web : nombre de sites Web analysés dans l'échantillon
* % Mono : pourcentage de sites Web jugés monolingues
* % Bi : pourcentage de sites Web jugés bilingues
* % Multi : pourcentage de sites Web avec plus de 2 versions linguistiques
* M.Moy : nombre moyen de versions de sites avec plus de 2 versions linguistiques
* M.Taux : Taux de multilinguisme
* Domaines : nombre de domaines en dessous du ccTLD[[27]](#footnote-27) pour tous les pays, sauf la Lettonie qui manquait et a été complétée par une autre source[[28]](#footnote-28).

**Tableau 2.** Répartition des sites Internet de l'UE par type de versions linguistiques analysées par [2]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pays** | **Population** | **Sites Web** | **% Mono** | **% Bi** | **% Multi** | **M.Moy** | **M.Taux** | **Domaines** |
| Autriche | 8,999,973 | 4,063 | 87,96 | 11h35 | 0,69 | 3,50 | 1.131 | 1,289,274 |
| Belgique | 11,579,502 | 4,063 | 81,20 | 15,92 | 2,88 | 3.34 | 1.227 | 532,005 |
| Bulgarie | 6,964,301 | 3,178 | 69,26 | 28.51 | 2.23 | 3.70 | 1.346 | 228,272 |
| Croatie | 4,112,131 | 3,489 | 69,30 | 27,66 | 3.04 | 3,85 | 1.363 | 137,901 |
| Chypre | 1,186,194 | 828 | 66,91 | 30,80 | 2.29 | 5.11 | 1.402 | 220,947 |
| Tchéquie | 10,705,712 | 4,084 | 86,46 | 12.34 | 1.20 | 3.76 | 1.156 | 1,510,721 |
| Danemark | 5,785,741 | 4,067 | 84,98 | 14.36 | 0,66 | 3.19 | 1.158 | 1,472,212 |
| Estonie | 1,327,561 | 3,556 | 67,32 | 23.06 | 9.62 | 3,50 | 1.471 | 215,250 |
| Finlande | 5,538,872 | 3,992 | 81,96 | 15,68 | 2.35 | 3.70 | 1.220 | 742,867 |
| France | 65,227,357 | 4,125 | 90,86 | 8.34 | 0,80 | 6.42 | 1.127 | 7,447,877 |
| Allemagne | 83,792,987 | 4,150 | 90,67 | 9.08 | 0,24 | 4.10 | 1.098 | 6,604,705 |
| Grèce | 10,439,436 | 3,953 | 68.13 | 29,65 | 2.23 | 3.70 | 1.357 | 174,018 |
| Hongrie | 9,668,737 | 3,993 | 84,22 | 14,70 | 1.08 | 3.30 | 1.172 | 289,796 |
| Irlande | 4,927,661 | 3,825 | 98,72 | 1.23 | 0,05 | 3,50 | 1.014 | 265,024 |
| Italie | 60,496,082 | 4,123 | 82,61 | 15.18 | 2.21 | 4.10 | 1.220 | 3,817,443 |
| Lettonie | 1,891,687 | 3,406 | 59,54 | 30,65 | 9.81 | 3.25 | 1.527 | 136,718 |
| Lituanie | 1,954,244 | 3,773 | 73,63 | 20,94 | 5.43 | 3.38 | 1.339 | 40,430 |
| Luxembourg | 623,897 | 2,876 | 72,95 | 21,73 | 5.32 | 3.72 | 1.362 | 192,571 |
| Malte | 441,161 | 444 | 95,72 | 3.15 | 1.13 | 6.60 | 1.095 | 66,264 |
| Pays-Bas | 17,123,478 | 4,133 | 87.10 | 12.10 | 0,80 | 3.39 | 1.140 | 4,340,730 |
| Pologne | 37,864,109 | 4,110 | 89,44 | 9,68 | 0,88 | 3.64 | 1.120 | 573,641 |
| Portugal | 10,205,235 | 4,084 | 74,34 | 16.67 | 2.25 | 3,80 | 1.163 | 299,126 |
| Roumanie | 19,266,079 | 3,975 | 70,36 | 28,48 | 1.16 | 3.57 | 1.314 | 330,703 |
| Slovaquie | 5,459,814 | 3,943 | 83.11 | 15.32 | 1,57 | 3.34 | 1.190 | 444,701 |
| Slovénie | 2,079,226 | 3,619 | 76.02 | 21.08 | 2,90 | 3.60 | 1.286 | 49,558 |
| Espagne | 46,767,543 | 4,088 | 86,89 | 10,96 | 2.15 | 3.44 | 1.162 | 2,172,046 |
| Suède | 10,081,948 | 4,084 | 85,70 | 13,96 | 0,34 | 3,00 | 1.146 | 961,089 |
| Royaume-Uni | 67,803,450 | 4,125 | 99,59 | 0,27 | 0,15 | 7.83 | 1.013 | 7,148,183 |

Note: L'étude a été réalisée avant le Brexit.

**Tableau 3.** Répartition des sites Internet de l'UE en anglais analysée par [2]

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pays** | **Sites Web** | **Mono** | **Bi** | **Multi** | **Mono Eng.** | **Bi Eng.** | **Multi-ingénierie** | **Total**  **Anglais** | **Versions totales** | **% Anglais** |
| Autriche | 4,063 | 3,574 | 461 | 28 | 68 | 454 | 28 | 550 | 4,594 | 12,0% |
| Belgique | 4,063 | 3,299 | 647 | 117 | 365 | 553 | 112 | 1,030 | 4,984 | 20,7% |
| Bulgarie | 3,178 | 2,201 | 906 | 71 | 139 | 891 | 71 | 1,101 | 4,276 | 25,7% |
| Croatie | 3,489 | 2,418 | 965 | 106 | 143 | 940 | 104 | 1,187 | 4,756 | 25,0% |
| Chypre | 828 | 554 | 255 | 19 | 421 | 255 | 19 | 695 | 1,161 | 59,9% |
| Tchéquie | 4,084 | 3,531 | 504 | 49 | 69 | 474 | 46 | 589 | 4,723 | 12,5% |
| Danemark | 4,067 | 3,456 | 584 | 27 | 218 | 577 | 26 | 821 | 4,710 | 17,4% |
| Estonie | 3,556 | 2,394 | 820 | 342 | 158 | 656 | 330 | 1,144 | 5,231 | 21,9% |
| Finlande | 3,992 | 3,272 | 626 | 94 | 157 | 578 | 94 | 829 | 4,872 | 17,0% |
| France | 4,125 | 3,748 | 344 | 33 | 46 | 335 | 31 | 412 | 4,648 | 8,9% |
| Allemagne | 4,150 | 3,763 | 377 | 10 | 62 | 370 | 10 | 442 | 4,547 | 9,7% |
| Grèce | 3,953 | 2,693 | 1,172 | 88 | 369 | 1,165 | 86 | 1,620 | 5,301 | 30,6% |
| Hongrie | 3,993 | 3,363 | 587 | 43 | 55 | 564 | 42 | 661 | 4,679 | 14,1% |
| Irlande | 3,825 | 3,776 | 47 | 2 | 3,763 | 47 | 2 | 3,812 | 3,876 | 98,3% |
| Italie | 4,123 | 3,406 | 626 | 91 | 88 | 595 | 89 | 772 | 5,031 | 15,3% |
| Lettonie | 3,406 | 2,028 | 1,044 | 334 | 188 | 817 | 331 | 1,336 | 5,202 | 25,7% |
| Lituanie | 3,773 | 2,778 | 790 | 205 | 121 | 747 | 196 | 1,064 | 5,051 | 21,1% |
| Luxembourg | 2,876 | 2,098 | 625 | 153 | 375 | 497 | 144 | 1,016 | 3,917 | 25,9% |
| Malte | 444 | 425 | 14 | 5 | 416 | 14 | 5 | 435 | 486 | 89,5% |
| Pays-Bas | 4,133 | 3,600 | 500 | 33 | 143 | 493 | 33 | 669 | 4,712 | 14,2% |
| Pologne | 4,110 | 3,676 | 398 | 36 | 47 | 387 | 34 | 468 | 4,603 | 10,2% |
| Portugal | 4,084 | 3,036 | 681 | 92 | 136 | 646 | 89 | 871 | 4,748 | 18,3% |
| Roumanie | 3,975 | 2,797 | 1,132 | 46 | 311 | 1,104 | 45 | 1,460 | 5199 | 28,1% |
| Slovaquie | 3,943 | 3,277 | 604 | 62 | 70 | 544 | 59 | 673 | 4,671 | 14,4% |
| Slovénie | 3,619 | 2,751 | 763 | 105 | 143 | 722 | 96 | 961 | 4,655 | 20,6% |
| Espagne | 4,088 | 3,552 | 448 | 88 | 91 | 392 | 81 | 564 | 4,751 | 11,9% |
| Suède | 4,084 | 3,500 | 570 | 14 | 230 | 540 | 12 | 782 | 4,682 | 16,7% |
| Royaume-Uni  I | 4,125 | 4,108 | 11 | 6 | 4094 | 11 | 6 | 4,111 | 4,177 | 98,4% |

À partir de ces deux tableaux, les résultats moyens pour toutes les nations de l'Union européenne sont calculés et résumés dans le tableau 4. Notez qu'il existe différentes manières de calculer la moyenne : moyenne simple, pondérée par la population du pays, pondérée par le nombre de sites explorés, et la dernière, probablement la plus appropriée pour obtenir un chiffre global pour l'Union européenne, pondérée par le nombre de domaines sous le ccTLD.

**Tableau 4.** Résultats moyens pour l'ensemble de l'Union européenne.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **% mono** | **% bi** | **% multi** | **M. Moy.** | **M.Taux** | **% Anglais** |
| Moyenne simple | 80,9 | 16,5 | 2.3 | 3.9 | 1.225 | 24,21% |
| Moyenne pondérée par la population du pays | 87,3 | 11.4 | 1.2 | 4.4 | 1.147 | 26,04% |
| Moyenne pondérée par le nombre de sites explorés | 81,3 | 16.2 | 2.3 | 3.7 | 1.216 | 24,81% |
| Moyenne pondérée par le nombre de domaines | 86,3 | 9.6 | 1.0 | 4.6 | 1.095 | 28,42% |

La dernière ligne des résultats du tableau 4 nous permet d'affirmer que, sur la base de l'échantillon établi par[2], 86 % des sites ccTLD de l'Union européenne sont monolingues, 10 % sont bilingues et 1 % ont plus de 2 versions linguistiques, avec une moyenne de 4,6. En suivant cette approche, 28,4 % des versions linguistiques des sites Web de l'Union européenne en 2020 étaient en anglais (ce pourcentage est obtenu en divisant le nombre total de versions linguistiques de sites Web en anglais, qu'il s'agisse d'un site Web monolingue en anglais ou d'une version anglaise d'un site Web multilingue, divisé par le nombre total de versions linguistiques).

Les différences assez importantes dans le taux de multilinguisme (M.Taux) entre les quatre modes de calcul de la moyenne présentés dans le tableau 4 pourraient indiquer que, en dehors de la liste Tranco du million de sites les plus visités, le taux de multilinguisme du Web n'est pas si élevé. Il se pourrait également, et beaucoup plus probablement, que le taux de multilinguisme des sites Web ccTLD soit logiquement inférieur à celui des sites Web à vocation mondiale (qui sont souvent dans des domaines non nationaux comme .com). On pourrait également s'attendre à ce que la présence de versions linguistiques en anglais augmente fortement dans les sites non ccTLD en même temps que le taux de multilinguisme. Une chose compensant l'autre, le pourcentage final d'anglais pourrait rester dans la fourchette de 20 à 30 %.

En ce qui concerne le taux global de multilinguisme du Web, il est à noter que les enregistrements de ccTLD ne représentent ensemble que 38 % du total des enregistrements de domaines. Cependant, souvent, même si ce n'est pas dans la région européenne, ils sont détournés de leur définition d'origine, associée à un pays, vers des propositions commerciales mondiales. Le meilleur exemple est celui du ccTLD .tk, des îles Tokelau, qui est le premier en termes d'enregistrements, même plus que le .cn pour la Chine. Évidemment, la plupart des sites Web .tk ne contiennent pas d'informations nationales sur Tokelau (voir sources[[29]](#footnote-29)). En fait, un coup d'œil au tableau des ccTLD dans l'article Wikipédia correspondant[[30]](#footnote-30) révèle qu'une grande partie d'entre eux sont désormais utilisés comme « hackers de domaine »[[31]](#footnote-31), un usage détourné de la définition originale propre à chaque pays. Les piratages de domaines sont généralement liés à une activité de commerce électronique, beaucoup plus enclins que les sites Web nationaux au multilinguisme.

En tout état de cause, les données de [2] ne modifient pas la discussion sur les biais de W3Techs qui sont liés à un échantillon différent, la liste Tranco. Les considérations sur les ccTLD de l'Union européenne, qui présentent un taux de multilinguisme relativement faible, ne peuvent pas non plus être généralisées à l'ensemble du Web, car les ccTLD sont par nature moins enclins au multilinguisme que les domaines génériques.

* 1. **La méthodologie W3Techs**

Selon la méthodologie W3Techs[[32]](#footnote-32), un algorithme de reconnaissance de langue est appliqué chaque jour à la page d'accueil des sites web de la liste Tranco. Cet algorithme identifie une langue unique pour chaque site web et procède à son comptage. Cette méthode implique que les sites web multilingues avec une version anglaise sont probablement comptabilisés uniquement comme anglais. Enfin, les pourcentages par langue sont calculés sur le total des sites web mesurés et sont appliqués pour les 40 premières langues (les autres langues étant mentionnées comme ayant moins de 0,1% du total).

Il est clair que la méthode considère le Web comme un espace monolingue, ce qui pourrait représenter un biais sérieux au vu des chiffres calculés ci-dessus pour le taux de multilinguisme du Web. Est-il possible de corriger ce biais, au moins pour le résultat de l'anglais ?

# Résultats

Il y a deux éléments à prendre en compte pour dé-biaiser les résultats de W3Techs :

* Pourcentage de sites Web non anglophones calculés par erreur comme étant en anglais. Cela peut se produire parce qu'il y a un certain nombre de mots anglais sur la page d'accueil à côté de l'autre langue principale. Cela peut également se produire parce qu'il s'agit d'un site Web non valide qui n'est pas détecté par l'algorithme en tant que tel et traité comme étant en anglais. Cela peut se produire parce que le message résultant de l'invalidité est en anglais.
* L'impact d'un taux de multilinguisme supérieur à un sur le calcul du pourcentage.

L'équation correctrice est : P' = (P-Err) / M.Taux, où :

* P' est le pourcentage non biaisé pour le contenu en anglais
* P est le pourcentage de sortie pour les contenus en anglais fournis par W3Techs
* Err est le pourcentage de sites Web calculés par erreur en anglais
* M.Taux est le taux de multilinguisme de l'échantillonnage.

L'analyse manuelle des 700 sites Web a montré qu'un certain pourcentage (environ 10 %) sont des sites Web non anglophones avec de nombreux mots anglais sur la page d'accueil. Sur les 27 % de sites Web jugés invalides, moins de 50 % étaient clairement identifiables comme tels (codes retournés 404, non trouvé, ou 403, accès interdit). La plupart du temps, le message invalide provenait du serveur hôte ou du gestionnaire de noms de domaine, et il était en anglais. Dans plus de 10 % des cas, il s'agissait simplement d'une courte phrase en anglais ou d'un message « site en construction ». Ne pas connaître les détails de l'algorithme rend impossible de déterminer si ces situations sont traitées comme non comptabilisées ou comptabilisées comme anglaises. Pour les besoins de l'équation, nous utiliserons un chiffre conservateur entre 5 % et 15 % (par exemple entre ¼ et ½ de ce chiffre de 27 %) comme pourcentage de sites Web comptabilisés à tort comme anglais.

La valeur moyenne de M.Taux est de 2,23 dans notre échantillon (tableau 1), fixer une valeur entre 1,5 et 2,5 semble raisonnable. Le tableau 5 propose alors trois scénarios pour les valeurs M.Taux et Err et utilise R= 56,1%, la valeur du 23/3/2023 pour les contenus en anglais du site W3Techs :

**Tableau 5.** Résultats de W3Techs avec correction des biais.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Faible** | **Moyen** | **Haut** |
| M.Taux de M | 1,5 | 2 | 2.5 |
| Err tromper | 5% | 10% | 15% |
| R' | 34,1% | 23,1% | 16,4% |

Il est assez intéressant de noter que le chiffre moyen pour R' est très proche des résultats d'OBDILCI[[33]](#footnote-33), l’un des acteurs du secteur avant 2009. OBDILCI a proposé une nouvelle méthode en 2017 et a publié ses résultats en 2022 pour 342 langues, donnant des contenus anglais autour de 20%, à l’instar des contenus chinois, et affichant un constat bien différent de celui de Statista : « *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*. »

Contrairement à W3Techs, la méthodologie utilisée pour les résultats OBDILCI a été révélée dans tous les détails et revue par des pairs [8] ainsi que les résultats produits [9]. Les indicateurs produits par OBDILCI sont désormais accessibles au public en libre accès (cc-by-sa 4.0) sous la forme d'une base de données dont la clef est le code ISO 639-3[[34]](#footnote-34). Les biais de la méthode d'OBDILCI existent et sont largement discutés dans [8]. Leur ampleur explique pourquoi les résultats sont présentés avec un large intervalle de confiance de –20% à +20%.

Quant aux 39 autres langues pour lesquelles W3Techs a donné des pourcentages, le manque de prise en compte du multilinguisme sur le web empêche de corriger les chiffres de W3Techs. Ils doivent être considérés comme peu fiables. En effet, le pourcentage pour les langues non anglaises présenté par W3Techs indique le pourcentage de sites web de la langue mentionnée sans version anglaise.

# Discussion

L'étude publiée dans [2] constitue un premier intérêt académique bienvenu, depuis longtemps, sur le sujet des langues sur le Web. Ce sujet a été laissé trop longtemps à très peu d'entités commerciales et d'organisations à but non lucratif. Cette situation a limité les discussions académiques sur un sujet d'intérêt général croissant, avec des implications bien au-delà de la linguistique, comme le commerce électronique, les politiques publiques, la géopolitique et la cyber-géographie [10].

L'algorithme programmé de [2] accorde une attention particulière au multilinguisme en vérifiant la langue de tous les liens internes du site Web analysé. Il pourrait être réutilisé, modifié et appliqué à la liste des sites Web de Tranco (au lieu de la liste des sites Web de l'Union européenne). Cela fournirait une preuve encore plus solide de notre conclusion selon laquelle si la nature multilingue du Web était prise en compte, les résultats de W3Techs pour l'anglais glisseraient de la fourchette de 55 à 65 % à la fourchette de 20 à 30 %.

Des tentatives de partenariat ont été faites dans cette optique, mais il semble que le département des arts audio et visuels de l'université ionienne ait logiquement d'autres priorités. Dans ce contexte, nous avons décidé de planifier la réalisation d'un projet de mesure directe des langues dans un échantillon de sites Web, en utilisant la détection de langue. Afin d'éviter le lourd investissement associé à l'exploration de millions de sites, nous prévoyons d'appliquer une approche statistique traditionnelle en sélectionnant au hasard, par exemple, une centaine de séries de 1 000 sites Web et en dérivant des résultats à partir de la distribution statistique des résultats.

Nous sommes en phase de préparation, testant différents algorithmes de détection de langues, réfléchissant à la meilleure façon de gérer le défi de la détection du multilinguisme et recherchant des financements. Dans le cadre de cette préparation, nous avons mis en place un nouvel échantillonnage de 5 ensembles de 100 sites Web sélectionnés au hasard sur Tranco qui sont utilisés pour les tests. Cela fournit un autre ensemble de résultats que nous laissons en libre accès[[35]](#footnote-35) et remplace les séries précédentes qui ont été perdues lors d'une panne informatique. Les nouveaux résultats, présentés dans le tableau 6, restent dans la même gamme que la première série, avec un M.Taux de 1,93 (au lieu de 2,23) et un pourcentage mesuré d'anglais de 29 % pour cet échantillon.

**Tableau 6.** Nouvel ensemble de mesures.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Série**  **1** | **Série**  **2** | **Série**  **3** | **Série**  **4** | **Série**  **5** | **Moy.** |
| **M.Taux** | 2.51 | 1,73 | 2.07 | 1,65 | 1,69 | 1,93 |
| **Invalide** | 31% | 30% | 28% | 28% | 28% | 29% |
| **Multi** | 28% | 16% | 25% | 19% | 18% | 21% |
| **M.Moy** | 5.47 | 5.10 | 4.53 | 3.36 | 3,85 | 4.46 |
| **Anglais %** | 23% | 30% | 29% | 30% | 34% | 29% |

Le projet, prévu pour 2025, servira à compléter les outils disponibles dans OBDILCI. Il permettra de réaliser des études spécifiques à certains sous-ensembles du Web (comme par exemple les ccTLD des pays lusophones) et, bien évidemment, il sera également appliqué à l'échantillon Tranco afin d'obtenir une correction définitive des résultats de W3Techs.

# Conclusions

Les preuves du multilinguisme croissant et déjà élevé du Web, notamment dans sa composante e-commerce, se multiplient. Il est de plus en plus évident que la lingua franca du Web n’est pas l’anglais mais la traduction, un processus de plus en plus assisté par l’informatique, avec des perspectives prometteuses offertes par la nouvelle génération d’outils d’intelligence artificielle.

Les chercheurs doivent faire preuve de prudence lorsqu’ils utilisent des sources commerciales qui ne sont pas appuyées par des publications scientifiques et des évaluations par les pairs, en particulier dans les domaines à la croisée des chemins entre la démolinguistique et les données du Web, où les intervalles de confiance sont assez larges.

Quant à la propriété historiquement unique et multilingue du Web, les preuves sont là, au coin de la rue virtuelle, et s’imposeront progressivement dans les années à venir, l’intelligence artificielle devenant un catalyseur et un amplificateur évident d’un multilinguisme en plein essor dans le domaine numérique.

L'activité prévue par OBDILCI, qui consiste à proposer une méthode alternative d'exploration d'un ensemble de sites Web pour la détection et le comptage des langues, en tenant compte du multilinguisme des sites Web, pourrait représenter une avancée dans l'étude des langues dans l’Internet. En plus d'offrir un large éventail de possibilités pour étudier la distribution des langues dans des segments ciblés du Web, en complément de sa méthode générale englobant l'ensemble du Web, elle pourrait apporter une réponse définitive à la question controversée de la réalité de la domination de l'anglais dans le monde numérique.

**Financement**

Cette recherche a été financée par la Direction Générale pour le français et les langues de France du ministère de la culture de la France.

**Déclaration du comité d'examen institutionnel**

Non applicable

**Déclaration de consentement éclairé**

Non applicable

**Déclaration de disponibilité des données**

Les fichiers Excel du dernier échantillonnage de 5 ensembles de 100 sites Web sélectionnés au hasard dans la liste Tranco sont disponibles à l'adresse suivante : https://www.obdilci.org/wp-content/uploads/2024/08/TRANCO-Sampling.zip

**Conflit d'intérêts**

L'auteur déclare n'avoir aucun conflit d'intérêt.

# Références

Callahan, E., Herring, S.C., 2012. Language choice on university websites: Longitudinal trends. International Journal of Communication. 6, 322–355. Available from: <http://ijoc.org/ojs/index.php/ijoc/article/view/1451/703>

Crystal, D., 2008. Two thousand million? English Today. 24(1), 3–6. Available from: <https://www.cambridge.org/core/journals/english-today/article/two-thousand-million/68BFD87E5C867F7C3C47FD0749C7D417>

Giannakoulopoulos, A., Pergantis, M., Konstantinou, N., et al., 2020. Exploring the dominance of the English language on the websites of EU countries. Future Internet. 12(4), 76– DOI: <https://doi.org/10.3390/fi12040076>

1. Kelly-Holmes, H., 2019. Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism. Annual Review of Applied Linguistics. 39, 24–39. DOI: https://doi.org/[10.1017/S0267190519000102](https://doi.org/10.1017/S0267190519000102)
2. Pimienta, D., Blanco, A., Müller de Oliveira, G., 2023. The method behind the unprecedented production of indicators of the presence of languages in the Internet, Frontiers in Research Metrics and Analytics. 8, DOI: https://doi.org/[10.3389/frma.2023.1149347](https://doi.org/10.3389/frma.2023.1149347)

Pimienta, D., 2022. Resource: Indicators on the presence of languages in Internet. Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages; 24-25 June 2022; Marseille, France. pp. 83–91. Available from: <https://aclanthology.org/2022.sigul-1.11/>

1. Pimienta, D., Müller de Oliveira, G., 2022. Cyber-geography of languages: Part 1: Method, results and focus on English; Part 2: The demographic factor and the growth of Asian languages and Arabic. The International Review of Information Ethics. Available from: <https://informationethics.ca/index.php/irie/article/view/488> <https://informationethics.ca/index.php/irie/article/view/491>

Danet, B., Herring, S.C., 2007. The multilingual internet: Language, culture, and communication online. Oxford University Press: New York, USA. Available from: <https://academic.oup.com/book/32471>

1. Leppänen, S., Peuronen, S., 2020. Multilingualism and the Internet. Wiley Online Library. DOI: https://doi.org/ [10.1002/9781405198431.wbeal0805.pub2](https://doi.org/10.1002/9781405198431.wbeal0805.pub2)

Pimienta D., Prado D., Blanco A., 2009. Twelve years of measuring linguistic diversity on the Internet: Balance and perspectives. UNESCO publications for the World Summit on the Information Society. CI-2009/WS/1. Available from: <http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016>

1. <https://www.emarketer.com/content/worldwide-ecommerce-sales-break-6-trillion> [↑](#footnote-ref-1)
2. Comme <https://motsdici.be/wp-content/uploads/2019/04/Article-ne-peux-pas-lire-ne-veux-pas-acheter.pdf> ou <https://www.t-works.eu/en/e-commerce-growth-and-the-importance-of-lingual/> [↑](#footnote-ref-2)
3. Chiffres issus du travail de <https://unicode.org> [↑](#footnote-ref-3)
4. <https://statista.com>, une société spécialisée dans la fourniture de statistiques. [↑](#footnote-ref-4)
5. [https://www.statista.com/chart/26884/langues-sur-internet/](https://www.statista.com/chart/26884/languages-on-the-internet/) [↑](#footnote-ref-5)
6. <https://w3techs.com>, une société spécialisée dans les enquêtes sur les technologies web [↑](#footnote-ref-6)
7. [https://w3techs.com/technologies/overview/content\_langue](https://w3techs.com/technologies/overview/content_language) [↑](#footnote-ref-7)
8. [https://w3techs.com/technologies/history\_overview/content\_lingual/ms/y](https://w3techs.com/technologies/history_overview/content_language/ms/y) [↑](#footnote-ref-8)
9. <https://en.wikipedia.org/wiki/Languages_used_on_the_Internet> [↑](#footnote-ref-9)
10. <https://w3techs.com/technologies> [↑](#footnote-ref-10)
11. https://obdilci.org [↑](#footnote-ref-11)
12. https://www.obdilci.org/projects/main/ [↑](#footnote-ref-12)
13. [https://www.netsweeper.com/government/top-linguals-commonly-used-interneto](https://www.netsweeper.com/government/top-languages-commonly-used-interneto) [↑](#footnote-ref-13)
14. https://www.obdilci.org/projects/main/englishweb/ [↑](#footnote-ref-14)
15. <https://www.ethnologue.com/Ethnologue-26-Global-Dataset-Doc.pdf> [↑](#footnote-ref-15)
16. https://en.wikipedia.org/wiki/Ethnologue [↑](#footnote-ref-16)
17. <https://en.wikipedia.org/wiki/List_of_polyglots> [↑](#footnote-ref-17)
18. [https://lawlinguists.com/fr/record-linguals-spoken-one-person/](https://lawlinguists.com/fr/record-languages-spoken-one-person/) [↑](#footnote-ref-18)
19. [https://en.wikipedia.org/wiki/List\_of\_countries\_by\_English- Speaking\_population](https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population) [↑](#footnote-ref-19)
20. https://www.statista.com/chart/26884/langues-sur-internet/ [↑](#footnote-ref-20)
21. <https://news.netcraft.com/archives/category/web-server-survey> [↑](#footnote-ref-21)
22. <https://www.worldwidewebsize.com> [↑](#footnote-ref-22)
23. <https://translate.google.com/?op=websites> [↑](#footnote-ref-23)
24. <https://tranco-list.eu> [↑](#footnote-ref-24)
25. Domaine de premier niveau avec code de pays [↑](#footnote-ref-25)
26. https://zenodo.org/record/3698008 [↑](#footnote-ref-26)
27. <https://domainnamestat.com/statistics/country/others> [↑](#footnote-ref-27)
28. <https://www.nic.lv/en/look-back-at-the-lv-in-2021> [↑](#footnote-ref-28)
29. <https://www.verisign.com/assets/domain-name-report-Q42019.pdf> et [https://www.verisign.com/en\_US/domain-names/dnib/index.xhtml?section =informations-supplémentaires](https://www.verisign.com/en_US/domain-names/dnib/index.xhtml?section%20=informations-supplémentaires) [↑](#footnote-ref-29)
30. <https://en.wikipedia.org/wiki/Country_code_top-level_domain> [↑](#footnote-ref-30)
31. <https://en.wikipedia.org/wiki/Domain_hack> [↑](#footnote-ref-31)
32. <https://w3techs.com/technologies> [↑](#footnote-ref-32)
33. <https://obdilci.org> [↑](#footnote-ref-33)
34. https://obdilci.org/Base [↑](#footnote-ref-34)
35. https://www.obdilci.org/wp-content/uploads/2024/08/TRANCO-Sampling.zip [↑](#footnote-ref-35)