Inventaire et comparaisons de toutes les méthodes de mesure des langues en ligne : implications pour la lingua franca d'Internet

Daniel Pimienta, OBDILCI, 10/2024, mise à jour 7/2025
pimienta@funredes.org

Ce rapport est une contribution à la Session 5 : Communication multilingue : traduction basée sur l'IA ou lingua franca ? à la deuxième Conférence internationale sur les technologies linguistiques pour tous (LT4ALL 2025) - Siège de l'UNESCO, Paris, France, 24 février 2025. L'auteur est membre du Comité de programme dans la section Politique.
Ce qui suit est une traduction de l'article originale en anglais avec des rajouts concernant des informations survenues après la réunion.

SYNTHESE 2
INTRODUCTION 2
I – APPROCHES ACTUELLES DE MESURE DES LANGUES EN LIGNE 2
APPROCHE 1 : W3TECHS 2
APPROCHE 2 : DATAPROVIDER.COM 3
APPROCHE 3 : NETSWEEPER 4
APPROCHE 4 : UNIVERSITÉ IONIENNE GRÈCE 5
APPROCHE 5 : MÉTHODE PRINCIPALE OBDILCI 6
APPROCHE 6 : ÉTUDES PRÉALABLES MECILDI@ OBDILCI 7
APPROCHE 7 : MECILDI@OBDILCII 7
SYNTHÈSE 8
II – COMPARAISONS ET DISCUSSION 9
2.1 – COMPARAISONS DES RÉSULTATS POUR L'ANGLAIS 9
2.2 - COMPARAISON DES PREMIÈRES LANGUES 10
2.3 QUE NOUS APPRENNENT CES COMPARAISONS ? 10
III INITIATIVES DE LA PREMIÈRE PÉRIODE (1996-2014) 11
Étude Xerox (1996-2000) 11
OBDILCI/Funredes (1998-2007) 11
ISOC Québec/Alis Technologies, suivi d'OCLC (1997, 1999, 2002) 12
INKTOMI (2000) 12
Google : Méthode du complément de l'espace vide (1988-2008) 12
Projet d'Observatoire des Langues – LOP (2003-2011) 13
UPC/IDESCAT (2003-2006) 13
DILINET/SEMACORE (2010-2014) 13
IV Évolution des indicateurs linguistiques clés de l'Internet 13
V CONCLUSION 16
IV RÉFÉRENCES 17

SYNTHESE

Sept approches existantes pour la mesure de la proportion des langues en ligne sont résumées avec les paramètres associés (institution, nombre de langues couvertes, méthode, mises à jour, évaluation par les pairs). L'analyse des biais amène à la conclusion d'une présence de l'anglais en termes de pourcentage de pages dans la Toile, entre 20% et 27 % et que l'Internet aujourd'hui se caractérise par le facteur de multilinguisme. La présence en ligne d'environ 750 langues permet à plus de 95 % de la population mondiale d'interagir dans le monde numérique en L1 ou L2. Cependant, il ne s'agit que de 10% des langues encore en vie ; le défi reste gigantesque pour assurer à toutes les langues le respect de leur droit d'être présentes dans le monde numérique. Le taux de multilinguisme du www, un paramètre clé encore inconnu à ce jour, pourrait avoir dépassé son équivalent humain (1,44). La lingua franca d'Internet est désormais la traduction boostée par l'IA.

INTRODUCTION

Nous exposons toutes les méthodes existantes et historiques de mesure de la proportion des langues dans la Toile. En comparant les méthodologies, les résultats démontrent que l'anglais représente désormais environ 20 à 27 % du contenu web, contredisant les affirmations répandues, mais inexactes, selon lesquelles il dépasserait les 50 %. Cette étude identifie et met en évidence les biais méthodologiques qui impactent ces mesures, notamment ceux liés à la non prise en compte du multilinguisme existant dans le monde des sites web, et traite de leurs implications pour l'évaluation de la diversité linguistique sur Internet.

Cinq approches actuelles ont été identifiées, émanant d'entreprises, d'universités et d'organisations de la société civile, qui proposent actuellement des chiffres sur la proportion de langues en ligne. Ce document les expose et tente de tirer des conclusions à partir de leurs similitudes, de leurs différences et de leurs biais potentiels. Ces méthodes sont comparées afin d'identifier les similitudes et les différences dans leur produit et les biais inhérents à chacune d'entre elles. De plus, les méthodes et les résultats obtenus dans la période initiale de la Toile (1997-2013) sont brièvement passés en revue, des analyses plus détaillées étant disponibles dans des publications antérieures (Pimienta et al., 2009).

L'analyse montre qu'en 2007, l'anglais représentait environ 50 % du contenu web. Aujourd'hui, des données concordantes indiquent un chiffre plus proche de 25 %, même si le discours public mésinforme à propos de cet indicateur. Si d'autres méthodes existantes n'ont pas été incluses dans cette étude, l'auteur invite leurs auteurs à le contacter pour une collaboration ou une inclusion dans de futures éditions.

I – APPROCHES ACTUELLES DE MESURE DES LANGUES EN LIGNE

APPROCHE 1 : W3TECHS

Source : https://w3techs.com/technologies/overview/content_language
Méthode: Application quotidienne d'un algorithme de détection de langue sur le million de sites Web les plus visités, répertoriés par TRANCO.
Institution: Société de services Internet spécialisée dans des études et sondages à propos des technologies du Web
Portée: Quotidien depuis 2011

Méthodologie exposée : partiellement (https://w3techs.com/technologies)
Méthodologie évaluée par les pairs : Non
Discussion sur les biais : Non
Intervalle de confiance des résultats : Non disponible
Nombre de langues couvertes : 40
Analyse : W3Techs attribue une seule langue à chaque site web de l'échantillon, l'anglais étant par défaut choisi si le site l'inclut parmi ses options linguistiques. Ce choix méthodologique introduit un biais important, surestimant fortement la présence de l'anglais en ligne. Il existe d'autres biais (comme l'extrapolation du million de visiteurs supplémentaires à l'ensemble du web, ce qui favorise l'anglais et les langues européennes), mais le biais principal, celui de ne pas prendre en compte le multilinguisme des sites web, peut conduire à une surestimation de l'anglais de l'ordre de 100 % (voir démonstration à https://www.obdilci.org/projets/principal/englishweb/)
Conclusion : Les données produites quotidiennement par W3Techs correspondent en réalité au pourcentage de sites web proposant l'anglais comme option linguistique. Le pourcentage fournit pour les 40 autres langues correspondent aux comptages des sites web dans ces langues, à l'exclusion de ceux proposant une version anglaise. Malgré d'importants biais méthodologiques, W3Techs est devenu une référence incontournable en matière de statistiques linguistiques sur le web, en grande partie grâce à sa longue histoire et à sa bonne réputation établie dans les enquêtes sur les technologies web. Lorsque les décideurs politiques et les chercheurs fondent leurs décisions sur des données aussi biaisées, cela devient un véritable problème de désinformation. La prudence est de mise et la préférence doit être accordée aux méthodes évaluées par les pairs.

APPROCHE 2 : DATAPROVIDER.COM

Source : https://www.dataprovider.com/blog/domains/what-languages-does-the-web-speak/
Méthode : Algorithme de détection de langue appliqué à près de 73 % de tous les sites Web existants (99 M sur 136 M).
Institution : Société de services Internet spécialisée dans l'analyse de données
Portée : Une seule fois en janvier 2023
Méthodologie exposée : Non. Cependant, ils ont gentiment répondu à toutes nos questions et autorisé la publication suivante. Ils explorent, en utilisant https://github.com/jmhodges/gocld3 pour la détection (un modèle identifiant un peu plus de 100 langues), l'ensemble de l'univers web (en 2023, 710 millions de sites web, dont 136 millions jugés valides). Il est à noter que leurs chiffres sont extrêmement cohérents avec les statistiques de Netcraft (https://www.netcraft.com/blog/october-2024-web-server-survey/). En 2023, ils ont appliqué la détection de langue à un sous-ensemble de 99 millions de personnes, filtré par pays (à ce stade, seuls 62 pays étaient inclus, soit moins de 30 %). Ils conservent les informations des différentes versions linguistiques, lorsque cela est spécifié dans l'instruction HTML hreflang=, pour une éventuelle utilisation ultérieure. Cependant, l'étude de 2023 ne détecte que la langue principale des sites web (même biais que W3Techs).
Méthodologie évaluée par les pairs : Non
Discussion sur les biais : Outre le même biais de multilinguisme qui s'applique aux W3Techs, il existe un biais résultant des pays exclus de la sélection.
Intervalle de confiance des résultats : Non disponible
Nombre de langues couvertes : 107
Analyse : Il s'agit d'une approche très intéressante et prometteuse puisque cette société dispose d'une base de données couvrant l'univers presque complet des sites web (aujourd'hui, 163 millions valides sur 856 millions au total) et a le potentiel d'appliquer, en partie, la mesure permettant de prendre en compte le multilinguisme des sites web (en partie, car d'après nos

études, le paramètre hreflang n'est utilisé que par 40 % des sites web multilingues). À ce stade, les résultats de la mesure doivent être interprétés avec la même prudence que ceux de W3Techs. Si l'on ne tient pas compte du biais de sélection des pays, les données pourraient confirmer que le pourcentage mondial de sites web dont l'une des versions est en anglais est comparable, bien que légèrement inférieur, à celui de l'échantillon Tranco analysé par W3Techs, ce qui est logique. En effet, il est probable que de nombreuses langues européennes, dont l'anglais, aient une probabilité plus élevée de figurer parmi les sites web les plus visités. Cette approche est à suivre avec intérêt car elle présente un potentiel d'amélioration vers des résultats qui évitent le biais de multilinguisme.
Conclusion : Dans l'hypothèse où l'entreprise investit dans une nouvelle campagne, incluant cette fois les informations d'une partie des sites multilingues, il est possible d'atténuer les deux biais restants. 1) Le biais de sélection : compte tenu de la liste des pays exclus, l'extrapolation des données manquantes est possible à partir de la combinaison du taux de connexion Internet par pays et du nombre de locuteurs par langue dans chaque pays, ce qui donnerait le pourcentage de locuteurs connectés, pour chaque langue, exclus des données. OBDILCI peut fournir, à partir de ces pourcentages, une correction multiplicative à appliquer à chaque compteur de langue et ainsi atténuer le biais (à noter que cette méthode est largement utilisée dans le modèle OBDILCI pour compléter les statistiques partielles).
2) Le biais résiduel multilingue : d'après nos statistiques approximatives, seuls 40 % des sites web multilingues utilisent l'instruction hreflang pour spécifier la liste des options linguistiques. L'exclusion de 60 % des cas constitue un biais important. Il peut cependant être considérablement réduit en multipliant simplement tous les compteurs de langues par 100/60 = 5/3, en supposant que le modèle obtenu avec 40 % se reproduise approximativement pour les autres. Bien sûr, l'hypothèse peut être erronée, mais le résultat sera beaucoup moins biaisé grâce à cette correction. Ce faisant, DataProvider.com pourrait produire la meilleure approximation jamais réalisée de la mesure de la proportion de langues dans les contenus web et, avec la même logique, pourrait produire la première approximation sérieuse d'une donnée clé, le taux de multilinguisme du Web, comparable à la même valeur pour les humains (la définition de cet indicateur est : le nombre total de versions linguistiques identifiées divisé par le nombre total de sites web analysés). OBDILCI et DataProvider.com ont convenu de coopérer pour la génération de chiffres impartiaux à partir de la base de données DataProvider.com en 2025.

Note de 7/25 : DataProvider.com a offert à OBDILCI un accès de courtoisie à sa base de données, ce qui a permis de développer une série d'études très intéressantes à propos du multilinguisme de la Toile. Voir https://www.obdilci.org/projets/autres/mlreports-2/.

APPROCHE 3 : NETSWEEPER

Source : https://www.netsweeper.com/government/top-languages-commonly-used-internet
Méthode : Algorithme de détection de langue appliqué sur 12 milliards de pages Web.
Institution : Société de services Internet spécialisée dans le filtrage Web à des fins de sécurité.
Portée : Une fois, en juin 2023
Méthodologie exposée : Non. Aucune réponse aux différentes tentatives de communication.
Méthodologie évaluée par les pairs : Non.
Discussion sur les biais : Non fourni. Si l'affirmation selon laquelle ils fonctionnent sur des pages web plutôt que sur des sites web était confirmée, le biais multilingue ne s'appliquerait donc pas. Un biais unique resterait à analyser dans la méthode NETSWEEPER : le biais de sélection. Douze milliards de pages web pourraient représenter 30 % de l'univers web total, ce qui est un chiffre élevé ; pourtant, selon la méthode de sélection, le biais résultant pourrait varier de presque nul à important ! Si les pages sont sélectionnées aléatoirement, le biais est presque nul. Si la sélection est effectuée sur tous les sites web, mais en la limitant à un sous-

ensemble des pages de chaque site, le biais pourrait également varier entre nul ou le biais multilingue, si la sélection privilégie les pages appartenant à la version anglaise. Le fait qu'ils calculent l'anglais à environ 25 % indique que ce biais a été probablement évité. Cependant, en l'absence d'informations sur le processus, cela reste indécidable pour le moment.
Nombre de langues couvertes : 47
Analyse : Si ce qui est affirmé est confirmé, il s'agit d'une méthode exempte du biais multilingue appliqué à une partie substantielle du Web (personne ne connaît réellement le nombre de pages Web, des chiffres d'environ 40 milliards sont donnés par https://www.worldwidewebsize.com), ce qui représenterait 30 % de l'univers. Si le biais de sélection était quasiment nul ou atténué par certaines techniques, cela pourrait devenir le résultat le plus prometteur sur le sujet. La coïncidence avec de nombreux chiffres de l'OBDILCI est frappante et plaide en faveur d'un biais de sélection contrôlé ; cependant, sans plus d'informations sur la méthode, cela reste une hypothèse à confirmer.
Conclusion : Il est dommage qu'ils n'aient jamais répondu à nos nombreuses demandes d'informations. Cette méthode reste un candidat sérieux pour la meilleure méthode de proportionnalité linguistique des contenus web.
A l'attention d'employés de Netsweeper : si par hasard vous lisez cet article, veuillez contacter l'auteur.

APPROCHE 4 : UNIVERSITÉ IONIENNE GRÈCE

Source : (Giannakoulopoulos, A. et al., 2020)
Méthode : L'étude se consacre à mesurer les contenus en anglais dans les domaines de premier niveau (ccTLD) des pays de l'Union européenne. Ils utilisent un algorithme de détection de langue sur 100 000 sites web de 27 pays. Ils évitent les biais multilingues en explorant tous les liens internes et en signalant toutes les langues.
Institution : Université (Département des arts visuels)
Portée : Une fois, en juin 2019
Méthodologie exposée : Oui, entièrement transparent
Discussion sur les biais : Non
Intervalle de confiance des résultats : Non disponible
Nombre de langues couvertes : une seule, l'anglais
Analyse : Il s'agit d'une expérience totalement fiable, mais limitée aux ccTLD européens. Elle peut néanmoins servir d'indicateur potentiel de la proportion de l'anglais à l'échelle mondiale.
Conclusion : Il s'agit de la première incursion, depuis très longtemps, et très bienvenue, du monde universitaire sur ce sujet. L'étude présente tous les attributs de la robustesse d'un travail universitaire évalué par des pairs. Cependant, elle ne cible qu'un sous-ensemble du Web et ses résultats ne peuvent être généralisés à l'ensemble du Web. Quoi qu'il en soit, elle apporte une preuve supplémentaire contre les résultats de W3Techs, qui indiquaient une proportion d'anglais supérieure à 50 %. Il n'y a aucune raison, au contraire, que le pourcentage moyen d'anglais sur les sites web des ccTLD de l'Union européenne avant le Brexit (y compris les pays anglophones : Royaume-Uni, Irlande et Malte) soit deux fois inférieur à celui de l'ensemble du Web, étant donné que la proportion d'anglophones dans ces pays est de 27 %, soit près du double de la valeur mondiale de 14 %.[1], et de plus, c'est la région la plus connectée au monde.
Note de 7/25: OBDILCI a mis à jour cette méthode, à l'aide de la base de données de DataProvider.com dans une étude récente disponible, en anglais, à https://www.obdilci.org/wp-content/uploads/2025/02/WebMultilingualismReport1.pdf. Les nouvelles mesures indiquent

[1] Les deux pourcentages sont calculés à partir des chiffres L1+L2 d'Ethnologue, en divisant la somme des anglophones des pays de l'UE par le total mondial des locuteurs L1+L2.

une nouvelle valeur de 19,76% pour le pourcentage de l'anglais dans sites web des ccTLD de l'Union européenne, incluant le Royaume Uni.

APPROCHE 5 : MÉTHODE PRINCIPALE OBDILCI

Source : https://www.obdilci.org/projects/main/
Méthode : Il s'agit d'une méthode indirecte basée sur la collecte et l'organisation de multiples indicateurs. Il ne s'agit pas d'une mesure, mais d'une approximation réaliste fondée sur des hypothèses et des sources solides, dont certaines impliquent des biais qui sont analysés en détail.
Institution : Organisation de la société civile, travaillant dans ce domaine depuis 1998
Portée : Depuis 2017, une ou deux nouvelles versions par an, la dernière, V6, de juillet 2025
Méthodologie exposée : Oui. (Pimienta et al., 2023)
Méthodologie évaluée par les pairs : Oui (Pimienta et al., 2023)
Discussion sur les biais : Oui, très détaillé et complet, (Pimienta et al., 2023)
Intervalle de confiance des résultats : grand, +-20% (estimé, non calculé)
Nombre de langues couvertes : 361
Analyse : Cette approche indirecte repose sur des données fiables concernant le nombre de locuteurs L1 et L2 de chaque langue par pays (Ethnologue) et le pourcentage de personnes connectées par pays (UIT). Elle suppose l'existence d'une loi économique naturelle reliant la demande (locuteurs d'une langue connectés) et l'offre (contenus pour cette langue). La modularité dépend d'un large ensemble de facteurs représentés par le plus grand ensemble possible d'indicateurs fiables (trafic, abonnements, présence des langues dans les interfaces et outils, préparation à la société de l'information…). Une hypothèse de simplification est formulée : tous les locuteurs d'une langue connectée d'un même pays partagent le même pourcentage de connectivité. Ceci représente le biais principal et la raison pour laquelle le modèle est limité à une grande population de locuteurs (L1 > 1M). Ce n'est donc pas une mesure, mais une estimation avec une plausibilité solide, mais dans un large intervalle de confiance et, tant que les autres méthodes ne sont pas sérieusement contrôlées en termes de biais, elle reste une option couvrant beaucoup plus de langues que les autres méthodes.
Conclusion : Les esprits intéressés et critiques pourraient sincèrement se demander : comment une telle méthode pourrait-elle approcher la réalité simplement en faisant la moyenne de nombreux indicateurs ? Sachant qu'elle repose sur une hypothèse très théorique (l'existence d'une loi inconnue reliant les internautes par langues et les contenus web par langues), cette loi inconnue pourrait-elle être décrite indirectement afin de permettre des chiffres approximatifs, mais fiables, en collectant de multiples indicateurs et en les traitant statistiquement en utilisant principalement des opérations de pondération ? Il existe une réponse intuitive à cette question raisonnable. De la même manière que l'analyse dimensionnelle permet d'approximer les équations physiques de phénomènes complexes, dans le big data, l'application de méthodes statistiques appliquées à un ensemble suffisamment ample d'indicateurs peut produire des métriques approximatives mais fiables. Dans un monde idéal, toutes les langues sont égales et la loi est linéaire : il y aurait proportionnalité des pourcentages de contenus avec ceux des locuteurs dans chaque langue. Le ratio que nous appelons *productivité des contenus* (pourcentage de contenus divisé par pourcentage de locuteurs connectés) serait égal à un pour chaque langue, ce serait une équation linéaire. Nous obtenons ces données linéaires en pondérant la matrice des locuteurs (langues vs pays) avec le vecteur de connectivité (pourcentage de connectés par pays). La complexité que reflète la série des autres indicateurs réside dans le fait que de nombreux facteurs influent sur une variation de ce ratio au-dessus ou en dessous de 1, selon chaque langue : au-delà de ses capacités technologiques, le nombre de locuteurs de L2 par pays, les tarifs Internet, la bande passante, l'éducation numérique, les applications d'administration en ligne, l'environnement économique, la présence dans les

principales applications, etc. Si l'on recueille suffisamment d'indicateurs de tous ces paramètres, il y a de fortes chances, en ne traitant que des données massives (langues comptant un grand nombre de locuteurs), que l'équation statistique résultante soit une approximation raisonnable. Il est à noter qu'une grande partie des facteurs dépendent des pays plutôt que des langues, mais l'existence de la matrice langues par pays permet de jouer le jeu, en fournissant certaines simplifications, qui entraînent certes des biais, mais ces biais peuvent devenir marginaux avec des grands nombres de locuteurs. De toute évidence, si les résultats du modèle OBDILCI pouvaient être confirmés par des mesures de données réelles, à condition que les biais soient contrôlés, cela renforcerait la confiance…

APPROCHE 6 : ÉTUDES PRÉALABLES MECILDI@ OBDILCI

Source : https://obdilci.org
Méthode : Il s'agit d'un effort manuel appliqué à une série de dix fois 100 sites choisis aléatoirement dans la liste TRANCO. Nous avons vérifié manuellement toutes les langues de chaque site web et la manière dont les options linguistiques sont implémentées, tant dans l'interface que dans le code source HTML, afin d'étudier la stratégie et les tactiques permettant d'envisager ultérieurement une approche d'exploration basée sur la détection de la langue (voir approche 7). Nous en avons profité pour approximer un indicateur clé totalement inconnu à ce stade : le *taux de multilinguisme du Web*, défini par le nombre total de versions linguistiques divisé par le nombre total de sites web (le même taux pour l'ensemble de l'humanité est mesuré à 1,443, d'après la source Ethnologue, et nous pensons que le Web présente un chiffre plus élevé et souhaitons le démontrer). Ce chiffre est essentiel pour évaluer l'ampleur du biais lié à la non-prise en compte du multilinguisme du Web : par exemple, si sa valeur est de 2, le biais correspond à une surestimation de 100 % de la proportion d'anglais. Les premières approximations lors de l'exploration manuelle de 1 000 sites web choisis aléatoirement dans la liste Tranco sont justement d'environ 2 (avec une variance élevée, donc à prendre avec prudence).
Institution : Organisation de la société civile
Portée : Deux fois en 2022 et 2024
Méthodologie exposée : Oui. Totalement transparent.
Méthodologie évaluée par les pairs : Oui voir (Pimienta, 2024)
Discussion sur les biais : Oui
Intervalle de confiance des résultats : Non
Nombre de langues couvertes : Anglais seulement
Analyse : Il s'agit simplement d'une étude intermédiaire créée par l'exploration humaine d'un sous-ensemble limité du Web, servant d'indication de tendance, notée avec moyenne et covariance. Ceci fait partie du projet MECILDI.
Source : https://obdilci.org

APPROCHE 7 : MECILDI@OBDILCII

Méthode : OBDILCI prévoit de créer un nouvel outil en 2025, un logiciel permettant la détection de la langue sur une série de sites web, en tenant compte systématiquement du fait que les sites peuvent être multilingues. Cet outil servira à divers projets et sera d'abord testé avec la liste Tranco. Des études préliminaires ont commencé à déterminer des stratégies et des tactiques pour tendre vers une prise en compte complète des langues des sites web multilingues. Il s'agit d'un problème complexe en raison de la diversité des solutions mises en œuvre sur les sites web, dont beaucoup ne se reflètent pas directement dans le code source visible. Les études préliminaires ont permis de déterminer des statistiques et des données approximatives utiles pour atténuer les biais : pourcentage de sites web utilisant les instructions lang=, pourcentage

de sites web utilisant les instructions hreflang=, pourcentage de sites web utilisant Google Translate intégré, pourcentages de disposition des options de langue dans l'interface (en haut, sur le côté, en bas, indirectement par pays, dans une page de configuration), modèles de codage utilisés pour le multilinguisme…). La complexité implique une combinaison de techniques et d'approches, incluant probablement l'utilisation d'intelligence artificielle. Notre puissance de calcul étant limitée, nous avons opté pour une approche statistique : au lieu d'analyser tous les sites web, nous allons créer 100 échantillons aléatoires de 1000 sites web et gérer la distribution statistique des résultats pour obtenir la moyenne, la variance et l'intervalle de confiance pour chaque langue et le reste des paramètres.
Institution : Organisation de la société civile
Portée : Futur (2025)
Méthodologie exposée : Elle le sera
Méthodologie évaluée par les pairs : Un article sera publié.
Discussion sur les biais : Oui
Intervalle de confiance de la figure : Sera calculé par méthode statistique
Nombre de langues couvertes : 141, les langues présentes à la fois dans le modèle Obdilci et dans Google Translate. Autrement dit, il s'agit du sous-ensemble de langues issues de 250 Google Translate comptant plus d'un million de locuteurs L1. Pourquoi ? Parce que pour les langues comptant peu de locuteurs, l'approche statistique choisie ne fournirait pas de résultats probants.
Analyse : Il s'agit d'un projet à réaliser en 2025 ouvrant des perspectives à de nouvelles recherches.

SYNTHÈSE

Le tableau suivant compare les sept approches détaillées précédemment, en se concentrant sur des aspects clés tels que l'évaluation par les pairs, la méthodologie, la taille de l'échantillon et la gestion des biais. Ces comparaisons mettent en évidence des différences substantielles en termes de fiabilité et de potentiel de mesure précise du langage web. À des fins de comparaison, les données de 2023 ont été sélectionnées, à l'exception de celles de l'Université Ionienne, dont la date est de 2020.

	W3TECHS	DataProvider .com	Netsweeper	UNIV. IONIEN.	OBDILCI PRINCIPAL	OBDILCI MECILDI
TYPE	COM	COM	COM	ÉDU	ORG	ORG
REVUE PAR LES PAIRS	NON	NON	NON	OUI	OUI	OUI
# LANGUES	40	107	47	1	361	141
TALLE échantillon	1 M sites (TRANCO)	99 M. sites *	12 B. pages **	0,1 M. ***	INDIRECT ****	TRANCO *****
Gestion des BIAIS	NON	NON	NON?	OUI	OUI	OUI
BIAIS	ÉNORME #	ÉNORME #	FAIBLE?	NON	FAIBLE	NON
Découvrabilité	ÉNORME	MOYENNE	FAIBLE	FAIBLE	CROISSANTE	
UTILISATION	ÉNORME	NON	NON	NON	FAIBLE	
% ANGLAIS	58%	51%	26%	28%	20%	29%
CHINOIS %	1,2%	10%	20%		19%	
Débiaisé	29%	25%##	26%	28%	24%	29%

Notes associées au tableau :
* : La base de données DataProvider.com couvre aujourd'hui l'ensemble de l'univers des sites Web (150 millions) mais au moment de l'enquête linguistique, elle ne couvrait que 99 millions, à l'exclusion de certains pays.
** : Netsweeper affirme couvrir 12 milliards de pages Web sur une période estimée à 40 ans. Aucune information ne permet de comprendre le biais de sélection.
*** : Les études de l'Université Ionienne se concentrent sur l'anglais dans les ccTLD des pays de l'UE.
**** : La méthode Obdilci n'est pas une mesure mais une approximation indirecte.
***** : Mecildi est une méthode en phase de construction qui utilisera Tranco avec une méthode dépassant le biais de multilinguisme.
# : Les biais sont nombreux, mais le biais majeur, appelé « biais de multilinguisme », résulte de la non-prise en compte du fait que les sites web peuvent contenir plusieurs langues. Cela entraîne une surestimation de l'anglais de 100 % si le taux de multilinguisme de l'échantillon est de 2.
: Obdilci se verra accorder par DataProvider.com l'accès à son immense base de données et sera en mesure de publier dans les prochains mois un résultat complet contrôlé par des biais.

II – COMPARAISONS ET DISCUSSION

2.1 – COMPARAISONS DES RÉSULTATS POUR L'ANGLAIS

W3TECHS 1/2023	**DataProvider.com 1/23**	**Netsweeper 6/23**	**Univ. IONIEN. 2020**	**OBDILCI Principal 5/2023**	**MECILDI** Étude préalable **5/2024**
57,7%	**51%**	**26,3%**	**28,4%**	**20%**	**29%**

Le biais de multilinguisme affecte les deux résultats à gauche du tableau avec un effet probable de surestimation de 100 % (Pimienta, 2024). La valeur la plus probable pour la proportion d'anglais dans toutes les pages web, en tenant compte de chaque méthode et de ses biais, était donc, en 2023, d'environ 25 %, avec une fenêtre probable de 20 % à 27 %. Quant aux données de l'université Ionienne, qui sont probablement les plus fiables, il est important de noter que le pourcentage de locuteurs anglais (L1+L2) dans l'Union européenne, y compris le Royaume-Uni, est le double du pourcentage mondial, tous deux calculés sur les populations L1+L2 (27 % sur 14 %) d'après la source Ethnologue. Cela indique qu'une extrapolation de l'UE au reste du monde donnerait un chiffre de contenu en anglais bien inférieur à 20 %.

2.2 - COMPARAISON DES PREMIÈRES LANGUES

	W3TECHS 11/2024	DataProvider 1/23	Netsweeper 6/23	OBDILCI 5/2024
1	**Anglais 49,4%**	**Anglais 51,3%**	**Anglais 26,3%**	**Anglais 20,4%**
2	**Espagnol 6%**	**Chinois 10,3 %**	**Chinois 19,8 %**	**Chinois 18,9 %**
3	**Allemand 5,6%**	**Allemand 7,3%**	**Espagnol 8,1%**	**Espagnol 7,7%**
4	**Japonais 5%**	**Espagnol 3,9%**	**Arabe 5%**	**Hindi 3,8%**
5	**Français 4,4%**	**Japonais 3,7 %**	**Portugais 4%**	**Russe 3,7%**
6	**Russe 4%**	**Français 3,4%**	**Malais 3,4 %**	**Arabe 3,7%**
7	**Portugais 3,8%**	**Russe 2,8%**	**Français 3,3%**	**Français 3,4%**
8	**Italien 2,7%**	**Portugais 2,7%**	**Japonais 3%**	**Portugais 3,1%**
9	**Néerlandais 2,1%**	**Néerlandais 2,0 %**	**Russe 2,8%**	**Japonais 2,2 %**
10	**Polonais 1,8%**	**Italien 1,9%**	**Allemand 2,1%**	**Allemand 2,2%**
11	**Turc**		**Coréen**	**Malais**
12	**Persan**		**Turc**	**Bengali**
13	**Chinois 1,2%**		**Italien**	**Turc**
14	**Vietnamien**		**Roumain**	**Italien**
15	**Malais**		**Persan**	**Vietnamien**

La proximité des premiers résultats entre Netsweeper, seule méthode ciblant directement les pages web, et Obdilci est vraiment impressionnante (sauf pour l'hindi). Les résultats de W3Techs pour le chinois, première langue en termes de locuteurs connectés à Internet, sont totalement improbables. Connaître plus de détails sur la sélection de DataProvider.com et NetSweeper permettrait d'atténuer leur biais de sélection. Comme DataProvider.com permettra à Obdilci d'accéder à sa base de données, il est probable que nous disposerons bientôt de chiffres concluants avec un biais de sélection contrôlé.

2.3 QUE NOUS APPRENNENT CES COMPARAISONS ?

1. La **prudence** est fortement recommandée lors de la lecture des chiffres sur le pourcentage de langues sur le Web, en particulier lorsqu'il s'agit de l'anglais, car il n'y a pas de concordance entre les différents résultats.

2. **Pourquoi les Chinois sont-ils si mal notés par W3Techs ?** Les résultats annoncés par W3Techs pour le chinois sont parfaitement invraisemblables. Où se situe la valeur réelle entre 10 % et 20 % ? Le chinois étant probablement utilisé sur de nombreux sites bilingues (chinois et anglais), la même règle que pour l'anglais pourrait s'appliquer. Il faudrait peut-être multiplier par 2 les chiffres de DataProvider.com pour obtenir un consensus autour de 20 %. Lors des études préliminaires de MECILDI, nous avons découvert qu'une forte proportion de sites web chinois (50 % de notre échantillon !) définissent le paramètre HTML « lang = » comme l'anglais plutôt que le chinois. Cela pourrait expliquer la sous-estimation si la méthode repose sur ce paramètre.

3. Jusqu'à présent, Netsweeper pourrait être considéré comme le résultat le plus fiable, car sa méthode ciblant les pages web plutôt que les sites web eux-mêmes évite le biais de multilinguisme. Netsweeper affirme couvrir 12 milliards de pages web, un chiffre colossal qui pourrait représenter 30 % de l'univers des pages web. Le manque d'informations laisse la question indécise à ce stade. La proximité entre les chiffres d'Obdilci et ceux de Netsweeper est frappante. Les principales différences concernent les langues indiennes (hindi, bengali, ourdou) avec toutes les autres méthodes. Ce point mérite d'être souligné compte tenu de l'importance démographique de l'Inde. Une étude approfondie réalisée en 2017[2]a affirmé que les internautes indiens avaient tendance à utiliser de plus en plus leur langue locale pour naviguer. Cette étude semble corroborer les chiffres d'Obdilci et il est possible que l'Inde n'ait pas été incluse dans l'échantillonnage de DataProvider.com et NetSweeper, le biais de sélection expliquant alors les différences.

4. Il est intéressant de comparer les prédictions d'OBDILCI avec les mesures de DataProvider.com pour les langues à faible contenu. Nous avons constaté des coïncidences extrêmes (galicien et basque) ainsi que des chiffres extrêmement éloignés (afrikaans, créole haïtien, irlandais et langues indiennes). Le biais de sélection des pays pourrait être une explication à étudier.

III INITIATIVES DE LA PREMIÈRE PÉRIODE (1996-2014)

Cette section résume brièvement, par ordre chronologique, les approches des premières mesures de langues sur le web. Pour une analyse détaillée, voir (Pimienta et al., 2009).

Étude Xerox (1996-2000)

Méthode : Approche linguistique basée sur l'occurrence de mots fréquents dans le corpus.
Source : (Grefenstette, G. et Noche, J, 2000)
Portée : Méthode réalisée une seule fois, non reproduite. Il s'agissait de la première tentative historique.
Discussion : Offre peu de pourcentages de langues en dehors de l'anglais.

OBDILCI/Funredes (1998-2007)

Méthode : Utilisait la capacité, fiable à cette époque, des moteurs de recherche à rapporter le nombre d'occurrences d'une chaîne de caractères dans l'ensemble des pages web indexées. Propose un vocabulaire comparatif, sélectionné avec un soin extrême en termes de correspondance syntaxique et sémantique, ainsi qu'une analyse des biais, pour un ensemble de langues sélectionnées : anglais, français, espagnol, italien, portugais, catalan, roumain et

[2] « Langues indiennes – Définir l'Internet indien », KPMG et Google, 2017 – (en anglais) https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

allemand. Utilisait des techniques statistiques pour obtenir des résultats en termes de pourcentage de chaque langue par rapport à l'anglais. Le pourcentage d'anglais était ensuite approximé par diverses techniques.
Source : Site historique d'Obdilci (https://funredes.org/lc/francais/inicio/)
Portée : Plusieurs mesures ont été réalisées entre 1998 et 2007, montrant un déclin de l'anglais de 80 % à 50 % et une croissance générale des langues européennes non anglaises. Il s'agissait de la deuxième tentative historique et de la seule, avec la LOP, à avoir proposé des observations en série sur une longue période.
Discussion : L'évolution des moteurs de recherche, faisant perdre totalement la crédibilité des réponses en termes d'occurrences, après 2007, a marqué la fin de cette méthode. OBDILCI/Funredes a poursuivi sa mission jusqu'en 2017, date à laquelle Funredes a cessé ses activités, avec des contributions sur le terrain principalement au sujet du français et de l'espagnol. La recherche d'une nouvelle approche a émergé en 2012, à partir de l'idée de Daniel Prado de mesurer la place des langues à partir d'une vaste collection d'indicateurs et de transformer les indicateurs pays en indicateurs langues par recoupement avec des données démolinguistiques. Cette nouvelle méthode a mûri en 2017 et ses biais ont été contrôlés en 2022.

ISOC Québec/Alis Technologies, suivi d'OCLC (1997, 1999, 2002)

Méthode : Une série de sites web est obtenue par génération aléatoire de 8 000 adresses IP. Un algorithme de détection de langue est appliqué à cette série et les pourcentages sont calculés, une seule fois. Cette méthode n'est pas statistiquement valide car il s'agit d'un essai unique, alors que plusieurs exécutions auraient été nécessaires pour obtenir une distribution aléatoire sur laquelle appliquer les lois statistiques (moyenne, variance, intervalle de confiance). Cette méthode a été reproduite à l'identique à deux reprises, en 1999 et 2002, avec le même défaut. Les trois mesures ont donné le même score de 80 %, stable pendant cinq ans, ce qui, grâce à un marketing efficace, a alimenté la désinformation pendant cette période, jusqu'à ce que les publications de l'UNESCO en 2009 incitent les médias à adopter la valeur de 50 %.
Portée : Trois mesures uniques en 1997, 1999, 2002.
Sources : https://web.archive.org/web/20010810234537/http://alis.isoc.org/palmares.en.html (Lavoie, BF, et O'Neill, ET, 1999) et (O'Neill, ET et al., 2003)

INKTOMI (2000)

En 2000, un moteur de recherche, INKTOMI, a annoncé avec une force marketing considérable ses mesures des langues sur le Web. Il présentait les 10 premières langues, l'anglais en tête, à 86 %. Un détail important que peu d'observateurs semblaient remarquer : le pourcentage total des 10 langues était de 100 %, un manque de rigueur mathématique qui ne prenait pas en compte l'existence de beaucoup d'autres langues.

Google : Méthode du complément de l'espace vide (1988-2008)

C'est ainsi que nous avons nommé une fonctionnalité découverte par hasard en mars 1998 avec AltaVista et reproduite par Google, qui permettait à l'époque de connaître la taille, par langue, de l'index du moteur de recherche. En effectuant une requête au moteur de recherche du type « - ggfdgfdyugfgvdgdv », où le premier terme est vide et le second une chaîne de caractères n'apparaissant nulle part dans les pages web, le chiffre renvoyé était le nombre total de pages web de l'index. Si une langue spécifique était sélectionnée en option, la réponse était le nombre total de pages dans cette langue. Les chiffres fournis par Google avec cette méthode était proche de 51 % pour l'anglais en 2008 (le même que celui d'Obdilci) et déjà d'environ 9 % pour le chinois.

Projet d'Observatoire des Langues – LOP (2003-2011)

Méthode : Application de la détection de langue à une partie du Web, généralement les ccTLD de pays et ciblant les langues locales. Ce projet, mené par un consortium d'universités dirigé par l'Université de Nagaoka, a porté l'espoir de voir enfin ce sujet important intégré à la communauté de recherche. L'adhésion commune de Funredes/Obdilci et de LOP au réseau MAAYA (Réseau mondial pour la diversité linguistique) était également le gage d'une coopération fructueuse. Cette coopération s'est renforcée fin 2010, lorsque Funredes s'est vu confier par la LOP les données nécessaires à l'exploration du ccTLD d'Amérique latine et a collaboré étroitement pour l'évaluation du matériel. Cependant, le tsunami catastrophique survenu au Japon en 2011 a brutalement mis fin à ce projet prometteur.
Sources :
https://dl.acm.org/doi/10.1145/1062745.1062833
https://en.wikipedia.org/wiki/Language_observatory

UPC/IDESCAT (2003-2006)

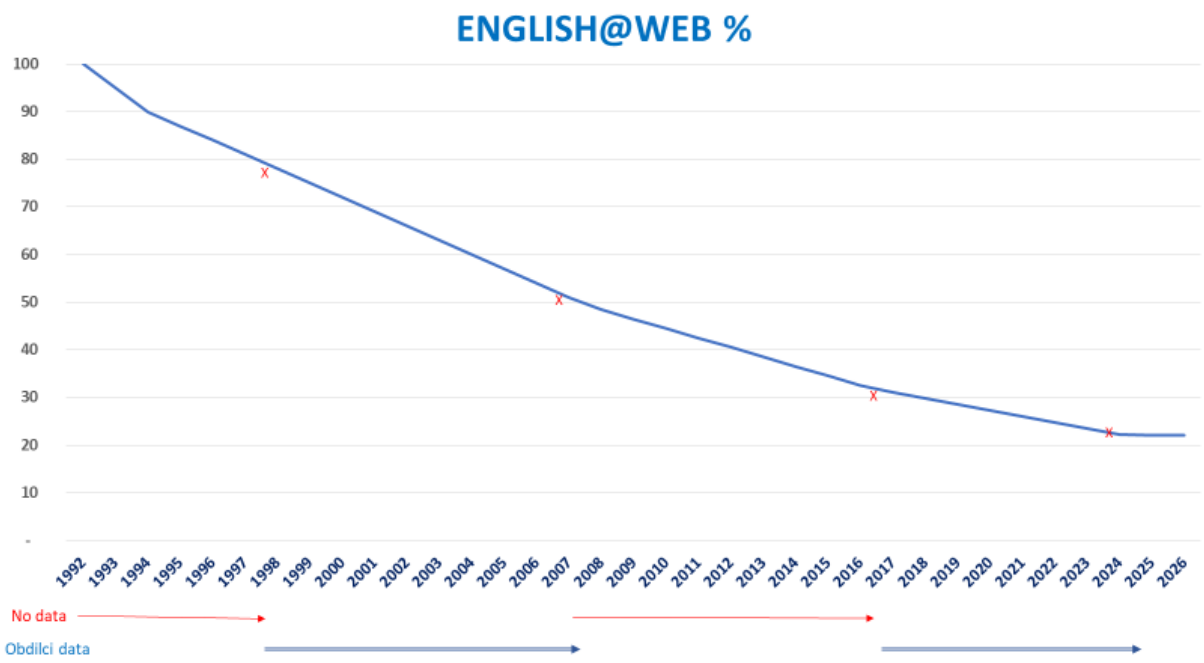
L'Université Polytechnique de Catalogne et l'Institut statistique de Catalogne ont organisé une base de données de 2 millions de sites Web pour vérifier la présence du catalan, avec détection de langue et ont présenté des résultats assez proches de Funredes/Obdilci en 2005 et moins proches en 2006.
Source : (Monrás, F. et al., 2006)

DILINET/SEMACORE (2010-2014)

En 2011, l'Union Latine, l'Organisation de la Francophonie et l'UNESCO ont uni leurs efforts pour permettre à MAAYA d'assurer la coordination de la création d'un consortium de recherche pour répondre à l'appel à propositions du Programme-cadre de recherche de l'Union européenne. Un consortium, composé de chercheurs et d'institutions de haut niveau et motivés, a été mis en place pour proposer une approche globale de très haut niveau sur le thème des langues dans l'Internet. Un projet intégré très ambitieux a été défini et soumis à l'appel à propositions ICT-2011.4.4 « Gestion intelligente de l'information ». Le projet a manqué d'un demi-point sa présélection, à un moment où le thème n'était pas considéré comme stratégique par l'UE. Une deuxième tentative a été menée en séquence, sous le nom de SEMACORE, avec une équipe plus concentrée et des objectifs relativement moins ambitieux, afin de diviser le budget par deux. Il a été soumis à l'appel à propositions ICT-2013.4.1 « Analyse de contenu et technologies linguistiques – Analyse de contenu cross-média ». Cet appel, le dernier du Programme-cadre, a reçu un nombre de propositions bien supérieur à la normale, ce qui a rendu la compétition davantage axée sur l'importance thématique que sur les qualités intrinsèques du projet. Encore une fois, le projet n'a pas réussi à obtenir de financement, laissant un grand vide dans ce thème pendant de nombreuses années.
Source : https://www.obdilci.org/projects/other/dilinet/.

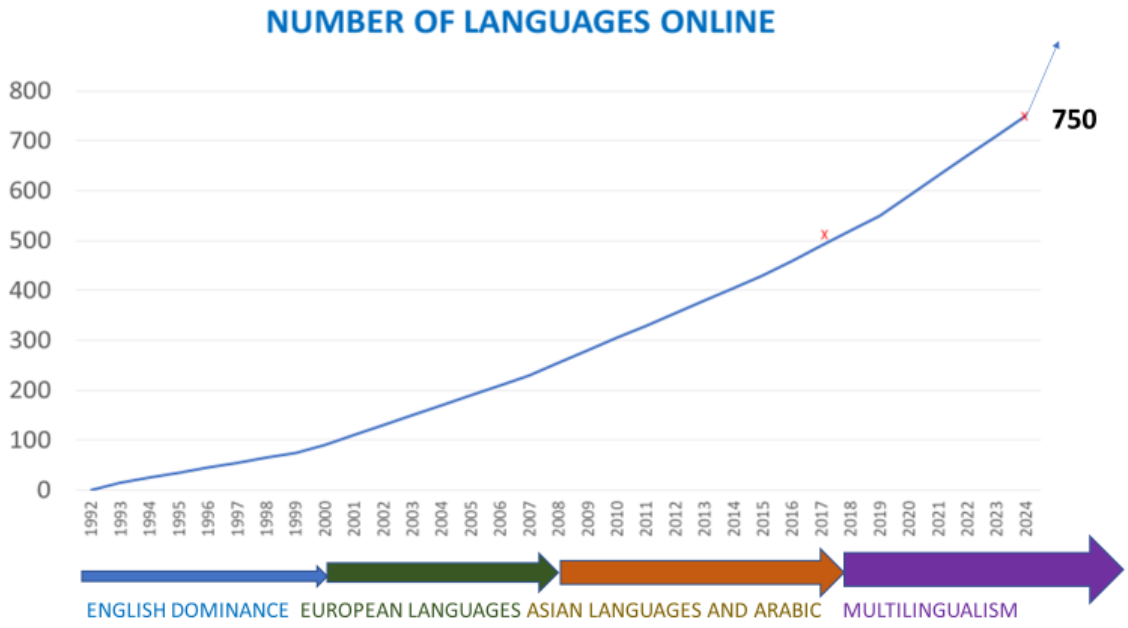
IV Évolution des indicateurs linguistiques clés de l'Internet

Cette section présente des tendances spéculatives, mais fondées sur des données probantes, concernant les indicateurs clés de l'évolution des langues dans l'Internet. Ces tendances sont issues des études de l'OBDILCI et étayées par d'autres analyses.



Cette courbe est supportée par une série de mesures réalisées par OBDILCI entre 1998 et 2007 et depuis 2017. Le reste de la courbe (la partie sans données) est extrapolé.

Ce pourcentage décroît de 100 % à la naissance du Web, en 1992, à un niveau asymptotique aujourd'hui, compris entre 20 et 25 %. Il pourrait rebondir vers un niveau asymptotique plus élevé, compris entre 25 et 30 %, lorsque la fracture numérique en Afrique sera comblée et, facteur plus important, lorsque le taux de multilinguisme du Web augmentera fortement, laissant à l'anglais un confortable rôle de leader comme deuxième langue de la majorité des sites web multilingues.



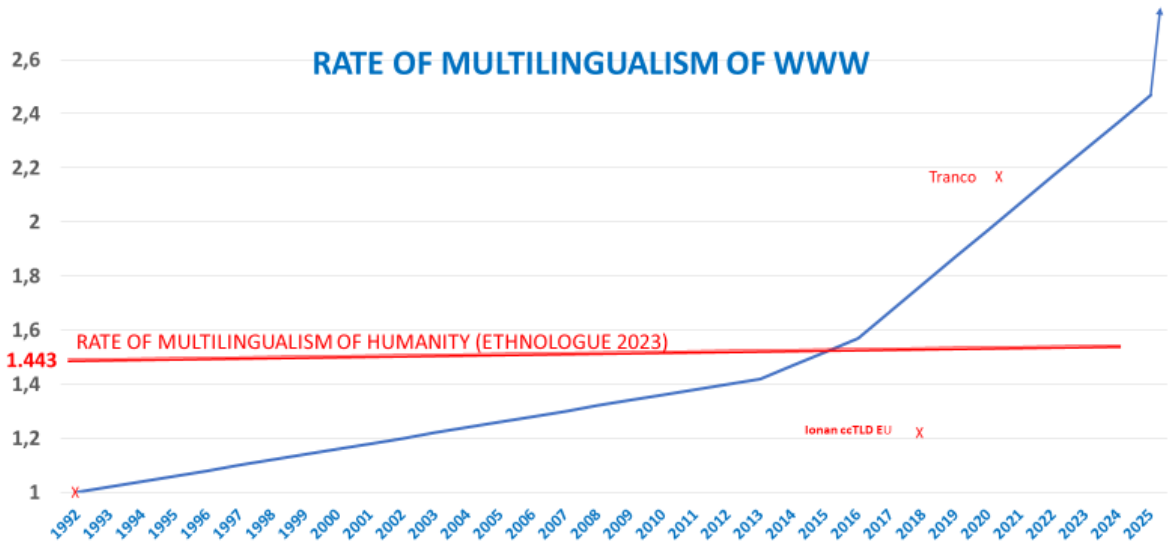
Cette courbe est spéculative ; seules deux tentatives de mesure ont été réalisées : en 2017 (500) et en 2024 (750). Les mesures étaient basées sur le nombre de langues enregistrées sur Unicode.org. Cette courbe illustre également l'évolution des langues dans l'Internet.

Le fait est qu'aujourd'hui, il y a de fortes chances que toutes les langues comptant plus d'un million de locuteurs fassent partie du monde numérique et aient des contenus (voir le modèle https://obdilci.org pour les indicateurs). Cela implique que, malgré le fait que moins de 10 % des langues ont une existence numérique (750 sur 7615), seulement 4 % de la population mondiale ne peut pas accéder à Internet ou trouver des contenus dans sa première ou sa deuxième langue, comme le montre le tableau suivant.

	L1>1M	L1<1M	TOTAL
LANGUES	336	6 890	7 226
%	**4,6%**	**95,4%**	100%
LOCUTEURS L1	7 062 906 326	370 592 893	7 433 499 219
%	**95,0%**	**5%**	100%
LOCUTEURS L1+L2	10 303 359 711	421 864 047	10 725 223 758
%	**96,1%**	**3,9%**	100%

Source : Élaboration réalisée à partir de la base de données Ethnologue #27, 2023, après regroupement des macro-langues

Cela ne signifie pas que les efforts ne doivent pas être accélérés pour inclure les 90 % de langues qui n'ont pas d'existence numérique, car la possibilité d'interagir avec l'Internet dans sa langue maternelle devrait devenir un droit citoyen accordé à tous les locuteurs de langues dans le monde.



Le taux de multilinguisme est défini comme le rapport entre le nombre total de versions linguistiques de tous les sites web et le nombre total de sites web. Il s'agit de la moyenne des versions linguistiques par site web. Cet indicateur clé est une grande inconnue du Web. Seules quelques mesures partielles ont été réalisées à ce jour : l'étude menée par l'Université Ionienne sur 100 000 sites web de l'Union européenne (ccTLD) (Pimienta, 2024) a obtenu une valeur

moyenne comprise entre 1,1 et 1,25 selon le mode de calcul. Dans (Pimienta, 2024), les mesures manuelles sur 7 échantillons aléatoires de 100 sites de la liste Tranco ont montré une moyenne de 2,23, avec les chiffres suivants :

- 75 % des sites Web sont monolingues[3]
- 10% des sites Web sont bilingues
- 4% des sites Web sont trilingues
- 10% des sites Web ont plus de 3 langues, avec une moyenne d'environ 11 langues.

La possibilité d'intégrer Google Translate aux sites web permet d'ajouter, gratuitement et assez facilement, 249 versions linguistiques à n'importe quel site web. Si seulement 1 % des sites web décidaient d'implémenter cette option, le taux de multilinguisme serait multiplié par 2,5. On pourrait toutefois légitimement affirmer qu'il ne s'agit pas de versions linguistiques obtenues par traduction, le résultat étant encore médiocre (Martin B., 2019), mais plutôt d'une aide à l'intercompréhension entre différentes langues.

Quoi qu'il en soit, la généralisation et les améliorations rapides des outils automatiques de gestion des langues basés sur l'IA vont évidemment stimuler le taux de multilinguisme du WWW dans les années à venir, car il est plus facile pour un site Web que pour un humain de s'exprimer dans plusieurs langues…

V CONCLUSION

Les progrès de l'IA ont considérablement amélioré l'accessibilité linguistique en ligne. Autrefois coûteuses et gourmandes en ressources, des tâches comme celles présentées ci-dessous sont désormais plus rapides et plus abordables, et génèrent des gains de productivité notables :

- **Génération de traductions initiales de documents** sans perdre le formatage
- **Création de versions multilingues de sites Web,** avec traduction automatique intégrée lors de la création de contenu.
- **Organiser des vidéos sur des plateformes comme YouTube**, où les utilisateurs peuvent facilement définir des sous-titres dans leur langue préférée (parmi les 249 langues prises en charge par Google Traduction). Bien que les traductions soient souvent approximatives, cette fonctionnalité est suffisamment rapide pour gérer la vitesse de parole et facilite grandement l'intercompréhension.
- **Intégrer l'interprétation automatique dans des plateformes comme Zoom** fournit une autre couche d'intercompréhension, même si elle est loin d'atteindre le niveau d'interprétation professionnelle en temps réel.
- **Étendre ces capacités aux conférences en face à face** avec des appareils permettant aux participants de choisir leur langue préférée, cela représente une avancée en matière d'accessibilité et d'inclusion.

Ces innovations transforment la communication mondiale, réduisent la prédominance de l'anglais comme lingua franca et favorisent la diversité linguistique. Une véritable révolution est en cours, qui transformera les réunions internationales, réduisant potentiellement la

[3] Ce chiffre est compris entre 81 et 87% dans l'étude de l'université ionienne, selon le mode de calcul de la moyenne.

prédominance de l'anglais comme lingua franca et supprimant ainsi les désavantages injustes dont souffrent ceux qui maîtrisent mal l'anglais (une langue comprise par moins de 20 % de l'humanité).

Dans les années à venir, nous pouvons nous attendre à de nouveaux perfectionnements et à une adoption généralisée de ces outils, entraînant un changement de paradigme en matière de diversité linguistique en ligne. Cela implique l'extension de ces services à davantage de langues et l'amélioration de la qualité des traductions pour les langues moins courantes, qui pourraient aujourd'hui être inférieures au seuil d'utilisabilité, comme l'ont suggéré certaines études (Martin, B., 2019).

À l'instar des avancées de l'IA dans d'autres domaines, la traduction automatique ne remplacera pas les professionnels qualifiés. Elle constituera plutôt un outil précieux pour améliorer leur productivité. La traduction et l'interprétation assistées par l'IA ne supprimeront pas le besoin d'interprètes et de traducteurs hautement qualifiés. Au contraire, elles offriront un support exceptionnel, peu coûteux et facile à utiliser pour l'intercompréhension mutuelle. Une fois les seuils de qualité améliorés dans toutes les langues, la portée de ces outils s'étendra encore davantage.

Ce n'est pas un hasard si ce changement de paradigme dans les technologies linguistiques s'accompagne de la transformation de l'Internet en l'espace le plus multilingue jamais vu. La métaphorique « tour de BabelIA » n'atteint peut-être pas le ciel, mais elle rapproche les gens en surmontant les barrières linguistiques, notamment dans l'Internet.

La lingua franca de l'Internet a peut-être été l'anglais lors de sa première période de développement, portée par les réseaux de recherche, conséquence naturelle de son statut de langue privilégiée pour les publications scientifiques et l'informatique. Aujourd'hui, et de plus en plus, la lingua franca d'Internet est le multilinguisme, stimulé par les outils d'IA pour les traductions entre langues.

Malheureusement, les langues avec une faible population de locuteurs, bien que représentant le plus grand nombre de langues[4], pourrait être exclues de ce changement de paradigme, et des solutions innovantes devront être recherchées par les spécialistes des technologies linguistiques. Quant à l'Internet d'aujourd'hui, plus de 96 % de la population mondiale peut y avoir accès en utilisant sa première ou sa deuxième langue[5].

IV RÉFÉRENCES

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

[4] D'après les données Ethnologue 27 de mars 2024, 81 % des 7615 langues existantes comptent moins de 100 000 locuteurs, 56 % moins de 10 000 et 30 % moins de 1 000 (chiffres calculés avant le regroupement des macro-langues).

[5] Chiffre obtenu à partir du modèle Obdilci de création des indicateurs pour la présence en ligne des langues comptant plus d'un million de locuteurs (https://obdilci.org/projets/principal).

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

Lavoie, B. F., and O'Neill, E. T. (1999). How “World Wide” is the Web? Annual Review of OCLC Research.
https://www.researchgate.net/publication/271903988

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

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

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

Pimienta, D., Prado, D., and Blanco, Á. (2009). Douze années de mesure de la diversité linguistique dans l’Internet – Paris/UNESCO publications pour le Sommet mondial pour la société de l’information.
https://unesdoc.unesco.org/ark:/48223/pf0000187016_fre

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

Martin B., (2019), Teach You Backwards: An In-Depth Study of Google Translate for 108 Languages, https://www.teachyoubackwards.com/

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

Pimienta, D. (2022). Resource: Indicators on the Presence of Languages in Internet, In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, Marseille. European Language Resources Association. 83–91. https://aclanthology.org/2022.sigul-1.11/
Version française :
https://obdilci.org/wp-content/uploads/2024/01/Res.Ind_.lang_.Internet.fr_.pdf

Pimienta, D., Müller de Oliveira, G., and Blanco, Á. (2023). The method behind the unprecedented production of indicators of the presence of languages in the Internet Front. Res. Metr. Anal., 17 May 2023 Sec. Research Methods Vol. 8
https://doi.org/10.3389/frma.2023.1149347
Version française : https://obdilci.org/wp-content/uploads/2024/01/METHODV3.fr_.pdf

Pimienta, D. (2024). Is it true that more than half the Web contents are in English? If Web multilingualism is paid due attention then no! Forum for Linguistic studies, Vol. 6 , Iss. 5
https://doi.org/10.30564/fls.v6i5.7144
Version française :
https://obdilci.org/wp-content/uploads/2024/06/English@www.en_.fr_.docx