MECILDI : une plateforme d’observation multilingue pour mesurer la distribution linguistique du contenu Web

Daniel Pimienta et Mike Field, aout 2026.

Mots-clés : diversité linguistique sur Internet, contenu web, multilinguisme web, webométrie, biais, Tranco, sociolinguistique informatique, métriques des langues numériques

Daniel Pimienta
Observatoire de la diversité linguistique et culturelle dans l’Internet (OBDILCI), Nice, France
pimienta@funredes.org
https://orcid.org/0000-0001-7314-6471

Mike Field,
Mike Field Enterprises, Toronto, Canada,
mike@mikefield.ca
https://orcid.org/0009-0009-5892-7937

Les travaux qui ont permis de développer la plateforme ouverte MECILDI ainsi que la rédaction de cet article de présentation de la méthode et des résultats ont été soutenu financièrement par la Délégation à la langue française et aux langues de France (DGLFLF) du Ministère de la culture de la France et ont compté également avec le soutien de l’Organisation Internationale de la Francophonie (OIF). La version complète de la plateforme MECILDI, dont les travaux viennent de commencer, reçoit le soutien financier principal de la part de l’Organisation de la Francophonie et compte aussi sur le soutien de la DGLFLF.

Résumé

La mesure précise de la distribution des langues dans l’Internet nécessite de prendre en compte le multilinguisme des sites web, le fait qu'un même site puisse proposer du contenu en plusieurs langues. Pourtant, la plupart des méthodes existantes n'attribuent qu'une seule langue par site, ce qui biaise fortement les résultats. Cette étude présente MECILDI, une plateforme informatique développée par l'Observatoire de la diversité linguistique et culturelle dans l’Internet (OBDILCI) qui mesure la distribution des langues et le multilinguisme au sein de toute série de sites web. La version présentée ici en est la première implémentation, s'appuyant principalement sur les annotations *hreflang* pour détecter les versions linguistiques et appliquant un facteur

d'extrapolation empirique pour compenser les sites web multilingues qui utilisent une autre technique que *hreflang*. MECILDI a été appliqué à un échantillon aléatoire de 100 000 domaines extraits du corpus Tranco, qui recense le million de sites web les plus visités. MECILDI révèle que 37 % des sites web sont multilingues, avec un taux de multilinguisme de 3,4. Bien que 67,5 % des sites web offrent une version en anglais, cette langue ne représente que 20,1 % de l'ensemble des pages web de l'échantillon. L'allemand, le français et l'espagnol représentent chacun environ 6 % des pages, et des dizaines d'autres langues contribuent de manière significative au contenu web. Tous les résultats sont calculés avec un intervalle de confiance de 99 % inférieur à 1 %. Ce pourcentage de pages web en anglais représente moins de la moitié des estimations obtenues par des méthodologies monolingues très médiatisées utilisant le même corpus. D'autres études corroborent une prévalence du contenu en anglais sur l'ensemble du web comprise entre 19 % et 26 % des pages. Appliqué à huit domaines génériques de premier niveau (gTLD), géographiques ou linguistiques, en lien avec la France, MECILDI révèle une variation importante des profils de multilinguisme, avec un taux de multilinguisme allant de 1,42 à 6,05. Par ailleurs, certaines langues régionales françaises, comme le breton et le corse, montrent une présence numérique notable, tandis que le domaine .eus affiche un bilinguisme basque - espagnol quasi total. Les analyses de sensibilité des facteurs clefs de la méthode, proposées pour un large éventail d'hypothèses méthodologiques plausibles, confirment la robustesse de la plateforme. Les résultats de MECILDI concordent, en ce qui concerne l'anglais, avec la modélisation prédictive d'OBDILCI des indicateurs de présence linguistique sur le Web, et également, pour plusieurs indicateurs, avec les études antérieures, réalisées par OBDILCI, à partir de la base de données de DataProvider.com. Cette étude confirme également que les indicateurs issus du concept des sites les plus visités, comme Tranco, ne reflètent pas la réalité linguistique de l'ensemble du Web, pour cause d'un biais important à l'encontre des langues non européennes. Le code source et les résultats de MECILDI sont publiques, sous licence CC-BY-SA 4.0. Enfin, les limites de cette approche sont exposées, les biais mis en évidence et une feuille de route pour MECILDI Version 2 est présentée. Cette dernière version qui intégrera un maximum de méthodes de détection, au-delà de *hreflang,* est présentée, ainsi que des stratégies pour étendre les études futures à l'ensemble du web.

Table des matières

Liste des tables et figures

1) Introduction

Les estimations actuelles de la répartition des langues sur le web varient considérablement, principalement en raison des différences méthodologiques concernant la sélection des corpus, les stratégies d'échantillonnage, le traitement des sites web multilingues et le périmètre du contenu mesuré. Si la plupart des études antérieures ont mesuré la prévalence des langues strictement au niveau du site web, rares sont celles qui ont pris en compte le fait qu'un même site puisse héberger plusieurs versions linguistiques complètes. Par conséquent, les estimations globales divergent fortement : certaines études concluent que plus de la moitié du contenu web est en anglais, tandis que d'autres l'estiment à seulement 26 % ou moins. De plus, les détails relatifs aux méthodologies de collecte de données des approches non universitaires existantes, qui représentent la plupart des estimations linguistiques disponibles sur le web, et les plus influentes, ne sont pas accessibles au public.

Pour illustrer comment le fait de ne pas tenir compte du multilinguisme d'un site web peut engendrer une telle divergence, prenons l'exemple d'un site proposant quatre versions linguistiques distinctes et parallèles : anglais, français, espagnol et portugais. En supposant que chaque version localisée soit entièrement traduite, un quart des pages de ce site sont en anglais, un quart en français, et ainsi de suite. Les méthodes d'analyse basées sur une classification monolingue ignorent cette réalité architecturale en attribuant une seule langue principale à l'ensemble du domaine, une simplification qui a un impact direct sur la validité des indicateurs de répartition linguistique sur un échantillon de sites web.

Face à la diversité des résultats et des méthodes employées, dresser un portrait précis du paysage linguistique du web demeure un défi. Ces indicateurs influencent pourtant les décisions des gouvernements, des chercheurs et des organisations privées, concernant les stratégies de localisation, l'accessibilité linguistique et les politiques publiques d'inclusion numérique. À titre d'exemple, Giannakoulopoulos et al. (2020) ont constaté une corrélation négative entre la disponibilité de l'anglais sur les sites web des ccTLDs de l'Union européenne et le PIB d'un pays. En mesurant le nombre exact de versions linguistiques localisées au sein de l'UE, leur étude a mis en évidence la nécessité d'étudier la nature multilingue des sites web.

Cette nécessité est devenue une urgence, face à la croissance rapide du commerce numérique (qui représente déjà 21 % des achats de détail mondiaux depuis 2025), à la proclamation par l'UNESCO de la Décennie internationale des langues autochtones (2022-2032) et à l'essor de l'intelligence artificielle générative qui rassemble des experts du monde entier pour relever de nouveaux défis linguistiques. Par conséquent, le domaine de la métrique des langues numériques doit évoluer et dépasser les modèles monolingues qui ne reflètent plus la réalité linguistique d’une Toile de plus en plus diversifiée linguistiquement.

Il est impératif de déployer une méthode transparente, documentée, dont les biais sont exposés et contrôlés, et dont la fiabilité soit établie selon des critères scientifiques, pour établir, à partir de

données sérieuses, les politiques linguistiques pour les langues dans le numérique et en mesurer les impacts. Cette méthodologie doit étendre son champ d'analyse à un éventail de langues beaucoup plus large, tout en fournissant des indicateurs pertinents du multilinguisme en ligne.

La présente étude introduit la mise en œuvre initiale de la plateforme méthodologique et informatique MECILDI (*MEsure CIblée des Langues Dans l'Internet)* qui mesure la distribution des langues sur n'importe quelle série de sites web, en prenant en compte le multilinguisme potentiel de chaque site.

Afin d'évaluer la validité de cette plateforme et de démontrer son application empirique, nous traitons les questions de recherche fondamentales suivantes :

> Si l'on considère toutes les versions linguistiques disponibles des sites web, l'anglais représente-t-il réellement 50 % des pages web ? Quelles sont les langues les plus représentées au niveau mondial ? Quelles langues dominent au sein d'une région, comme la Corse, ou d'un territoire particulier, comme la Nouvelle-Calédonie ? Quels indicateurs clés permettent de définir et de suivre l'évolution du multilinguisme sur la Toile ? Combien de langues sont actuellement présentes en ligne avec des contenus, et ce nombre est-il en augmentation ? Quelles sont les limites inhérentes à la mise en œuvre actuelle, et où peut-on envisager des améliorations empiriques ?

En nous appuyant sur les modélisations antérieures de Pimienta (2022) et Pimienta et al. (2023), nous avons établi un référentiel théorique pour la répartition linguistique. Selon ces projections, environ 20 % des pages web devraient être en anglais, 19 % en chinois (macro-langue), 7 % en espagnol et environ 3,5 % en hindi, russe, arabe, français et portugais. Afin de vérifier ces hypothèses, nous avons appliqué notre plateforme au corpus Tranco (Le Pochat et al., 2019), ainsi qu'à une sélection de domaines génériques de premier niveau (gTLD) géographiques et linguistiques associés à la France (par exemple, .alsace , .paris , .bzh).

Sur la liste Tranco, la proportion observée de pages web en anglais (20,13 %) se situe dans la fourchette attendue. En revanche, le classement observé des autres langues diffère sensiblement du modèle. L'allemand (6,51 %), le français (6,13 %) et l'espagnol (6,09 %) occupaient les deuxièmes places. Le chinois n'apparait qu'en dixième position, avec seulement 2,92 % des pages web.

Sur l'ensemble de données des gTLD, le français occupe logiquement la première place dans la quasi-totalité des domaines, atteignant un taux de présence de 95 % sur les sites web des domaines actifs de Nouvelle-Calédonie (.nc). Seul le domaine de la Guadeloupe (.gp) présente davantage de contenu en anglais qu'en français, mais avec une différence de moins de 1 %. Dans le gTLD basque (.eus), l'espagnol et le basque sont les langues les plus représentées, avec des pourcentages de pages web presque identiques. Enfin, les langues régionales françaises telles que le breton et le

corse sont présentes sur moins de 4 % des domaines .bzh et .corsica, soulignant les défis persistants liés à la vitalité numérique des langues.

2) Contexte et recherches antérieures

2.1 Perspectives historiques

Parmi les ouvrages fondateurs de la linguistique numérique figure *English as a Global Language* (1997) de David Crystal, qui retrace l'évolution historique de l'anglais dans la société moderne et met en lumière son hégémonie précoce sur l'Internet naissant. Dès les premiers stades du développement du web, Crystal exhortait les chercheurs à aborder les indicateurs de répartition des langues avec prudence. Il soulignait notamment que les affirmations contemporaines – telles que l'idée largement répandue selon laquelle « environ 80 % des informations stockées électroniquement dans le monde sont actuellement en anglais » – devaient être « interprétées avec prudence » (p. 115).

Quelques années plus tard, Grefenstette et Nioche (2000) ont introduit une méthodologie permettant de déterminer la distribution linguistique d'un corpus en mesurant la fréquence des mots-clés spécifiques à chaque langue au sein d'un texte. Appliquant cette technique au moteur de recherche AltaVista, ils ont identifié 48 millions de pages en anglais, suivies de 3,3 millions en allemand, 2,7 millions en français et 1,89 million en espagnol. Point crucial, les auteurs ont constaté qu'AltaVista n'indexait alors qu'environ 16 % du web accessible, ce qui introduisait un biais de sélection important favorisant de manière disproportionnée les sites web commerciaux et américains.

Durant cette même période, de 1996 à 2007, l'OBDILCI avait proposé un suivi de la répartition des langues sur le web, en se concentrant sur les langues romanes, l'allemand et l'anglais. Sa méthodologie reposait sur le traitement statistique du nombre de résultats des moteurs de recherche pour un ensemble de mots-clés soigneusement sélectionnés afin de maintenir l'équivalence sémantique et linguistique entre les langues. Grâce à cette approche, un déclin constant de la part de l'anglais sur le web avait été constaté, passant de 80 % en 1996 à 50 % en 2007. La méthode initiale de l'OBDILCI, ainsi qu'un aperçu complet des méthodologies contemporaines de mesure du web, ont été détaillés dans une publication de l'UNESCO de 2009 (Pimienta et al., 2009). Cependant, cette méthodologie de comptage des résultats est devenue obsolète en raison de l'évolution des algorithmes d'indexation des moteurs de recherche et de la perte de fiabilité des comptages de résultats qui en découle.

2.2 Initiatives et indicateurs de performance contemporains en matière de mesure du Web

Les études universitaires évaluées par les pairs sur la distribution des langues sur le web sont rares. Par conséquent, la plupart des données largement citées sur la présence des langues en ligne proviennent d'initiatives commerciales. Bien qu'utiles pour le suivi technique, ces méthodes

commerciales n'ont pas été conçues pour l'analyse sociolinguistique et leurs mécanismes sous-jacents sont rarement rendus publics.

Parmi ces entreprises, W3Techs se distingue. Depuis 2011, cette entreprise publie des enquêtes quotidiennes sur les technologies web, et complémente avec les langues utilisées dans le contenu web. W3Techs déploie un algorithme propriétaire de détection de la langue sur le corpus Tranco dont les résultats sont souvent repris comme étant valide pour la Toile entière. En 2026, ils ont annoncé que 49,7 % de ces sites web utilisaient l'anglais, suivi de l'espagnol (6,1 %) et de l'allemand (6,0 %). Selon W3Techs, le pourcentage d'anglais est resté relativement stable entre 50 % et 60 % depuis 2011. Le chinois ne figure pas parmi les cinq premières langues ; il se classe en 13e position avec un faible pourcentage de 1,2 %. À part leur recours au corpus Tranco, W3Techs ne fournit aucune documentation publique concernant leurs mécanismes d'échantillonnage ou leur méthodologie de détection de la langue, ce qui rend inconnues les bases techniques précises de ces chiffres.

Déjà citée, l'étude de Giannakoulopoulos et al. (2020) a analysé la prédominance de l'anglais sur environ 100 000 sites web au sein des domaines de premier niveau nationaux (ccTLD) de l'Union européenne. Les auteurs ont constaté que 25,64 % des sites web des États membres non anglophones de l'UE proposaient du contenu en anglais. Pour ce faire, ils ont extrait les ccTLD cibles du corpus Common Crawl, exploré leurs pages d'accueil et leurs sous-pages (un seul niveau de profondeur), puis évalué la disponibilité multilingue en croisant l'attribut structurel <html> Lang= avec une analyse textuelle algorithmique. Plus précisément, ils ont utilisé une bibliothèque N-gram basée sur PHP (Schur, 2025) capable de détecter 110 langues dans des textes dépassant 150 caractères, permettant ainsi à l'étude de documenter de manière fiable les domaines multilingues.

Par la suite, une évaluation à grande échelle menée par Netsweeper (2023) a analysé la répartition linguistique de 12 milliards de pages web individuelles (environ 30% de la taille estimée du Web en termes de pages). Cette étude a révélé que 26,3 % des pages indexées étaient en anglais, suivies du chinois (19,8 %) et de l'espagnol (8,1 %). Cependant, la méthodologie de Netsweeper étant propriétaire et non documentée publiquement, il demeure impossible de reproduire la sélection de leur corpus ou d'évaluer d'éventuels biais d'échantillonnage géographique et mécanismes de détection de la langue. Néanmoins, leurs résultats globaux concordent étroitement avec les données du second modèle d'OBDILCI présenté dans la section 2.4, ce qui suggère que des biais structurels sous-jacents ont pu être partiellement atténués ou pris en compte dans leur processus. Il faut souligner qu'en prenant comme unité d'analyse la page web (et non le site) Netsweeper évite le biais inhérent aux méthodes qui imposent une seule classification linguistique à un site web entier.

Parallèlement, DataProvider.com (2023) a analysé un corpus de 99 millions de domaines actifs, constatant que 51,3 % des sites web indexés étaient en anglais, suivis du chinois (10,3 %) et de l'allemand (7,3 %). Pour analyser cet ensemble de données, ils ont implémenté le détecteur de

langue compact v3 de Google (Google Inc., 2016), basé sur un réseau neuronal, via le « wrapper » de source ouverte *gocld3* (Hodges, 2025). Sur le plan méthodologique, DataProvider.com n'a attribué qu'une seule langue par domaine, ce qui n'intègre pas le multilinguisme intra-site, malgré l'enregistrement des métadonnées *hreflang* . L'enregistrement des attributs *hreflang* sur l'ensemble de leur échantillon fournit néanmoins une métrique structurelle alternative cruciale pour la cartographie, qui demeure très précieuse pour la recherche comparative.

En 2025, DataProvider.com a accordé à OBDILCI un accès gratuit à sa base de données, permettant ainsi la réalisation de sept études approfondies sur le multilinguisme de la Toile (OBDILCI, 2025a), dont les résultats ont depuis été validés par la plateforme MECILDI. La première de ces études (OBDILCI, 2025b) comprenait une analyse reproduisant la méthodologie de Giannakoulopoulos et al. (2020), mais portant sur un échantillon considérablement élargi de 16 millions de sites, contre 100 000 initialement. Cette étude a établi que la proportion de pages web en anglais sur l'ensemble des ccTLD de l'UE avant le Brexit a chuté de 28,4 % en 2019 (un chiffre calculé directement à partir des données de Giannakoulopoulos et al.) à 19,8 % en 2025. De plus, la même étude a montré une croissance significative des indicateurs de multilinguisme de l'échantillon, le taux de multilinguisme passant de 1,14 à 1,83 et la proportion globale de sites multilingues augmentant de 11 % à 17 %.

Pour une analyse complète de ces méthodologies contemporaines et de leurs limites architecturales, voir Pimienta (2025).

2.3 Initiatives académiques antérieures et le manque de métriques

Il convient de noter qu'au début des années 2000, deux initiatives académiques très prometteuses ont émergé parallèlement aux analyses de marché de l'époque. La première était le Language Observatory Project (LOP), lancé en 2003 à l'Université de technologie de Nagaoka au Japon, qui a créé un vaste consortium d'universités internationales. Utilisant un moteur statistique N-gram capable d'identifier 350 langues, le LOP a cartographié avec succès environ 100 millions de pages web sur les ccTLD asiatiques et africains entre 2006 et 2007, recueillant ainsi des données essentielles sur la répartition des langues locales. Parallèlement, en 2003, l'Institut de statistique du gouvernement catalan (IDESCAT) a collaboré avec l’Université polytechnique de Catalogne (UPC) pour élaborer un référentiel linguistique dédié. Le projet UPC/IDESCAT a appliqué des algorithmes de reconnaissance de langue à un sous-ensemble de deux millions de noms de domaine extraits d’un vaste référentiel de 30 millions de domaines, produisant des données jusqu’en 2006, lesquelles ont fourni un point de comparaison essentiel aux études OBDILCI de l’époque.

Ces deux projets universitaires proposaient une alternative spécifiquement adaptée à la recherche publique et scientifique, intervenant à un moment où les données webométriques commerciales – bien qu'utiles pour les applications axées sur le marché – étaient fréquemment mal interprétées ou déformées par des tiers cherchant à répondre à des questions sociolinguistiques plus vastes, pour lesquelles les données commerciales n'ont jamais été conçues. Cependant, ces deux initiatives ont pris fin en 2011, laissant un vide dans l'étude académique des indicateurs linguistiques du contenu

web et du multilinguisme. Si des chercheurs contemporains comme Giannakoulopoulos ont depuis fait progresser la littérature sur la localisation web, leurs travaux se sont principalement concentrés sur des contextes régionaux spécifiques. Par conséquent, les projets d'OBDILCI visent à combler cette lacune en modernisant la méthodologie de la cartographie linguistique à grande échelle.

2.4 Contextualisation de la présente étude

Depuis 2017, OBDILCI développe et maintient un modèle analytique destiné à cartographier le paysage linguistique du web. Ce modèle produit des indicateurs empiriques pour les 362 langues comptant plus d'un million de locuteurs L1 (Pimienta 2022 ; Pimienta et al. 2023). Plutôt que de s'appuyer sur des explorations directes du web, ce modèle, qui a atteint sa maturité méthodologique en 2022, utilise une approche indirecte fondée sur trois grandes catégories de données : des données démolinguistiques issues d'*Ethnologue*, les taux de pénétration d'Internet par pays, fournis par l'Union internationale des télécommunications (UIT), et une vaste série d'indicateurs numériques relatifs aux langues, aux pays et aux internautes.

Le modèle repose sur le postulat économique qu'une relation structurelle unit le contenu numérique dans une langue (l'offre) à la population mondiale de locuteurs L1 +L2 connectés de cette langue (la demande). Cette vision a été inspirée par des années d'observation de terrain du ratio contenu par langue divisé par le nombre d'utilisateurs connectés par langue (productivité du contenu), ratio qui s'est généralement maintenu entre 0,5 et 2. Le modèle approxime cette loi économique sous-jacente en combinant et pondérant judicieusement un ensemble d'indicateurs numériques. Sur le plan méthodologique, les deux premières catégories d'entrées (données démolinguistiques et taux de connexion Internet) établissent une base de référence initiale d'« économie parfaite ». La troisième catégorie d'entrées, plus multifactorielle, calibre ensuite ces projections de référence afin de refléter les conditions économiques réelles, en tenant compte de l'infrastructure technologique supportant les langues et du niveau de préparation global à la société de l'information des pays d'origine des locuteurs.

Structurellement, le modèle repose sur une hypothèse simplificatrice : tous les locuteurs d’une langue donnée, au sein d’un pays spécifique, bénéficient du même taux de connexion Internet. Bien que cette distribution uniforme introduise un biais connu – limitant de fait la fiabilité du modèle aux grandes populations linguistiques (L1 > 1 million) –, elle constitue un cadre de plausibilité mathématiquement rigoureux, assorti d’un intervalle de confiance d’environ 20 %. En définitive, bien qu’il ne s’agisse pas d’une mesure empirique directe, cette approche offre un modèle théorique très inclusif, couvrant un nombre de langues nettement supérieur à celui des méthodes d’exploration web commerciales limitées par le nombre de langues identifiables par les algorithmes.

Outre son rôle de complément au modèle théorique, le programme MECILDI a été conçu avec deux objectifs distincts. Premièrement, il réalise une mesure empirique contrôlée, transparente et entièrement reproductible sur des corpus identiques à ceux analysés par W3Techs, confirmant ainsi le biais mathématique induit par la négligence du multilinguisme sur le Web (Pimienta, 2024).

Deuxièmement, il établit un cadre d'analyse pour des sous-ensembles ciblés du Web (tels que des ccTLD ou gTLD spécifiques) que les modèles macro ne peuvent isoler. L'analyse intermédiaire de la base de données DataProvider.com a constitué une transition providentielle entre le modèle et le projet MECILDI, fournissant des données préalables qui ont permis d'affiner les indicateurs clés de MECILDI pour le multilinguisme sur le Web — notamment *Multi%* , *MultiR* et *AvgL* — et offrant un point de référence pour les estimations actuelles. Ensemble, le modèle théorique, l'analyse de DataProvider.com et le cadre MECILDI offrent une perspective complète sur le multilinguisme web, bien que les deux derniers partagent explicitement une hypothèse fondamentale sous-jacente qui reste à prouver : un taux d'adoption de base estimé à 40 % pour les métadonnées *hreflang*.

2.5 Technologies Web pour l'identification de la langue

D'un point de vue technique, les sites web multilingues peuvent indiquer la disponibilité de plusieurs versions linguistiques grâce à divers mécanismes, standardisés ou non. Parmi les approches courantes, on trouve les annotations de liens HTML *hreflang*, l' attribut *lang* de l'élément <html>, les sitemaps XML spécifiques à chaque langue, les sous-domaines spécifiques à chaque langue (par exemple, fr.example.com et es.example.com), les chemins d'URL spécifiques à chaque langue (par exemple, example.com/fr/ et example.com/es/), les cadres de localisation basés sur JavaScript, les plugiciels multilingues pour systèmes de gestion de contenu, les interfaces de sélection de langue personnalisées et les « widgets » de traduction automatique tels que Google Traduction. Contrairement aux autres mécanismes mis en place par les éditeurs et mentionnés ici, les « widgets » de traduction automatique sont considérés comme une simple aide à l’intercompréhension linguistique et non comme une traduction authentique ; ils ne sont donc pas pris en compte dans les indicateurs de multilinguisme.

La diversité de ces techniques d'implémentation fait qu'aucun indicateur technique unique ne permet d'identifier avec certitude tous les sites web multilingues. Certains signaux, comme les annotations *hreflang* et l'attribut HTML *lang* , sont standardisés et peuvent être détectés de manière fiable par analyse automatisée. D'autres, tels que les structures d'URL, les sous-domaines spécifiques à une langue, la localisation basée sur JavaScript et les sélecteurs de langue personnalisés, sont considérablement plus difficiles à identifier par programmation. Par exemple, la détection de modèles d'URL ou de sous-domaines spécifiques à une langue nécessite l'évaluation de nombreuses combinaisons possibles de codes de langue et peut générer des faux positifs, tandis que les implémentations JavaScript personnalisées ne suivent pas de conventions standardisées et ne peuvent donc pas être détectées de manière fiable par un robot d'exploration web générique.

Par conséquent, la présente étude utilise une méthodologie en deux étapes. Premièrement, un jeu de données de référence annoté manuellement intègre de multiples indicateurs de multilinguisme afin d'établir la prévalence réelle des sites web multilingues et de quantifier les limites des différentes méthodes de détection. Deuxièmement, une analyse informatique à grande échelle s'appuie principalement sur l'annotation *hreflang* standardisée comme indicateur de

multilinguisme, les estimations obtenues étant calibrées ensuite à l'aide d'un facteur d'extrapolation empirique dont les paramètres sont issus de l'étude de référence manuelle.

Note : Alors que ce cadre actuel (MECILDI V1) repose principalement sur les balises *hreflang*, les attributs *lang* et l'extrapolation statistique, le futur MECILDI V2 visera l'exhaustivité méthodologique en capturant tout le spectre des indicateurs techniques multilingues.

2.6 Terminologie et clarifications conceptuelles

Cette section établit la terminologie fondamentale utilisée tout au long de l'étude, en délimitant les frontières du contenu numérique, les contraintes opérationnelles, les formules mathématiques et la portée du corpus.

2.6.1 Définir les limites du contenu numérique

Nous distinguons trois niveaux emboîtés d'information numérique :

- **Contenu numérique (cyberespace) :** Il s'agit de la catégorie la plus vaste, englobant toutes les données stockées sous n'importe quel format numérique, quel que soit leur emplacement. Elle comprend à la fois les données connectées à l'Internet et les informations hors ligne (telles que le stockage local d'un ordinateur, les bases de données d'entreprise non connectées ou les supports de stockage hors ligne). Même hors ligne, une grande partie de ce contenu conserve une voie latente vers le monde en ligne, soit par le biais de téléchargements ultérieurs, d'une intégration via l'Internet des objets, soit par son ingestion dans des corpus utilisés pour l'entraînement de modèles d'intelligence artificielle.
- **Contenu Internet (contenu en ligne) :** ce terme désigne les informations numériques publiées dans l'Internet et accessibles aux utilisateurs via des protocoles de communication courants (par exemple, HTTP/S, SMTP, FTP et SSH). Il comprend les médias en ligne, les courriels, les transferts de fichiers et les journaux de serveurs distants, sous forme de texte, d'image, d'audio et de vidéo.
- **Contenu Web (Toile) :** Sous-ensemble distinct du contenu Internet, défini spécifiquement comme l'information accessible via un navigateur Web. Cela inclut les pages Web standard, les documents intégrés, les applications Web, les services en ligne et les fichiers multimédias hébergés.

Lors de l'analyse des langues dans ces domaines, les chercheurs doivent distinguer la langue des internautes (utilisateurs de l'Internet) de celle du contenu. Les utilisateurs sont par nature dynamiques et souvent multilingues ; chaque personne possède une langue maternelle (L1) et potentiellement une ou plusieurs langues secondaires (L2). L'internaute utilise l'ensemble de son répertoire linguistique (L1 + L2) en fonction de ses besoins numériques. À l'inverse, la langue du contenu Internet désigne strictement la langue spécifique utilisée au sein d'une unité de média numérique isolée (par exemple, une page web, le corps d'un courriel ou une piste audio).

2.6.2 Simplifications et contraintes méthodologiques

Mesurer la répartition mondiale des langues sur le web nécessite plusieurs simplifications pratiques pour garantir la faisabilité informatique et la cohérence empirique :

- **La page web comme unité de granularité :** l'identification de la langue s'effectue fondamentalement au niveau de la page web. Par conséquent, les systèmes de messagerie électronique ne sont évalués que lorsqu'ils sont archivés sur des pages web statiques et indexées publiquement (telles que les forums web publics ou les archives de listes de diffusion).

- **Contenu statique vs. contenu dynamique :** seul le contenu statique des pages web est traité. Le contenu généré dynamiquement par les utilisateurs en temps réel (comme les flux algorithmiques personnalisés) est exclu.

- **Gestion des pages Web multilingues :** lorsqu'une seule page Web contient plusieurs langues, l'analyse syntaxique pratique exige que seule la langue dominante principale, déterminée soit par l'attribut HTML lang, soit par la classification dominante de l'algorithme de détection de la langue, soit traitée.

- **Dissocier le site de l'utilisateur :** Si l'on considère généralement qu'un utilisateur humain possède une langue première (L1) et plusieurs langues secondes (L2), ce concept ne s'applique pas aux sites web. Un site multilingue présente des configurations linguistiques distinctes, dont aucune ne peut être considérée comme la langue « principale ». De fait le choix de langue se fait dynamiquement, soit implicitement par sélection de la langue de travail du système opérationnel du mobile ou de l'ordinateur de l'utilisateur, soit explicitement par un menu à travers lequel l'utilisateur exprime son choix de langue ou de pays, ce qui détermine la langue. Les pages d'accueil affichant souvent un texte d'introduction en plusieurs langues, les utiliser comme indicateur de « langue principale » d'un site entraîne des risques sérieux d´erreur de classification. Par conséquent, chaque version linguistique d'un site web doit être traitée comme une entité équivalente et indépendante, même si certaines versions sont plus nourries que d'autres.

- **Pondération du volume :** Dans cette version du programme MECILDI (V1), nous utilisons une hypothèse simplificatrice standard selon laquelle toutes les configurations linguistiques d'un site web sont considérées comme structurellement équivalentes. Ainsi, une version localisée de 10 pages d'un site a le même poids que son homologue de 10 000 pages. La mise en œuvre d'une moyenne pondérée basée sur le nombre de pages demeure un objectif prioritaire de MECILDI V2.

2.6.3 Métriques opérationnelles et formules

Afin de garantir la cohérence tout au long de cette étude, nous définissons et appliquons les termes et indicateurs clés suivants :

- **Domaine vs. Site web :** Un nom de domaine enregistré ne garantit pas un site web actif ; de nombreux domaines sont en attente, réservés ou renvoient des erreurs de serveur. La plateforme MECILDI ne prend pas en compte les domaines inactifs ou ceux présentant des erreurs de connexion fatales empêchant l'affichage dans le navigateur.
- **Série (de sites web) :** Liste ciblée et active de sites web sélectionnés pour la mesure. Une série peut englober une zone de domaine de premier niveau entière (par exemple, tous les domaines d'un gTLD ou d'un ccTLD), un corpus complet ou un sous-ensemble d'un corpus (par exemple, le million de domaines de la liste Tranco). Le système exclut automatiquement les domaines enregistrés inactifs ou renvoyant des erreurs de connexion fatales au sein d'une série.
- **%PAGES(L) :** Pourcentage de pages web rédigées en langue *L* au sein d'une série de sites web donnée. La somme de cette métrique pour toutes les langues est égale à 100 %. Par simplification, ce pourcentage est calculé en divisant le nombre de versions linguistiques identifiées d'un site par le nombre total de versions linguistiques de la série.

$$\%PAGES(L) = \frac{Nombre\ de\ versions\ linguistiques\ dans\ la\ langue\ L}{Nombre\ total\ de\ versions\ linguistiques\ de\ la\ série}$$

- **%SITES(L) :** Le pourcentage de sites web au sein d'une série qui hébergent une configuration linguistique dans la langue *L.* Étant donné qu'un seul site peut prendre en charge plusieurs langues, la somme cumulée de ces pourcentages dans toutes les langues dépassera généralement 100 %.

$$\%SITES(L) = \frac{Nombre\ de\ versions\ linguistiques\ en\ langue\ L}{Nombre\ total\ de\ sites\ web\ validés\ de\ la\ série}$$

- **Nombre moyen de langues par site web multilingue (AvgL) :** Le nombre total de versions linguistiques de tous les sites web multilingues divisé par le nombre de sites web multilingues :

$$AvgL = \frac{Nombre\ total\ de\ versions\ linguistiques\ de\ tous\ les\ sites\ multilingues}{Nombre\ total\ de\ sites\ multilingues}$$

- **Pourcentage de sites web multilingues (%Multi) :** Le pourcentage de tous les sites web qui ont été détectés comme multilingues :

$$\%Multi = \frac{Nombre\ total\ de\ sites\ multilingues}{Nombre\ total\ de\ sites}$$

- **Taux de multilinguisme (MultiR) :** nombre moyen de versions linguistiques prises en charge par site web au sein de la série considérée. Ce taux est calculé en additionnant toutes

les valeurs %SITES(L), ou, de manière équivalente, en divisant le nombre total de configurations linguistiques par le nombre total de sites web dans la série.

$$MultiR = \frac{Nombre\ total\ de\ versions\ linguistiques}{Nombre\ total\ de\ sites}$$

Le taux de multilinguisme peut également être exprimé en fonction de $AvgL$ et $\%Multi$:

$$MultiR = (1 - \%Multi) + AvgL \times \%Multi$$

Pour illustrer ce propos : supposons qu'un corpus contienne deux sites web. Le site A propose du contenu en anglais et en français ; le site B propose du contenu uniquement en allemand. Dans ce corpus, %SITES(anglais) = 50 % (un seul des deux sites web possède une version anglaise), %PAGES(anglais) = 33 % (une seule des trois versions linguistiques est en anglais), et le MultiR est de 1,5 (trois versions linguistiques réparties sur les deux sites web).

Remarque : Bien que le %SITES(L) soit la métrique généralement utilisée dans la plupart des initiatives de mesure du Web, le %PAGES(L) offre une représentation bien plus adéquate de la véritable répartition linguistique en ligne. Cependant, la mesure au niveau de la page est souvent méthodologiquement irréalisable en raison du volume considérable de données : alors qu'il existe environ 200 millions de sites Web actifs dans le monde, le Web indexable comprend environ 40 milliards de pages Web. Bien que la plateforme MECILDI mesure directement le %PAGES(L), elle repose sur un ensemble spécifique d'hypothèses opérationnelles détaillées ci-dessous.

2.6.4 Exclusions du champ d'application

Cette plateforme ne mesure pas :

1. **Contenu d'interface dynamique :** éléments d'interface utilisateur localisés et spécifiques à l'utilisateur, mis à jour en temps réel.
2. **Flux de médias sociaux générés par les utilisateurs :** Bien que les pages d'interface administrative, d'accueil et statiques d'une plateforme de réseau social soient analysées, le flux continu de publications des utilisateurs est exclu.
3. **Canaux de communication privés :** courriels personnels, messages directs et protocoles de communication sécurisés.
4. **Diffusion de données audio et vidéo en continu :** la distribution linguistique au sein des pistes audio, des fichiers vidéo ou des flux en direct est exclue, même si l'hébergement se fait sur des plateformes web standard.

Exemples d'application de la portée :

- **Cas 1 (Facebook.com) :** MECILDI prend en compte et mesure toutes les langues d'interface officielles prises en charge et servies par facebook.com, mais il n'indexe ni

n'évalue la distribution linguistique des publications individuelles générées par les utilisateurs.

- **Cas 2 (YouTube.com) :** La plateforme mesure les différentes configurations d'interface localisées prises en charge par youtube.com, mais elle n'analyse ni ne compte les langues parlées dans les fichiers vidéo ou audio hébergés par youtube.com.

3) Méthodologie

L'ensemble des données brutes d'évaluation au niveau du domaine, utilisées pour étayer les résultats de cette étude, sont archivées de manière permanente sur Zenodo (DOI: 10.5281/zenodo.21878762). Des listes de références et des tables de correspondance supplémentaires (correspondances des codes ISO, termes de détection des sites « en construction » et listes de fournisseurs d'hébergement) sont disponibles en annexe du manuscrit (fichier des documents supplémentaires).

3.1 Établissement d'une base de référence

Afin d'établir une base de référence empirique, nous avons examiné manuellement un échantillon aléatoire de 500 domaines extraits du corpus Tranco. Cet ensemble de données de référence, constitué manuellement, a permis d'éviter le recours à des méthodes de classification automatisées, bien qu'il reste soumis aux caractéristiques d'échantillonnage géographique des listes Tranco sous-jacentes.

Pour chaque domaine, nous avons relevé les caractéristiques multilingues, notamment le nombre de balises *hreflang* , la valeur de l'attribut *lang* dans l'élément <html>, la présence de contenu en anglais, l'intégration du « widget » Google Traduction et la disponibilité de menus de navigation localisés. Ces observations ont été compilées dans un tableur structuré pour analyse.

L'annotation manuelle initiale a été réalisée en 2023 selon un protocole de codage prédéfini. Lors de la phase de validation en 2025, quatre divergences entre les annotations manuelles et les classifications MECILDI ultérieures ont été relevées. Après examen indépendant, il a été établi que ces quatre divergences étaient dues à des variations temporelles attendues dans la configuration des sites web sur une période de deux ans, et non à des erreurs de codage humain, confirmant ainsi la fiabilité de l'annotation manuelle initiale avant les analyses finales.

L'audit manuel de ces 500 sites Web a révélé un %PAGES(anglais) de 29 % et un taux de multilinguisme (MultiR) de 2,23 (Pimienta, 2024), en supposant des versions linguistiques entièrement parallèles sans tenir compte des variations de taille du site Web.

L'étude a également révélé que seulement 40 % des sites web multilingues confirmés utilisaient les balises *hreflang*. Puisque l'analyse à grande échelle identifie le multilinguisme uniquement par le biais de *hreflang*, ce taux de détection de 40 % a servi à calculer ce que nous appellerons désormais le facteur d'extrapolation (FE). L'application de ce facteur permet d'ajuster le $FE =$

$1/0.40 = 2.5$. L'application de ce facteur d'extrapolation sur le multilinguisme observé, basé sur *hreflang*, permet une approximation de la prévalence réelle des sites web multilingues, dans l'hypothèse où la relation entre l'utilisation de *hreflang* et le multilinguisme effectif dans l'échantillon manuel serait représentatif du corpus global. Pour garantir la robustesse de ce paramètre, la valeur de référence de 40 % a été vérifiée et validée à l'aide de requêtes indépendantes soumises à plusieurs modèles de langues, formulées de manière neutre afin d'éviter toute suggestion susceptible de provoquer un biais de confirmation.

Cette calibration part du principe que les taux d'adoption *de hreflang* sont relativement stables selon les catégories de sites web, les régions géographiques et les niveaux de popularité. De même, l'extrapolation repose sur l'hypothèse que les sites multilingues non détectés partagent exactement la même distribution linguistique que ceux analysés avec succès via *hreflang*, indépendamment des mécanismes de localisation technique alternatifs qu'ils pourraient utiliser. Ces deux hypothèses introduisent des biais potentiels dans la méthodologie actuelle, qui sont examinés plus en détail ci-dessous.

3.2 Sélection du corpus

Après avoir établi la base de référence, nous avons préparé les jeux de données utilisés pour l'analyse informatique. Afin d'assurer la comparabilité avec l'étude W3Techs et l'analyse de référence, nous avons adopté le même corpus source : le jeu de données standard proposé par Tranco.

Le choix d'une approche par échantillonnage statistique plutôt que le traitement de l'ensemble des domaines a été motivé par les contraintes financières et techniques d'une petite ONG opérant dans un environnement d'hébergement mutualisé aux ressources limitées. Un échantillon de 100 000 domaines (répartis en 100 séries indépendantes de 1 000 sites web) a été déterminé afin d'obtenir des résultats fiables grâce aux méthodes statistiques standardisées pour les distributions aléatoires, notamment pour les langues dont la proportion est supérieure à 0,01 %. Par conséquent, si la détection des langues en dessous du seuil de 0,01 % reste précise, leur représentation globale au sein des sous-ensembles fluctue. Plutôt que de refléter un manque de fiabilité statistique, ces langues à faible densité apparaissent de manière intermittente et aléatoire (de 1 à 5 fois au sein d'une même série de 1 000 sites web), ce qui explique la sélection différente de langues minoritaires recensées dans chaque série.

Pour la partie de l'étude portant sur les domaines de premier niveau génériques et de pays associés à la France, des corpus distincts ont été constitués pour les domaines suivants : .alsace , .bzh (breton), .corsica , .eus (basque), .gp (Guadeloupe), .nc (Nouvelle-Calédonie), .paris et .quebec . Les fichiers de zone provenaient principalement du Service centralisé de données de zone (CZDS) de l'ICANN. Le service commercial ZoneStats.io a été utilisé comme alternative lorsque l'accès direct aux fichiers de zone au niveau du registre était impossible. Chaque corpus contenant moins de 100 000 domaines, il a été traité comme un ensemble de données agrégées unique, plutôt que

subdivisé en sous-ensembles de 1 000 domaines. En effet, l'analyse de la totalité des éléments de la série rend inutile le recours à un échantillonnage statistique.

Après avoir évalué plusieurs algorithmes de détection de langue, nous avons sélectionné le moteur MachineTranslation.com (de Tomedes), qui prend en charge 216 langues, pour l'identification linguistique. Les identifiants de langue ont ensuite été associés à leurs codes ISO 639-1 à deux lettres correspondants, lorsqu'ils étaient disponibles. La liste complète des langues et des codes, y compris ISO 639-3, est fournie dans les documents supplémentaires, section S1.

3.3 Infrastructure technique et configuration du robot d'exploration

Pour mesurer informatiquement les métadonnées linguistiques de chaque ensemble de données, un système dédié a été mis en œuvre sur un serveur Linux partagé, développé en PHP et intégrant à la fois une interface web et une chaîne de traitement côté serveur.

L'interface web permet le téléchargement de fichiers d'entrée en texte brut, le lancement d'un traitement en arrière-plan, la surveillance en temps réel via une interface de journalisation, l'interruption et la reprise de l'exécution, ainsi que l'exportation des résultats finaux.

Le robot d'exploration a été configuré avec une chaîne agent-utilisateur personnalisée conçue pour identifier clairement l'objectif de la recherche et fournir une URL de référence :

Mozilla/5.0 (compatible ; MecildiLanguageDetector /1.0 ; +https://www.obdilci.org/projects/mecildi/detector/)

Une fois lancé, le processus en arrière-plan a ingéré séquentiellement les fichiers d'entrée et traité les domaines par lots de 10 afin de respecter les limites des ressources du serveur et les seuils de sécurité standard. Le traitement d'un corpus complet de 100 000 domaines prend un peu plus de 21 heures.

Pour chaque domaine, le système fait l’essai de jusqu'à quatre requêtes, par ordre de préférence de protocole et de sous-domaine (*https://www., https://, http://www., http://*), et sélectionne la première requête renvoyant une réponse HTTP 200 ou 404 comme point de terminaison canonique. Le fichier robots.txt récupéré, lorsqu'il est présent, est analysé et stocké dans la base de données. Un cache local des réponses robots.txt est maintenu pendant sept jours. Chaque ensemble de données n'étant traité qu'une seule fois, les requêtes répétées pour un même domaine sont rares et ne se produisent que si un domaine apparait dans plusieurs échantillons Tranco aléatoires. Le robot d'exploration respecte les bonnes pratiques d'exploration éthiques en se conformant aux directives spécifiées dans robots.txt, et les domaines explicitement interdits par ce fichier ne sont pas consultés.

Après résolution du protocole, et lorsque l'accès n'est pas bloqué par le fichier robots.txt, le système émet une requête pour récupérer la page d'accueil du domaine. Les résultats de la récupération (y compris les erreurs de contenu et de réseau) sont enregistrés, et les métadonnées multilingues sont extraites comme décrit dans la section suivante. Les résultats sont stockés au format JSON comme

des fichiers texte, puis convertis en feuilles de calcul Excel, compressés dans des archives ZIP et mis à disposition pour téléchargement.

3.4 Gestion des erreurs

Lors de la récupération d'une page d'accueil, le robot d'exploration peut rencontrer plusieurs situations : délais d'attente dépassés, échecs de résolution DNS, réponses d'erreur HTTP, redirections HTTP ou récupération réussie du contenu. Les requêtes dépassant un délai d'attente de 10 secondes sont enregistrées comme des erreurs de délai d'attente. Les échecs de résolution DNS et les domaines non enregistrés sont également conservés comme des erreurs DNS, tandis que les réponses HTTP avec des codes d'état d'erreur sont consignées à l'aide de leurs codes d'état correspondants. En cas de réponse de redirection, le robot suit automatiquement la chaîne de redirection jusqu'à la destination finale.

Seuls les contenus récupérés avec succès font l'objet d'une analyse linguistique. Avant traitement, chaque page est évaluée à l'aide d'une série de contrôles de qualité heuristiques visant à exclure les pages ne présentant pas de contenu pertinent. Les pages sont exclues de l'analyse si elles répondent à l'une des conditions suivantes :

1. **Réponses vides**, quand une requête HTTP a abouti mais aucun contenu de page n'a été renvoyé.
2. **Pages de challenge du pare-feu d'application Web (WAF)**, identifiées grâce à des signatures caractéristiques générées par des services tels que Cloudflare, BitNinja et Sucuri.
3. **Pages d'espace réservé DNS**, indiquant qu'un domaine a été résolu avec succès mais n'a pas été configuré avec un contenu de site Web valide.
4. **Pages par défaut du serveur web**, telles que les pages d'index par défaut générées par les installations Apache, cPanel ou Plesk.
5. **Pages de stationnement ou de monétisation de domaines**, y compris les pages faisant la publicité de domaines à vendre ou affichant du contenu publicitaire généré par le registraire.
6. **Pages en construction**, identifiées à l'aide d'une bibliothèque multilingue de phrases génériques courantes.
7. **Les pages d'espace réservé courtes**, définies comme des pages contenant moins de 250 caractères de texte visible, accompagnés d'un indicateur « en construction », parce qu’elles ne permettent pas un traitement valide par le détecteur de langue.

Les bibliothèques d'expressions utilisées pour identifier les pages de parking de domaines et les pages en construction sont gérées comme des fichiers de configuration externes, permettant ainsi la mise à jour des règles de détection indépendamment du robot d'exploration. Pour une liste complète des expressions de recherche relatives aux pages en construction et aux fournisseurs d'hébergement web, voir les sections S2 et S3 des documents complémentaires.

Tous les résultats d'erreur sont consignés à l'aide des abréviations définies dans le tableau suivant. Notez que ERR-CFG englobe les conditions 3 à 5 ci-dessus, ERR-BLK englobe la condition 2 ainsi que les rejets côté serveur, et ERR-HTTP inclut les réponses vides et autres rejets côté serveur.

Table 1 : Abréviations des codes d'erreur

Abréviation	Signification
ERR-DNS	Domaine inaccessible (délai d'attente dépassé, non enregistré)
ERR-BLK	Accès bloqué (WAF, robots.txt, rejet du serveur)
ERR-CFG	Domaine non configuré (en attente, en construction)
ERR-HTTP	Erreur HTTP du site web (301, 503, etc.)
ERR-LANG	Échec de la détection de la langue
ERR-TOT	Taux d'erreur total

La robustesse des mécanismes de détection d'erreurs, notamment l'identification des sites web en construction, est essentielle à la précision des résultats. Dans la grande majorité des cas, ces pages sont susceptibles d'être classées à tort comme des sites monolingues en anglais. Une grande partie de ces sites en construction contiennent des messages d'accueil par défaut en anglais, ce qui, dans certains domaines génériques de premier niveau géographiques (gTLD) français, a eu un impact disproportionné sur les données initiales, gonflant artificiellement les pourcentages de sites en anglais. Il a donc fallu mettre en place un système de filtrage pour isoler et corriger ces cas spécifiques.

3.5 Extraction des métadonnées linguistiques

Une fois la page web récupérée validée par les filtres de gestion des erreurs décrits dans la section précédente, le système extrait les métadonnées linguistiques à l'aide d'une stratégie de détection hiérarchique. Le flux de travail est résumé ci-dessous.

1. Toutes les valeurs *hreflang s*ont extraites du document.
2. Les valeurs extraites sont normalisées en ne conservant que leurs deux premiers caractères (code de langue ISO 639-1), en excluant la valeur spéciale *x-default* et en supprimant les entrées en double.
3. Si au moins deux valeurs *hreflang* uniques subsistent après la normalisation, le site web est classé comme multilingue, les codes de langue détectés sont enregistrés et le traitement pour ce domaine est conclu.
4. Si aucune valeur *hreflang* n’est présente ou bien seulement une valeur unique, il est procédé aux indicateurs heuristiques décrits ci-dessous.
5. Les valeurs de tous les attributs *Lang= s*ont extraites et normalisées selon la même procédure.

6. La présence d'un « widget » Google Traduction est recherché.
7. Des structures d'URL et des sous-domaines spécifiques à la langue (par exemple, example.com/fr/ ou fr.example.com) sont recherchées.
8. Des implémentations JavaScript personnalisées susceptibles d'indiquer une fonctionnalité de changement de langue multilingue sont recherchées.
9. Avant la détection de la langue, les éléments HTML non liés au contenu (<script>, <style>, <noscript> , <header>, <footer> et <nav>) sont supprimés du document.
10. Les 1 500 premiers caractères du texte visible restant sont soumis au service de détection de langue MachineTranslation.com (Tomedes). Les pages contenant moins de 250 caractères visibles sont exclues de la détection de langue car elles ne fournissent pas suffisamment de texte pour une classification fiable.
11. Toutes les métadonnées extraites sont enregistrées au format Excel pour une analyse ultérieure.

Les indicateurs heuristiques évalués aux étapes 6 à 8 ont été inclus car de nombreux sites web multilingues n'utilisent pas d'annotations *hreflang* standardisées. Ces indicateurs sont pondérés afin de calculer un score heuristique de multilinguisme. Ce score est conservé uniquement à des fins de validation et n'impacte pas les analyses statistiques finales, lesquelles reposent exclusivement sur des annotations *hreflang* standardisées, ajustées à l'aide du facteur d'extrapolation décrit dans la section 3.1.

Note : Les données relatives à l'existence de noms de domaine internationalisés (IDN) sont également comptabilisées.

3.6 Attribution de la langue d'un site monolingue

Pour les sites multilingues, les langues sont déterminées directement à partir des valeurs de la balise *hreflang*. Pour les sites monolingues, la langue est attribuée à partir du premier attribut *lang* trouvé dans le code HTML. Si aucun attribut *lang* n'est présent, le résultat de l'algorithme de détection de la langue est utilisé.

Lorsque l'algorithme de détection de la langue diverge de la valeur de l'attribut *`lang`*, son résultat prévaut et l'écart est consigné. Cette priorité tient compte du fait que les valeurs de l'attribut *`lang`* contiennent fréquemment des erreurs dues à une saisie manuelle incorrecte.

3.7 Disponibilité et reproductibilité du code

Cette étude présente les résultats obtenus avec MECILDI v1.8. Afin de garantir leur reproductibilité à long terme, le code source, les fichiers de configuration et les jeux de données de référence utilisés pour cette analyse sont archivés de façon permanente sur Zenodo. Les chercheurs souhaitant reproduire ou vérifier ces résultats doivent utiliser la version archivée v1.8, car les versions ultérieures du dépôt actif pourraient introduire des modifications méthodologiques susceptibles d'altérer les résultats.

- Code archivé (v1.8) : https://doi.org/10.5281/zenodo.21212299
- Dépôt actif : https://github.com/mikefieldenterprises/mecildi

3.8 Validation de la plateforme et évaluation comparative des outils

Pour valider la plateforme, MECILDI a d'abord été exécuté sur l'ensemble de données de référence annoté manuellement, composé de 500 domaines et décrit dans la section 3.1.

Table 2 : Performances de validation de MECILDI par rapport à un corpus de référence de 500 domaines

Métrique	Valeur
Précision	100%
Rappel	43,8%
Précision	86,6%
Score F1	60,9%

La précision parfaite indique l'absence de faux positifs dans les classifications multilingues. Le taux de rappel de 43,8 % s'explique par le fait que de nombreux sites web multilingues n'utilisent pas d'annotations *hreflang* standardisées et ne sont donc pas détectés par le système dans sa configuration actuelle ; il s'agit d'une limitation connue, prise en compte par le facteur d'extrapolation décrit dans la section 3.9 et détaillé dans la section 5.2.

3.9 Cadre mathématique pour l'extrapolation

Cette section établit le cadre mathématique utilisé pour extrapoler les données observées et compenser la sous-détection des sites web multilingues. S'appuyant sur les formules définies dans la section 2.6.3, ces algorithmes ajustent les distributions linguistiques observées afin de se rapprocher de la réalité linguistique réelle du corpus.

3.9.1 Facteur d'extrapolation (EF) :

Comme indiqué dans la section 3.1, les analyses informatiques de cette étude ont utilisé un facteur d'extrapolation de 2,5, initialement calculé à partir du constat de base selon lequel environ 40 % des sites web multilingues affichent des annotations *hreflang* standardisées. Lors de la validation, la correction de quatre erreurs d'annotation manuelle a permis d'augmenter le rappel estimé de MECILDI à 43,8 %. Cette modification étant mineure et toutes -les analyses à grande échelle ayant déjà été réalisées avec la valeur initiale, le FE a été conservé à 2,5 pour tous les résultats présentés.

Note : En principe, si tous les sites web multilingues pouvaient être détectés avec une précision parfaite grâce à tous les signaux techniques disponibles, aucune extrapolation ne serait nécessaire et les valeurs observées refléteraient directement le multilinguisme réel. Les versions futures de

MECILDI intégreront un plus large éventail de signaux de multilinguisme, ce qui rendra à terme l'extrapolation superflue. Cette transition offrira également une précieuse opportunité d'évaluer rétrospectivement la précision et la validité de la méthodologie d'extrapolation actuelle.

3.9.2 Valeurs observées vs. valeurs extrapolées :

MECILDI produit deux ensembles de valeurs pour chaque langue : *les valeurs observée* , dérivées directement du corpus mesuré, et *les valeurs extrapolées*, calculées à l'aide des formules décrites ci-dessous. Nos analyses principales reposant sur ces projections modélisées, toutes les valeurs présentées dans cette étude sont des valeurs extrapolées, sauf indication contraire.

3.9.3 Trois niveaux de calculs

Pour effectuer cet ajustement avec précision sur de vastes corpus de sites web, MECILDI traite les données sur trois niveaux de calcul distincts :

- **Niveau du domaine :** Les données structurelles de base capturées à partir de sites web individuels.
- **Niveau de sous-échantillon (S_i) :** Fichiers discrets contenant des lots de 1 000 domaines.
- **Niveau de synthèse :** Le traitement agrégé des 100 sous-échantillons (représentant l'espace de domaine complet de 100 000) pour calculer les moyennes finales, les écarts types et un intervalle de confiance à 99 %.

Dans chaque sous-échantillon de 1 000 sites (S_i), les occurrences de langue sont calculées séparément pour les sites monolingues (M_o) et les sites multilingues observés (M_u, définis comme des sites avec une présence *hreflang* valide indiquant plus d'une langue).

3.9.4 Extrapolation spécifique à la langue

Pour obtenir la véritable prévalence estimée d'une langue spécifique L au sein d'un sous-échantillon, la plateforme recalcule les distributions monolingues et multilingues observées par rapport aux proportions globales ajustées imposées par le facteur d'extrapolation ($FE = 2{,}5$).

Tout d'abord, les proportions de référence observées de sites monolingues (Mono_L) et multilingues (Multi_L) sont définies par la présence d'indicateurs *hreflang valides* (% H) :

$$\%H = \frac{Nombre\ total\ de\ sites\ avec\ hreflang\ spécifié\ pour\ plus\ d'une\ langue}{Nombre\ total\ de\ sites\ traités\ sans\ erreur}$$

et de ce fait,

$$Multi_L = \%H$$

$$Mono_L = 1 - \%H$$

En appliquant le facteur d'extrapolation (EF), les proportions cibles ajustées ($Mono'_L$et $Multi'_L$) sont établies de telle sorte que :

$$Multi'_L = \%H \times FE$$

$$Mono'_L = 1 - (\%H \times FE)$$

En utilisant ces contraintes, la valeur extrapolée finale des occurrences linguistiques (M_e) pour toute langue *L donnée* est calculée comme suit :

$$M_e = M_o \times \frac{1 - (\%H \times FE)}{1 - \%H} + (M_u \times FE)$$

3.9.5 Vérification croisée des matrices et détection des erreurs

L'agrégation de ces calculs sur plus de 230 langues distinctes pouvant engendrer de subtiles différences d'arrondi ou des désalignements de matrices, le cadre applique une vérification croisée par double équation au niveau du sous-échantillon de 1 000 sites. Le nombre total de versions linguistiques extrapolées pour l'ensemble du sous-échantillon T_e et calculé par deux méthodes indépendantes.

Voie A : La somme des langues agrégées

Le total standard est obtenu en additionnant directement les occurrences extrapolées (M_e) de toutes les langues individuelles :

$$T_e = \sum_L M_e(L)$$

Voie B : Vérification croisée de la matrice structurelle

Simultanément, une formule de vérification croisée directe calcule T_eà partir des totaux bruts des sites monolingues observés (T_o) et des sites multilingues observés (T_u) au sein du lot de 1 000 sites :

$$T_e = T_o + (FE \times T_u) - \%H \times (FE - 1)$$

Note méthodologique : La soustraction de *% H × (FE – 1)* est structurellement nécessaire pour éviter un double comptage. Ce terme tient compte des sites de référence initialement classés comme monolingues, mais convertis en sites multilingues par le processus d'extrapolation, et qui sont donc déjà pris en compte dans le bloc d'expansion $(FE \times T_u)$.

Lors de l'exécution, si la valeur T_e dérivée de la voie A ne correspond pas à la valeur de la voie B jusqu'à la sixième décimale, il est signalé une incohérence de matrice dans les paramètres de la feuille de calcul, servant de contrôle automatisé d'intégrité des données. Cette technique a joué un rôle essentiel dans la phase de test de la plateforme permettant une détection rapide des erreurs et offrant une validation immédiate des corrections.

3.9.6 Proportions normalisées finales

Une fois que les valeurs M_e et T_e sont vérifiées et validées pour un sous-échantillon, les métriques standardisées finales pour *%PAGES* et *%SITES* pour chaque langue sont obtenues :

La proportion de toutes les versions linguistiques appartenant à une langue L est obtenue en divisant son occurrence extrapolée par le nombre total de versions linguistiques validées :

Pourcentage de pages Web en langue L : $\%PAGES(L) = \frac{M_e}{T_e}$

La proportion de sites Web contenant la langue *L* est obtenue en divisant son occurrence extrapolée par le nombre total de sites accessibles et sans erreur ($N_{sites} \leq 1000$) au sein du sous-échantillon :

Pourcentage de sites web ayant une version en langue L : $\%SITES(L) = \frac{M_e}{N_{sites}}$

Par construction, la somme de tous *les %PAGES* dans le corpus est exactement égale à 100%, tandis que la somme de tous *les %SITES* donne 100% × MultiR, préservant parfaitement le taux structurel du multilinguisme, sauf, là aussi, des différences d'arrondi entre les deux modes de calcul.

3.9.7 Sites Web inaccessibles et sites Web comportant des erreurs :

Dans tous les calculs ci-dessus, les sites web inaccessibles et ceux présentant des erreurs ont été exclus. Par exemple, sur les 100 000 sites web analysés, si 40 000 étaient inaccessibles, le nombre total de sites web serait de 60 000 (et non de 100 000).

3.10 Score heuristique de multilinguisme

Pour les sites web qui ne présentent pas d'annotations *hreflang* standardisées, le système calcule un score heuristique de multilinguisme comme indicateur auxiliaire du multilinguisme probable.

Ce score a été implémenté comme un outil heuristique conçu strictement pour valider la méthodologie d'extrapolation et son facteur d'extrapolation (FE) correspondant de 2,5. Utilisé exclusivement lors du développement et de la validation du système, cette métrique n'a pas été intégrée dans les estimations finales de la distribution des langues, qui reposent uniquement sur les annotations *hreflang* extrapolées.

Le score est calculé comme la somme des pondérations heuristiques sélectionnées empiriquement et associées à chaque indicateur détecté, plafonnée à une valeur maximale de 1. Les quatre indicateurs et leurs pondérations respectives sont présentés dans le tableau suivant.

Table 3 : Indicateurs heuristiques du multilinguisme

Indicateur	Heuristique Poids	Raisonnement
Chemins d'URL ou sous-domaines spécifiques à une langue	0,5	Probabilité forte de l'existence de versions localisées du site web (par exemple, /fr/, fr.example.com)
Plusieurs valeurs *de lang*= distinctes	0,4	Probabilité forte de plusieurs versions linguistiques sur un seul domaine
Menu de sélection de la langue	0,08	Probabilité faible ; taux de faux positifs élevés lors des tests préliminaires
Localisation JavaScript indicateurs	0,07	Probabilité faible ; spécifiques à la mise en œuvre et sujettes aux faux positifs

Ces pondérations ont été calibrées spécifiquement pour correspondre à l'échantillon de référence, ce qui a permis d'obtenir un score de référence de 2,5. Par conséquent, si le jeu de données Tranco, plus vaste, produit un facteur d'extrapolation heuristique (voir ci-dessous) proche de 2,5, cela validera à la fois notre facteur d'extrapolation de 40 % (FE = 2,5) et notre hypothèse sous-jacente concernant l'invariance des distributions linguistiques.

Voici quelques exemples :

- Indicateur d'URL uniquement → 0,50
- URL + attributs *de langue multiples → 0,90*
- Les quatre indicateurs → 1,00 (le total des quatre est de 1,05, mais plafonné à 1,00)

3.11 Facteur d'extrapolation heuristique

Pour valider le facteur d'extrapolation supposé de 2,5, nous avons élaboré une estimation ascendante complémentaire basée sur les scores heuristiques de multilinguisme calculés pour chaque domaine. Alors que le facteur d'extrapolation de 2,5 est une estimation descendante dérivée des performances de rappel de MECILDI sur le corpus de référence de 500 domaines, le facteur d'extrapolation heuristique (HDEF) est calculé indépendamment à partir des scores heuristiques observés lors de l'exploration à grande échelle. La concordance entre les deux estimations confirme la pertinence du facteur d'extrapolation supposé pour un corpus donné.

Le HDEF d'un ensemble de données S est défini comme suit :

$$HDEF(S) = \frac{FE^2 \times N_m}{\sum_{w \in S} H(w)}$$

où:

- $HDEF(S)$ est le facteur d'extrapolation dérivé de l'heuristique pour la série S
- FE est le facteur d'extrapolation de base (2,5)

- N_m est le nombre total de sites web multilingues confirmés dans la série S
- $\sum_{w \in S} H(w)$ est la somme des scores heuristiques de multilinguisme sur l'ensemble des sites web dans la série S

Si le HDEF d'un corpus particulier est proche de 2,5, cela confirme le choix du facteur d'extrapolation pour ce corpus. Des valeurs sensiblement supérieures ou inférieures à 2,5 indiquent que le niveau réel de multilinguisme dans ce corpus diffère de l'hypothèse de base : soit parce que le multilinguisme est plus fréquent (HDEF > 2,5), soit parce qu'il est moins fréquent (HDEF < 2,5).

3.12 Études de sensibilité

Trois analyses de sensibilité ont été réalisées afin d'évaluer la robustesse des principaux résultats face aux hypothèses méthodologiques clés.

En-tête Accept-Language

Certains sites web proposent un contenu différent selon des signaux contextuels tels que la langue préférée du navigateur (en-tête AcceptLanguage) ou la géolocalisation IP. Nous avons testé l'effet de la variation de l'en-tête AcceptLanguage sur les pourcentages des indicateurs obtenus. Lors de la première exécution, cet en-tête n'était pas renseigné. Deux exécutions supplémentaires ont été réalisées sur le corpus Tranco : l'une avec l'en-tête défini comme anglais (en), et l'autre avec l'en-tête défini avec la séquence prioritaire chinois (zh), espagnol (es), français (fr) et portugais (pt). Les résultats de ces exécutions sont comparés dans la section 4.3.

Facteur d'extrapolation

Afin de quantifier la sensibilité des résultats au choix du facteur d'extrapolation, les calculs initiaux ont été répétés sur le corpus Tranco avec des valeurs de facteur d'extrapolation de 1 (aucune extrapolation), 1,5, 2, 2,5 et 3, correspondant à des taux d'adoption de *hreflang* supposés de 100 %, 67 %, 50 %, 40 % et 33 % respectivement. Les résultats sont présentés dans la section 4.3.

Randomisation

Afin de vérifier la validité statistique du plan d'échantillonnage 100 × 1 000, l'analyse complète a été réalisée sur deux échantillons indépendants sélectionnés aléatoirement dans le même corpus Tranco. Pour la deuxième exécution, nous avons vérifié si les valeurs produites se situaient dans l'intervalle de confiance à 99 % de la première exécution, afin de confirmer que les intervalles de confiance rapportés caractérisent correctement la variabilité des estimations. Les résultats sont présentés dans la section 4.3.

4) Résultats et discussion

Tous les résultats sont accessibles au public sous forme de fichiers Excel téléchargeables sous licence CC-BY-SA 4.0 :

- Pour TRANCO : https://www.obdilci.org/wp-content/uploads/2026/04/MECILDI-Results1.xlsx
- Pour les gTLD liés à la France : https://www.obdilci.org/wp-content/uploads/2026/04/MECILDI-DGLFLFV1.8.xlsx

Remarque : le fichier précédent est une synthèse des résultats pour les 8 gTLD traités. Pour des résultats détaillés pour chaque gTLD, les fichiers spécifiques sont accessibles sur la page suivante : https://www.obdilci.org/projets/mecildi/france-gtlds-2/

Pour tenter d'obtenir des indicateurs approximatifs pour l'ensemble du web, nous avons fusionné les données issues des 3 principales sources (modèle OBDILCI, études OBDILCI avec DataProvider.com et MECILDI V1) : https://www.obdilci.org/projets/mecildi/www-state-of-multilingualism-2/

Remarque : Certains tableaux ci-après sont construits à partir de résultats intermédiaires et présentent donc des chiffres légèrement différents de ceux de la version finale 1.8. Une fois le point établi, il n'a pas été nécessaire de le reproduire pour chaque version successive jusqu'à la version finale.

4.1 Résultats principaux : Tranco (novembre 2025)

Le tableau suivant présente les résultats de la distribution des langues primaires obtenus à partir de 100 000 domaines sélectionnés au hasard dans le corpus Tranco en novembre 2025, exprimés en %PAGES ainsi que les intervalles de confiance correspondants à chaque indicateur.

Table 4 : Résultats du corpus Tranco (novembre 2025)

Langue	%PAGES	IC à 99 %
Anglais	22,07%	0,80%
Allemand	6,59%	0,23%
Français	6,35%	0,22%
Espagnol	6,34%	0,23%
Italien	4,08%	0,16%
Russe	3,96%	0,14%
Portugais	3,90%	0,14%
Néerlandais	3,15%	0,15%
Japonais	3,05%	0,14%
Chinois	2,65%	0,13%

Polonais	2,57%	0,12%
Turc	1,78%	0,11%
Suédois	1,77%	0,11%
Coréen	1,65%	0,11%
Arabe	1,59%	0,10%
Tchèque	1,52%	0,09%
Indonésien	1,49%	0,09%
Danois	1,39%	0,10%
Ukrainien	1,30%	0,10%
Finlandais	1,30%	0,09%
Hongrois	1,23%	0,08%
Roumain	1,22%	0,08%
Grec moderne	1,12%	0,09%
Thaïlandais	1,08%	0,09%
Vietnamien	1,06%	0,09%
OTHERL	1,00%	0,15%

Notes :

- Les valeurs de la colonne *IC à 99 %* représentent la marge d'erreur à un niveau de confiance de 99 %.
- *OTHERL* désigne un code de langue trouvé dans *lang* ou *hreflang* et non inclus dans la liste des langues gérées.

Sur 100 000 domaines traités, 49,30 % étaient inaccessibles ou ont renvoyé des erreurs et ont été exclus de l'analyse, ce qui a permis de constituer un corpus de travail de 50 700 sites web accessibles. Parmi ceux-ci, 33,78 % étaient classés comme multilingues, avec en moyenne 7,00 versions linguistiques par site multilingue et un MultiR global de 3,03. Dans ce corpus de travail, l'anglais était la langue la plus représentée (22,07 % des versions linguistiques), suivi de l'allemand (6,59 %), du français (6,35 %) et de l'espagnol (6,34 %). Le regroupement relativement étroit de ces trois langues européennes est remarquable et sera analysé plus en détail dans la section 4.5, en tenant compte des biais géographiques connus du corpus Tranco. La méthode s'est assurée que les résultats obtenus sur ce corpus de travail peuvent légitimement être représentatif du corpus initial complet d'un million de sites, à l'intérieur d'un intervalle de confiance qui a été calculé pour chaque indicateur et qui est toujours resté inférieur à 1%.

Au total, 148 langues ont été détectées avec au moins une occurrence dans l'échantillon ; 66 des 216 langues détectables par Tomedes n'ont jamais été relevées.

4.2 Validation : Facteur d'extrapolation dérivé de l'heuristique

Pour valider le facteur d'extrapolation supposé de 2,5, le facteur d'extrapolation heuristique (FE) a été calculé indépendamment pour chaque jeu de données. Comme décrit dans la section 3.11, le FE fournit une estimation ascendante du facteur d'extrapolation basée sur les scores heuristiques de multilinguisme observés et peut être comparé à la valeur supposée descendante de 2,5 afin d'évaluer si le facteur d'extrapolation est approprié pour un corpus donné.

Remarque : Ce mécanisme a été conçu essentiellement en tenant compte de la liste Tranco et a permis la confirmation dans ce cas précis. Il n'est pas surprenant qu'appliqués à des séries très différentes de Tranco en termes de multilinguisme, les chiffres divergent et qu'il soit difficile d'en tirer des conclusions, si ce n'est que ce chiffre indique une proximité du multilinguisme avec Tranco. En réalité, le multilinguisme peut s'exprimer soit par un pourcentage élevé de sites multilingues, soit par un taux élevé de multilinguisme. Le nombre moyen de langues par site multilingue joue un rôle clé dans l'influence de ce taux (par exemple, .eus est extrêmement multilingue en termes de pourcentage de sites, mais pas en termes de taux, car il s'agit en fait d'un bilinguisme presque parfait, proche de 100 %). Dans ce contexte, en dehors du fait qu'il s'agit d'une mise en garde, il est difficile de tirer des conclusions de cette divergence.

Pour le corpus Tranco, le HDEF était de 2,32 ± 0,08, validant ainsi le facteur d'extrapolation supposé de 2,5. La forte concordance entre les deux estimations – l'une issue d'une annotation manuelle et l'autre d'une évaluation heuristique automatisée – confirme la fiabilité de l'approche d'extrapolation pour les corpus web généralistes. Par ailleurs, la base de référence ayant été établie en septembre 2024, de légères variations des indicateurs de multilinguisme ont pu survenir au cours des deux années suivantes, telles qu'une augmentation générale de l'adoption du web multilingue accompagnée d'une diminution du pourcentage de pages en anglais (%PAGES).

4.3 Analyse de sensibilité

Trois analyses de sensibilité ont été réalisées pour évaluer la robustesse des résultats principaux.

En-tête Accept-Language

Étant donné que certains sites web affichent un contenu différent selon la langue préférée du navigateur, nous avons testé l'effet de la variation de l'en-tête de requête Accept-Language. Notre première exécution a laissé cet en-tête vide. Deux exécutions supplémentaires ont été réalisées : l'une avec l'en-tête défini sur l'anglais, et l'autre avec les langues suivantes, par ordre de priorité : chinois (zh), espagnol (es), français (fr) et portugais (pt).

Table 5 : Résultats de la modification de l'en-tête Accept-Language

En-tête Accept-Language	%PAGES (anglais)	%PAGES (chinois)	MultiR	%Multi	AvgL	ERR-TOT	ERR-BLK
zh , es, fr , pt	22,07%	2,71%	3.03	33,78%	7.00	49,30%	12,38%
en	22,28%	2,65%	3.02	33,79%	7.00	48,81%	11,60%

% de variation	0,22%	2,1%	-0,36%	0,01%	-0,55%	-0,49%	-0,79%

L'impact a été marginal. Le fait de définir l'en-tête en anglais n'a augmenté que de 0,22 % le pourcentage de pages web en anglais, tandis que le fait de le définir en chinois l'a augmenté de 2,1 %. Tous les autres indicateurs ont varié de moins de 1 %. Ces résultats confirment que l'en-tête Accept-Language a un effet négligeable sur les mesures de répartition des langues au niveau du corpus et que nos principaux résultats ne sont pas significativement affectés par ce paramètre. La plateforme n'a donc plus utilisé ce paramètre.

Facteur d'extrapolation

Pour évaluer la sensibilité des résultats au choix du facteur d'extrapolation, les calculs principaux ont été répétés en utilisant des valeurs de facteur d'extrapolation de 1 (pas d'extrapolation), 1,5, 2, 2,5 et 3.

Table 6 : Résultats des variations du facteur d'extrapolation (FE) sur le pourcentage de pages Web en langue *L* (%PAGES)

FE	% Extrapolation	Anglais	Allemand	Français	Espagnol	Italien	Russe	Portugais	Néerlandais
1	**Aucune extrapolation**	32,3%	5,6%	5,1%	5,7%	3,2%	4,9%	3,7%	2,6%
1.5	**66,7%**	27,6%	6,0%	5,7%	6,0%	3,6%	4,4%	3,8%	2,8%
2	**50,0%**	24,4%	6,4%	6,1%	6,2%	3,9%	4,2%	3,8%	3,0%
2.5	**40,0%**	22,1%	6,6%	6,4%	6,3%	4,1%	4,0%	3,9%	3,1%
3	**33,3%**	20,3%	6,8%	6,6%	6,4%	4,2%	3,8%	3,9%	3,3%

Table 7 : Résultats des variations du facteur d'extrapolation (FE) sur le pourcentage de pages Web en langue *L* (%PAGES) (suite)

FE	Japonais	Chinois	Polonais	MultiR	%Multi
1	3,3%	2,6%	2,2%	1,81	13,5%
1.5	3,2%	2,6%	2,3%	2.22	20,3%
2	3,1%	2,7%	2,5%	2,62	27,0%
2.5	3,1%	2,7%	2,6%	3.03	33,8%
3	3,0%	2,7%	2,6%	3,43	40,5%

En fonction de la valeur de FE entre 33 % et 100 %, le pourcentage de pages web en anglais varie de 20,3 % à 32,3 %. Il est important de noter que pour toutes les valeurs du facteur d'extrapolation supérieures ou égales à 1,5 (correspondant à un taux d'adoption de l'attribut *hreflang* supposé inférieur ou égal à 67 %, ce qui constitue une limite supérieure prudente), la part de l'anglais reste

inférieure à 28 %. Cette analyse de sensibilité confirme la robustesse de la conclusion selon laquelle l'anglais représente nettement moins de 50 % du contenu web, et ce, pour toutes les valeurs plausibles du facteur d'extrapolation. Le choix du facteur d'extrapolation a un impact tout aussi modeste sur les autres langues européennes, et quasiment aucun sur le chinois, langue rarement présente sur les sites multilingues et dont la présence varie donc peu, quelle que soit l'extrapolation appliquée.

Randomisation

Afin de vérifier la validité statistique de la méthode d'échantillonnage, l'analyse a été menée sur deux échantillons indépendants sélectionnés aléatoirement dans le même corpus Tranco. Sur l'ensemble de ces analyses, 97,8 % des valeurs mesurées lors de la seconde analyse se situaient dans l'intervalle de confiance à 99 % de la première analyse. Parmi les quatre valeurs hors intervalle de confiance, l'écart maximal était de 0,07 %. Ce résultat confirme la validité du plan d'échantillonnage 100 × 1 000 et démontre que les intervalles de confiance rapportés caractérisent fidèlement la variabilité des estimations.

Analyse des tendances

Tranco offre la facilité de télécharger des listes établies dans des dates antérieures. Cette fonctionnalité a été utilisée pour réaliser une analyse des tendances. Il est important de comprendre que, lors du traitement d'une série de sites supposés être les plus visités il y a quelques années, le programme MECILDI extrait les données telles qu'elles sont aujourd'hui, et non telles qu'elles étaient à la date de création de la liste. Cette réalité empêche d'établir une analyse objective, car le contenu des sites web a pu évoluer depuis. On peut toutefois considérer ces données comme un indicateur secondaire des tendances, une sorte de tendance minimale, car il est logique de supposer que les sites web tendent à évoluer vers un multilinguisme accru, et non l'inverse. Par conséquent, la tendance réelle est probablement plus marquée que les chiffres obtenus, et ces derniers pourraient être considérés comme une estimation basse.

Nous avons mesuré les résultats pour janvier 2020, mars 2023 et juin 2026, et ils montrent une nette tendance à un multilinguisme accru et à une présence moindre de l'anglais.

Table 8 : Analyse des tendances

	Janvier 2020	Mars 2023	Juin 2026
%PAGES(anglais)	32.38%	26.80%	20.89%
MultiR	2.30	2.65	3.28
%Multi	24.70%	29.20%	36.42%
AvgL	6.28	6.62	7.21
HDEF	1.69	2.06	2.63

Grande variabilité des indicateurs

Les résultats concernant les gTLD confirment les conclusions d'études antérieures utilisant la base de données DataProvider.com sur les gTLD linguistiques européens et le multilinguisme web en général (OBDILCI, 2025c), et démontrent que les valeurs des indicateurs ont une variance très haute d'un corpus web à l'autre.

Table 9 : Variation des résultats

Indicateur	Le plus haut	Série	Le plus bas	Série
MultiR	6.05	.gp	1.42	.nc
%Multi	96.54%	.eus	16.42%	.bzh
AvgL	9.28	.gp	2.45	.quebec
Google Translate	1.47%	.corsica	0.00%	.gp
HDEF	11.301	.quebec	1.22	.bzh
Présence de *Lang*=	86.87%	.bzh	58.23%	.quebec
Erreurs avec *Lang*=	11.33%	.eus	0.00%	.corsica
IDN	7.88%	.quebec	0.00%	plusieurs
ERR-DNS	67.62%	.quebec	12.78%	.nc
ERR-BLK	12.38%	Tranco	1.70%	.corsica
ERR-CFG	22.18%	.paris	1.18%	.nc
ERR-http	10.16%	Tranco	2.50%	.quebec
ERR-LANG	14.46%	.gp	1.22%	Tranco
ERR-TOT	82.34%	.quebec	42.16%	.nc

Note : Ce tableau a été établi à partir de résultats intermédiaires ; les différences avec les résultats finaux sont minimes. Les taux d'erreur sont plus élevés dans les corpus provenant du service de données de zone centralisé de l'ICANN (CZDS), qui inclut tous les domaines enregistrés (y compris les domaines parqués), alors que ZoneStats.io les exclut probablement.

Analyse des sites multilingues

Il est intéressant d'analyser ces mêmes indicateurs sur les sites multilingues. L'analyse qui suit porte seulement sur les sites dont l'attribut *hreflang* est présent dans plusieurs langues et ne tient donc pas compte de l'extrapolation.

Table 10 : Répartition des langues dans les sites multilingues (Tranco)

	%PAGES	%SITES
TOTAL	**100.00%**	**715.60%**
Anglais	12.66%	90.58%
Allemand	7.26%	51.94%
Français	7.24%	51.81%

Espagnol	6.63%	47.48%
Italien	4.87%	34.82%
Portugais	4.13%	29.55%
Néerlandais	3.75%	26.81%
Russe	3.05%	21.83%
Polonais	2.96%	21.18%
Japonais	2.86%	20.50%
Chinois	2.71%	19.38%
Suédois	2.08%	14.88%
Coréen	2.03%	14.53%

Le nombre moyen de langues par site multilingues de la série Tranco analysée est de 7,156. Il convient de souligner l'écart important entre les pourcentages de pages et de sites, logiquement liés ici par la formule %Sites = AvgL x %Pages. Plus de 90 % des sites multilingues de la série Tranco ont une version anglaise, et environ la moitié d'entre eux proposent des versions allemande, française et espagnole. Pourtant, le pourcentage de pages web en anglais est inférieur à 13 %. Fait intéressant, ce chiffre est proche de la proportion mondiale de locuteurs de l'anglais (L1 + L2), soit 14,1 % de la population mondiale de locuteurs, selon les données d'Ethnologue de 2024. Ces chiffres, en apparence contradictoires, ne le sont pas ; leur divergence importante est plutôt une conséquence mathématique directe du multilinguisme dense, où les sites web de ce sous-ensemble multilingue prennent en charge en moyenne plus de sept langues par site.

Même si l'anglais était universellement présent sur chaque site web multilingue, une moyenne de sept langues par site limiterait mathématiquement la part maximale possible de pages en anglais à 14,3 % (100 divisé par 7).

La quête infructueuse de contenu chinois dans Tranco

Sachant que Chrome était largement utilisé en Chine, avec une prévalence d'environ 50 % des utilisateurs, nous avons sélectionné Crux (Chrome User Experience Report), une des sources de Tranco, comme source unique pour un échantillonnage personnalisé. Cette fonctionnalité de Tranco nous permettait d'espérer observer une présence plus importante du chinois dans les résultats. Nous avons également profité de l'occasion pour définir un paramètre spécifique, différent de la liste standard de Tranco : « n'inclure qu'un seul domaine (le plus populaire) pour une organisation donnée ». Ce paramètre empêche, par exemple, que google.com et google.fr apparaissent simultanément dans la série.

Paradoxalement, les résultats ont montré une prévalence du chinois encore plus faible et ont souligné l'importance des sites web internationaux à succès pour faire grimper l'AvgL, un phénomène que nous avons observé lors de l'étude des gTLD, notamment pour le .gp où le multilinguisme élevé est en partie dû aux nombreuses redirections vers d'importantes entreprises proposant plusieurs versions linguistiques de leur site web.

Table 11 : Résultats avec CRUX

INDICATEUR	CRUX	STANDARD	DIFFÉRENCE
MultiR	2.45	3.28	**-33.79%**
%Multi	35.29%	36.42%	-3.20%
AvgL	5.09	7.21	**-41.64%**
HDEF	2.34	2.63	-12.33%
Lang=	71.96%	69.14%	3.92%
%PAGES			
Anglais	24.18%	20.89%	**13.63%**
Espagnol	6.67%	6.25%	6.39%
Français	6.49%	6.29%	3.09%
Allemand	6.26%	6.58%	-5.18%
Indonésien	4.67%	1.84%	**60.52%**
Italien	4.09%	4.05%	0.76%
Portugais	3.91%	3.90%	0.18%
Japonais	3.43%	2.97%	13.23%
Russe	3.31%	3.54%	-6.79%
Néerlandais	2.95%	3.23%	-9.43%
Polonais	2.41%	2.70%	-11.94%
Chinois	1.98%	2.86%	**-43.98%**

Cette expérience permet en outre de comprendre la fragilité des projections linguistiques réalisées pour le www entier à partir des chiffres concernant le million de sites les plus visités, puisque les résultats dépendent fortement des sources utilisées.

4.4 Résultats des gTLD francophones

MECILDI a été appliqué à huit domaines de premier niveau francophones et régionaux associés directement ou indirectement à la France : .alsace , .bzh , .corsica , .eus , .gp , .paris et .quebec , ainsi qu'à .nc (Nouvelle-Calédonie). Vous trouverez ci-dessous un résumé des résultats.

Table 12 : Résumé de l'étude gTLD

gTLD	MultiR	HDEF	%PAGES(premières langues)	Présence des langues régionales
.alsace	1,73	2,34	Français (51%), anglais (21%)	Aucune présence significative
.bzh	1,44	1.22	Français (58%), anglais (18%)	Breton : mesurable (2,91 %)
.corsica	1,85	2.21	Français (43%), anglais (27%)	Corse : mesurable (3,55 %)

.eus	2,57	11.3	Espagnol (36%), basque (36%)	Basque : fort (36,38 %)
.gp	6.05	5,45	Anglais (13%), français (13%)	Créole français : Présence non significative
.nc	1.42	1,54	Français (66%), anglais (17%)	Langues kanak : aucune présence significative
.paris	2,48	3,91	Français (31%), anglais (27%)	Sans objet
.quebec	1,72	4,71	Français (46%), anglais (40%)	Sans objet

Plusieurs conclusions méritent d'être tirées.

Le français étant la langue dominante dans tous les domaines (à l'exception du .eus), il atteint son niveau le plus élevé en .nc (66%) et son niveau le plus bas en .gp (Guadeloupe), où l'anglais le devance de moins d'un point de pourcentage (12,81 % contre 12,57 %). La quasi-parité de l'anglais et du français en .gp reflète probablement la forte orientation de ce domaine vers le tourisme international.

Le domaine .eus (basque) a présenté le profil de multilinguisme le plus marqué de toutes les séries étudiées dans cette étude, ce qui concorde avec les résultats d'une étude antérieure (OBDILCI, 2025c). Plus précisément, 96,54 % de ces sites web du .eus ont été classés comme multilingues, ce qui correspond à un MultiR élevé de 2,57. L'espagnol (36,53 %) et le basque (36,38 %) étaient quasiment à égalité en tant que langues dominantes, reflétant le statut co-officiel du basque au Pays basque espagnol.

4.5 Comparaison avec W3Techs

Le tableau suivant compare les résultats de MECILDI sur le corpus Tranco avec les chiffres correspondants publiés par W3Techs (2026), qui utilise le même corpus source.

Table 13 : Comparaison des résultats avec W3Techs

MECILDI	MECILDI (%PAGES)	W3TECHS	W3TECHS	DIFF. ABSOLUE	DIFF. %
Anglais	22,07%	Anglais	49,50%	-27,43%	-55%
Allemand	6,59%	Espagnol	6,00%	0,34%	6%
Français	6,35%	Allemand	6,00%	0,59%	10%
Espagnol	6,34%	Japonais	5,00%	-1,95%	-39%
Italien	4,08%	Français	4,50%	1,85%	41%
Russe	3,96%	Portugais	4,10%	-0,20%	-5%
Portugais	3,90%	Russe	3,60%	0,36%	10%
Néerlandais	3,15%	Italien	2,80%	1,28%	46%
Japonais	3,05%	Néerlandais	2,20%	0,95%	43%
Chinois	2,65%	Polonais	1,80%	0,77%	43%

Polonais	2,57%	Turc	1,60%	0,18%	11%
Turc	1,78%	Chinois	1,20%	1,45%	121%
Suédois	1,77%	Indonésien	1,00%	0,49%	49%
Coréen	1,65%	Vietnamien	1,00%	0,06%	6%
Arabe	1,59%	Tchèque	0,90%	0,62%	69%
Tchèque	1,52%	Persan	0,90%	-0,60%	-67%
Indonésien	1,49%	Coréen	0,90%	0,75%	84%
Danois	1,39%	Ukrainien	0,70%	0,60%	86%
Ukrainien	1,30%	Hongrois	0,60%	0,63%	105%
Finlandais	1,30%	Arabe	0,60%	0,99%	164%
Hongrois	1,23%	Suédois	0,50%	1,27%	255%
Roumain	1,22%	Roumain	0,50%	0,72%	145%
Grec	1,12%	Grec	0,50%	0,62%	125%
Thaïlandais	1,08%	Danois	0,40%	0,99%	247%
Vietnamien	1,06%	Finlandais	0,40%	0,90%	225%
AUTRES	1,00%	Hébreu	0,40%	0,08%	20%

Le résultat le plus frappant, bien qu'attendu, est l'écart de 27,43 points de pourcentage entre le chiffre de W3Techs pour l'anglais (49,50 %) et celui de MECILDI (22,07 %). Comme indiqué dans la section 4.6, cet écart est principalement dû à la différence méthodologique entre la classification monolingue et la mesure prenant en compte le multilinguisme, plutôt qu'à des différences dans la sélection des corpus ou la détection des langues.

Pour la plupart des langues européennes autres que l'anglais, MECILDI affiche systématiquement des proportions supérieures à celles de W3Techs, et souvent de manière significative. Le suédois, par exemple, est mesuré à 1,77 % par MECILDI contre 0,50 % par W3Techs (soit une différence relative de 255 %), et le norvégien à 0,75 % contre 0,20 % (275 %). Ces écarts relatifs importants s'expliquent par le fait que ces langues apparaissent principalement sur des sites web multilingues contenant également de l'anglais ; selon l'approche monolingue de W3Techs, ces sites sont généralement classés comme étant en anglais, masquant ainsi la présence de l'autre langue.

Le chinois présente un profil différent. MECILDI mesure 2,65 % du chinois et W3Techs 1,20 %, soit une différence relative de 121 %. Ces deux chiffres sont toutefois bien inférieurs aux attentes démographiques. Comme nous l'expliquerons dans la section suivante, cela reflète une limitation structurelle commune au corpus Tranco plutôt qu'une différence méthodologique entre les deux systèmes. Néanmoins, en l'absence de documentation précise sur la méthodologie de W3Techs, il demeure difficile de comprendre pourquoi MECILDI a enregistré plus du double que W3Techs pour le chinois, notamment dans un contexte caractérisé par un très faible multilinguisme sur les sites web chinois.

Toute comparaison directe avec W3Techs est sujette aux informations méthodologiques limitées qu'ils rendent publiques. W3Techs ne fournissant que des descriptions générales de son processus de détection de la langue, il est impossible de déterminer précisément comment son système gère les erreurs, les redirections ou les domaines inactifs. Une partie des différences observées pourrait donc refléter des variations dans la gestion des erreurs ou la précision de la détection, plutôt que la seule prise en compte du multilinguisme. Néanmoins, l'ampleur de l'écart pour l'anglais est bien trop importante pour être expliquée par ces facteurs secondaires.

Cette comparaison ne constitue pas une critique de la méthodologie W3Techs, développée dans un but analytique différent. Elle illustre plutôt comment les choix méthodologiques, notamment le traitement des sites web multilingues, peuvent influencer considérablement les estimations de la répartition des langues. Ces résultats soulignent l'importance de prendre en compte le multilinguisme lorsque la répartition des langues est l'objet principal de l'étude.

4.6 Réévaluation de la domination de l'anglais sur le Web à travers %PAGES

La principale conclusion de cette étude est sans équivoque et confirme par la mesure la démonstration mathématique de (Pimienta, 2024) : lorsque le multilinguisme des sites web est correctement pris en compte, l’anglais représente environ 20 % des pages web de l’échantillon Tranco, soit bien moins que les 50 % souvent cités. Ce chiffre de 20% provient de la version finale 1.8 de MECILDI, qui constitue la version la plus aboutie et la mieux calibrée de la plateforme.

Toutefois, si l'on utilise le pourcentage de sites web proposant une version anglaise (%SITES) comme indicateur principal, ce chiffre demeure considérablement élevé (67,5 %) et nous avons constaté que plus de 90 % des sites multilingues possèdent une version anglaise, cependant le pourcentage de pages concernées n'est que de 13 %, ce qui témoigne de l'existence de nombreuses autres langues.

La question du pourquoi le chiffre de MECILDI, pour %SITES, de 67.5% est bien supérieur au chiffre de 49.50% de W3Techs, un résultat surprenant, est difficile à répondre. La réponse est liée à la définition de ce que les méthodes sans prise en compte du multilingue entendent comme « langue principale » d’un site, un concept exporté de la linguistique humaine, nous l’avons expliqué, qui en lui-même porte une incohérence avec la réalité de la linguistique des langues numérique, et en tous cas une grande ambiguïté, ce qui n’est pas le cas des concepts de %SITES et %PAGES,

En définitive, cela démontre que la conception méthodologique détermine fondamentalement le résultat ; malgré les hypothèses opérationnelles, la mesure de la distribution au niveau de la page fournit une mesure beaucoup plus précise et logiquement cohérente dont il faut tenir compte dans toute étude sérieuse du multilinguisme sur le Web.

4.7 Prévalence de l'anglais : vers une estimation à l'échelle de la Toile entière

Le corpus Tranco reflète le million de domaines les plus visités au monde, ce qui ne constitue pas un échantillon neutre linguistiquement de l'ensemble du web. Il est donc essentiel de se demander si la part de pages en anglais (%PAGES) de 20 % observée dans le million de sites web listé par Tranco est susceptible d'être supérieure ou inférieure à la proportion réelle d'anglais sur l'ensemble des quelques 200 millions de sites web actifs dans le monde.

Deux forces antagonistes influencent le pourcentage de pages web en anglais dans la Toile complète. Les sites web à fort trafic investissent généralement davantage dans une infrastructure multilingue que la moyenne ; par conséquent, le taux de multilinguisme observé dans l'échantillon Tranco (MultiR = 3,41) est très supérieur à celui attendu pour l'ensemble du web. Une étude parallèle menée par OBDILCI à partir de la base de données DataProvider.com (OBDILCI, 2025b) a établi une valeur de %Multi = 14,4 % pour l'ensemble du web, entre 2 et 3 fois moins que la mesure MECILDI de 37 % pour Tranco.

Bien que le nombre moyen de langues par site multilingue (AvgL) pour l'ensemble du web demeure inconnu, une estimation prudente de 4,5 – nettement inférieure à la moyenne de 7,5 observée dans le corpus Tranco (à haute densité) – donne un MultiR global plausible d'environ 1,5. Ce chiffre est proche du taux de multilinguisme de l'humanité, définit comme le ratio du nombre total de locuteurs L1+L2 par le nombre de locuteurs L1 (le nombre d'humains), qui a la valeur de 1,44 (selon les données de Ethnologue de 2024).

Ainsi, un MultiR plus faible, indiquant moins de variantes linguistiques non anglaises par site web, la part relative des pages en anglais augmente mécaniquement. Cet élément suggère que la part réelle de pages en anglais sur l'ensemble du web pourrait être supérieure aux 20 % observés dans le corpus Tranco.

D'un autre côté, il faut prendre en compte le fort biais de sélection de Tranco qui surreprésente, à cause d'un biais géographique de ses sources vers les pays occidentaux, les langues européennes, et sous-représente, en particulier, le chinois, qui est la langue comptant le plus grand nombre de locuteurs connectés au monde. Un corpus non biaisé comprendrait une plus grande proportion de contenu en chinois et dans d'autres langues non européennes, ce qui réduirait considérablement la part proportionnelle de l'anglais.

Lequel de ces deux facteurs contradictoires l'emporte ?

Les comparaisons externes apportent une corroboration utile. Netsweeper (2023) a mesuré la répartition des langues directement au niveau des pages web sur 12 milliards de pages – évitant ainsi tout biais lié au multilinguisme – et a constaté que l'anglais représentait 26,3 % des pages web. Ce chiffre est probablement lui-même surestimé en faveur de l'anglais, étant donné que l'infrastructure d'exploration de Netsweeper est principalement orientée vers des clients dans des pays industriels. DataProvider.com (2023), bien que ne tenant pas compte du multilinguisme, a trouvé que l'anglais représentait 51,3 %, contre 57,7 % pour W3Techs à la même date. En

supposant qu'ils partagent la même *définition* de « langue principale », l'écart de 6,4 points de pourcentage entre ces deux études monolingues serait attribuable aux différences de corpus (un million de sites versus l'ensemble du web). L'application d'une correction de corpus proportionnellement équivalente à la base de référence de 20 % de MECILDI donnerait une estimation pour l'ensemble du web d'environ 18,7 % de pages en anglais.

L'ensemble de ces comparaisons permet une estimation révisée de la prévalence des pages web en anglais sur l'ensemble du web, qui se situerait entre 19 % et 26 %, la valeur la plus probable étant plus proche de 20 % que de 26 %. Cette fourchette concorde à la fois avec le modèle d'OBDILCI, qui prévoit une prévalence de l'anglais de 20,1 % (juillet 2025), et avec les estimations empiriques les plus récentes établies par Pimienta (2025).

4.8 Limites du corpus Tranco

Bien que le corpus Tranco soit largement reconnu comme une référence précieuse et robuste pour la mesure du Web et la recherche en sécurité, ses sources introduisent des limites notables lorsqu'il est appliqué à la cartographie linguistique mondiale. Par exemple, le chinois n'occupe que la dixième place dans le corpus Tranco, représentant seulement 2,92 % des pages Web analysées. Ce chiffre contraste fortement avec les 19 % projetés par le modèle d'OBDILCI et les 19,8 % mesurés par Netsweeper sur un corpus Web plus vaste, non soumis à des restrictions de popularité. Cet écart n'est pas dû aux méthodologies de détection ; il reflète plutôt une limitation involontaire des sources à partir desquelles Tranco est construit.

Ce biais géographique et linguistique découle directement de la méthodologie utilisée pour établir les classements de popularité. Tranco agrège les données de trafic provenant de sources commerciales et d'infrastructures, notamment CrUX, Farsight, Majestic, Radar et Umbrella. Ces sources établissent leurs listes de sites web « les plus visités » à partir de métriques de navigateur et de journaux de requêtes DNS inégalement répartis à l'échelle mondiale (Ruth et al., 2022). Ces outils de suivi, fortement dépendants d'infrastructures et de bases d'utilisateurs concentrées dans les pays occidentaux, privilégient implicitement les langues européennes dominantes et sous-représentent les écosystèmes de réseaux localisés, tels que les plateformes web chinoises à routage interne (par exemple, Baidu, Weibo, Taobao), qui fonctionnent en grande partie en dehors des panels de suivi occidentaux.

En définitive, la conclusion méthodologique est simple : il est statistiquement incorrect d'extrapoler les résultats obtenus à partir de la liste Tranco à l'ensemble du Web. Le concept de liste des « sites les plus visités », combiné à l'infrastructure géographiquement biaisée utilisée pour l'établir, surreprésente les langues européennes dominantes. Si Tranco demeure un outil précieux pour analyser les domaines à forte visibilité et accessibles dans le monde entier, il n'est pas conçu pour représenter la véritable diversité de la Toile. Si l'objectif de la recherche est de mesurer la répartition réelle des langues sur l'ensemble du Web, un corpus basé sur le classement par popularité comme Tranco ne convient pas pour représenter l'écosystème Web dans son ensemble.

4.9 Implications pour les politiques et la recherche

La persistance du chiffre tant médiatisé de 50 % d'anglais sur la Toile, et son corollaire : *« l'anglais est la lingua franca de l'Internet* » a des conséquences bien réelles. Elle influence les conceptions de l'inclusion numérique, détermine les langues à privilégier dans les données d'entraînement des IA et alimente les débats politiques sur la diversité linguistique en ligne. Un chiffre autour de 20 % révèle une tout autre réalité. Si l'anglais demeure de loin la langue la plus représentée sur le web, en particulier, si l'on considère le pourcentage de sites web proposant une version anglaise, le paysage numérique est véritablement et largement multilingue. Des dizaines de langues représentent désormais entre 1 % et 7 % du contenu total des pages web, tandis que des centaines d'autres se situent entre 1 % et 0.001%. Fait significatif, des analyses de tendances récentes suggèrent que cette diversification s'accélère.

La répartition en termes de pages web reflète plus fidèlement la diversité linguistique des populations connectées que ne le suggère le discours dominant. Elle implique également que les utilisateurs recherchant du contenu dans des langues autres que l'anglais bénéficient d'un accès au web bien plus performant que ne l'indiquent les indicateurs monolingues ; et inversement, que les défis de l'inclusion numérique auxquels sont confrontés les locuteurs de langues aux ressources limitées sont plus complexes qu'une simple vision de la domination de l'anglais ne le laisse entendre.

Pour les chercheurs, l'implication méthodologique est claire. Les études qui attribuent une seule langue par domaine ne mesurent pas la distribution linguistique sur le web, mais le profil linguistique des pages d'accueil des domaines. Ce sont deux grandeurs liées, mais distinctes. Les confondre produit des résultats trompeurs dès lors que la question de recherche porte sur la disponibilité réelle du contenu plutôt que sur la classification des domaines. Les travaux futurs dans ce domaine devraient adopter une mesure prenant en compte le multilinguisme comme norme de base et tenir explicitement compte des biais liés aux corpus, notamment la sous-représentation du chinois et des autres langues non occidentales dans les classements de domaines par popularité.

Enfin, les résultats concernant les gTLD démontrent que la distribution des langues sur le Web n'est pas un phénomène global unique, mais un phénomène à forte variance. Cette variabilité souligne l'importance du choix du corpus dans toute étude de la distribution des langues sur le Web et suggère que les données globales agrégées doivent toujours être accompagnées d'une analyse désagrégée au niveau de domaines, de régions ou de communautés linguistiques spécifiques.

5) Limites et travaux futurs

La version 1 de MECILDI a été conçue dans un souci de simplicité, en raison de contraintes budgétaires. Ce choix de conception introduit plusieurs biais connus, dont chacun est analysé ci-dessous en fonction de son impact probable sur les résultats. Le développement de la version 2 de MECILDI vise principalement à corriger ces limitations pour offrir des données parfaitement

fiables et mises à jour avec la fréquence nécessaire pour que la construction de politiques publiques puisse s'appuyer, tant en termes de diagnostic que de mesure de l'impact des politiques.

5.1 Biais statistique

L'analyse d'un échantillon de 100 000 domaines représente environ 10 % du corpus Tranco, qui compte un million de domaines. En principe, ce taux d'échantillonnage pourrait introduire un biais ; en pratique, le cadre statistique employé (100 sous-ensembles indépendants de 1 000 domaines chacun) garantit que ce n'est pas le cas, notamment pour les langues dont le pourcentage de pages dépasse 0,01 %. L'analyse de sensibilité de la randomisation a confirmé que des échantillons tirés indépendamment produisent des résultats se situant dans le même intervalle de confiance à 99 % dans 97,8 % des cas, démontrant ainsi la robustesse du plan d'échantillonnage, à condition que le corpus de domaines cible ne dépasse pas un seuil de plusieurs millions d'unités.

La principale limitation pratique réside dans le bas du classement linguistique. Les langues dont la part mesurée est inférieure à environ 0,01 % ne sont représentées que par quelques occurrences dans un échantillon donné, et les langues spécifiques apparaissant à ce niveau peuvent varier d'une analyse à l'autre en fonction du tirage aléatoire. Les résultats concernant les langues de cette catégorie doivent donc être considérés comme indicatifs plutôt que précis. Ce seuil pourrait être amélioré dans de futurs travaux en augmentant la taille du sous-ensemble pour les corpus où la détection des langues de faible fréquence est prioritaire.

Concernant la limite supérieure de cette architecture d'échantillonnage, l'application du cadre de 100 x 1000 domaines à des corpus dépassant le million de domaines risque de faire sortir les estimations des langues minoritaires des intervalles de confiance acceptables. Pour des populations de taille moyenne (de 2 à 5 millions de domaines), cette conception reste potentiellement valable pour mesurer les 10 à 20 langues dominantes les plus fréquentes ; cependant, pour les corpus dépassant 5 millions de domaines – ou lors de l'évaluation de langues à faible densité – l'erreur d'échantillonnage dépasse probablement les seuils de confiance standard. Définir et valider empiriquement cette limite supérieure sur des ensembles de données de ccTLD plus vastes constitue une question primordiale pour nos recherches futures.

5.2 Biais d'extrapolation

Le biais d'extrapolation est la limitation la plus importante de la version 1 et la justification la plus convaincante du développement de la version 2. Il comporte deux dimensions distinctes.

Le premier point concerne le taux d'adoption supposé *de hreflang* de 40 %. Ce chiffre a été obtenu empiriquement à partir d'une étude de référence portant sur 500 domaines, mais sa variation selon les corpus, les régions géographiques et les catégories de sites web est inconnue. L'analyse HDEF présentée dans la section 4.2 montre que l'hypothèse de 40 % est raisonnable pour les corpus généralistes tels que Tranco, mais pas pour les domaines fortement bilingues comme .eus et .quebec , ni pour les domaines à faible multilinguisme comme .bzh et .nc . L'analyse de sensibilité de la section 4.3 quantifie l'impact de cette incertitude : pour des valeurs allant de 1,5 à 3,

%PAGES(anglais) varie de 27,6 % à 20,3 %, confirmant ainsi la robustesse du principal résultat – l'anglais représente nettement moins de 50 % du contenu web – pour toutes les valeurs plausibles.

La seconde dimension concerne l'hypothèse selon laquelle la répartition linguistique des sites web multilingues détectés par *hreflang* est représentative de celle des sites multilingues utilisant d'autres mécanismes. La procédure d'extrapolation reproduit fidèlement le schéma linguistique observé *pour hreflang* sur les quelque 60 % de sites web multilingues qui n'utilisent pas *hreflang*. La validité pratique de cette hypothèse ne peut être vérifiée avec la seule version 1 ; elle ne pourra être évaluée qu'avec la version 2, intégrant une détection complète du multilinguisme. Toutefois, même si cette hypothèse introduit une certaine distorsion, le biais qui en résulte est nettement inférieur à celui induit par l'ignorance totale du multilinguisme.

5.3 Biais ISO 639-3

La version 1 traite uniquement les codes de langue à deux caractères conformes à la norme ISO 639-1. Bien que les langues les plus répandues dans l'Internet soient couvertes par cette norme, certaines ne sont identifiables que par des codes à trois caractères (ISO 639-3) et ne sont donc pas prises en compte par le système actuel. Une analyse de toutes les valeurs *`lang`* et *`hreflang`* extraites de l'échantillon Tranco le plus récent a révélé que les codes ISO 639-3 sont présents mais rares, apparaissant plus fréquemment dans les attributs ` *hreflang*` que dans les attributs *`lang`* . Les langues les plus susceptibles d'être affectées par cette limitation que nous avons identifiés sur l'échantillon de travail sont le philippin, le hmong, le cebuano et l'hawaïen, mais l'impact estimé dans chaque cas est inférieur à 0,2 %, une valeur qui reste dans la marge d'erreur due à la variabilité d'échantillonnage. Ce biais n'affecte que les langues les moins représentées et n'a aucun impact significatif sur les langues les plus représentées ni sur les principaux résultats de cette étude. La prise en charge complète de la norme ISO 639-3 est prévue pour la version 2, ainsi qu'un alignement rigoureux de l'analyse syntaxique avec les normes BCP 47 et RFC 5646 qui régissent les balises de langue internationales en HTML.

5.4 Biais de détection d'erreurs

Bien que de nombreuses erreurs soient directement signalées par le protocole HTTP, d'autres doivent être déduites par programmation. La plus importante d'entre elles est la détection des pages en construction et des pages en attente. Les pages en construction sont particulièrement problématiques car elles sont fréquentes (dans certaines zones gTLD, elles représentent plus de 10 % des domaines) et parce que leur texte d'espace réservé est souvent en anglais, quelle que soit la langue cible du site. Ne pas détecter et exclure ces pages aurait pour conséquence de gonfler artificiellement la part de l'anglais mesurée.

La simple correspondance par mots-clés sur des expressions telles que « en construction » ou « site en construction » s'est avérée insuffisante, générant trop de faux positifs. L'approche adoptée dans la version actuelle combine la présence d'expressions « en construction » dans plusieurs langues avec une liste d'hébergeurs connus, ne signalant une erreur que lorsque les deux signaux sont

présents. Ceci reflète le constat que les messages « en construction » sont fréquemment générés par les hébergeurs plutôt que par les propriétaires du site eux-mêmes. Les listes complètes des expressions et des hébergeurs utilisés sont disponibles dans les sections 3 et 4 des documents complémentaires.

Cette approche a permis de résoudre les cas les plus extrêmes rencontrés dans les ensembles de données des gTLD francophones et de réduire la part de l'anglais mesurée dans Tranco d'environ un point de pourcentage. Les listes de phrases relatives aux domaines en construction et aux fournisseurs d'hébergement restent toutefois incomplètes, notamment pour les environnements d'hébergement non européens, et seront complétées au fur et à mesure que de nouveaux cas seront identifiés.

5.5 Biais de l'algorithme de détection de la langue

Aucun algorithme de détection de langue n'est parfait. Des erreurs surviennent notamment entre les langues possédant des alphabets similaires et des vocabulaires proches, comme par exemple le portugais et le galicien, l'allemand et le suisse allemand, ou encore l'italien et le corse. Le détecteur de langue MachineTranslation.com (de Tomedes) a été sélectionné pour ses excellentes performances lors des tests préliminaires, et sa couverture de 216 langues est plus étendue que celle de la plupart des outils comparables. Néanmoins, certaines erreurs de classification sont inévitables, et leur impact est plus marqué pour les langues moins bien dotées en ressources et pour les textes courts, pour lesquels le seuil minimal de 250 caractères contribue à atténuer les erreurs de détection, sans toutefois les éliminer complètement. L'évaluation d'autres outils de détection de langue et d'approches d'ensemble est prévue pour la version 2.

6) Conclusions

Cette étude présente MECILDI, une plateforme méthodologique et informatique, avec la vocation de se transformer en observatoire pérenne, permettant de mesurer la distribution des langues et les indicateurs de multilinguisme, de n’importe quelle série de sites, en tenant explicitement compte du multilinguisme de chaque site. En détectant et en comptabilisant indépendamment chaque version linguistique d'un site web, au lieu d'attribuer une seule langue par site, MECILDI offre une représentation fondamentalement différente et plus précise du paysage linguistique du web que les approches de classification monolingues existantes.

Appliqué à 100 000 domaines sélectionnés aléatoirement dans le corpus Tranco, MECILDI constate que l'anglais représente environ 20 % du volume total de pages web, contre 50 % selon W3Techs pour le même corpus source. Cet écart de 30 points de pourcentage n'est pas imputable à des différences dans la sélection du corpus ou la détection de la langue ; il résulte directement de la prise en compte du multilinguisme, comme l'a prédit mathématiquement Pimienta (2024). Les méthodologies monolingues traditionnelles ont historiquement classé les sites web multilingues comme étant en anglais dès lors que l'anglais figurait parmi les langues prises en charge, ce qui a

gonflé la part mesurée de l'anglais et effacé la présence des langues co-occurrentes. MECILDI corrige ce biais en traitant chaque version linguistique comme une unité de mesure distincte.

Les principaux résultats concordent avec le modèle construit par OBDILCI, qui estime la part de l'anglais à environ 20 % du contenu web mondial. La corroboration par des études externes, notamment la mesure de Netsweeper (26,3 % au niveau des pages web) et une estimation corrigée issue de DataProvider.com, confirme une prévalence de l'anglais sur l'ensemble des pages web comprise entre 19 % et 26 %, la valeur la plus probable étant plus proche de 20 %. Le web n'est pas à moitié anglophone. Il est véritablement et largement multilingue, des dizaines de langues contribuant chacune de manière significative au contenu disponible.

Les résultats concernant les gTLD francophones confirment la forte variance de la répartition des langues sur le web selon les domaines et les communautés linguistiques, comme l'avaient déjà constaté des études antérieures utilisant DataProvider.com. Le taux de multilinguisme s'échelonne de 1,42 en .nc à 6,05 en .gp . Les langues régionales telles que le breton et le corse, bien que présentes sur seulement 2 à 4 % de leurs domaines respectifs, témoignent d'une vitalité numérique notable. Le domaine .eus affiche un bilinguisme quasi parfait, l'espagnol et le basque étant présents en proportions à peu près égales sur 96 % des sites web. Ce constat a des implications directes pour la politique linguistique et l'inclusion numérique au Pays basque.

La version 1 de MECILDI a été conçue dans un souci de simplicité, fonctionnant dans un environnement d'hébergement partagé et s'appuyant sur les annotations *hreflang* comme principal signal de multilinguisme, ajustées par un facteur de 2,5 déterminé empiriquement. Les analyses de sensibilité et la validation heuristique présentées dans cette étude confirment que cette approche produit des résultats robustes et reproductibles pour les corpus généralistes tels que Tranco. Cependant, le facteur d'extrapolation s'est avéré peu adapté aux corpus très spécialisés, notamment ceux présentant des profils de multilinguisme extrêmes comme .eus.

Il convient de noter que MECILDI offre une plateforme transparente et reproductible dont les hypothèses peuvent être testées, améliorées et recalibrées à mesure que de meilleurs jeux de données et méthodes de détection deviennent disponibles. Le facteur d'extrapolation, le taux d'adoption *de hreflang* et la sélection du corpus sont tous explicitement documentés et ouverts à la discussion, ce qui est précisément l'objectif. Les progrès dans ce domaine nécessitent non seulement de meilleures mesures, mais aussi de meilleures méthodes de mesure.

L'axe principal des travaux futurs est le développement de MECILDI version 2, qui visera à détecter l'ensemble des signaux de multilinguisme disponibles en HTML — notamment les structures d'URL spécifiques à chaque langue -, les structures de localisation JavaScript, les plugiciels linguistiques des CMS et les sélecteurs de langue personnalisés — réduisant ainsi la dépendance au facteur d'extrapolation et améliorant sensiblement la précision et fiabilité des données produites. Remédier à cette limitation de la détection est essentiel pour parvenir à une mesure de la distribution des langues sur le Web à la fois exhaustive et exempte des biais qui ont caractérisé le domaine jusqu'à présent.

En définitive, cette étude démontre que la généralisation des modèles linguistiques et multilingues des sites web à fort trafic à l'ensemble du web présente des limites intrinsèques et conduit à des conclusions erronées. Cartographier la réalité multilingue du web dans son intégralité demeure l'objectif ultime, mais le passage à l'échelle de centaines de millions de domaines actifs pose des défis opérationnels considérables. Une exploration à grande échelle doit gérer des exigences d'infrastructure massives, telles que la bande passante pour des flux parallèles et la puissance de calcul nécessaire au rendu JavaScript, ainsi que des barrières externes comme la limitation du débit, les pièges à robots d'exploration dynamiques, le filtrage des doublons et les protections anti-bots rigoureuses. Pour surmonter ces contraintes à grande échelle, les solutions futures s'appuieront probablement sur des partenariats stratégiques, que ce soit par le biais de collaborations directes avec des équipes de recherche reconnues, l'intégration de robots d'exploration existants ou des initiatives conjointes avec des communautés open source.

Les enjeux de la diversité linguistique, qui doivent prendre la place qu'ils méritent dans la gouvernance de l'Internet (Pimienta, 2026), ainsi que les défis concernant les langues, qui émergent de l'explosion de l'IA et des besoins de corpus de grande taille qui lui sont associés, plaident pour une professionnalisation de l'observation des langues dans le numérique pour pouvoir y guider lucidement les politiques linguistiques.

Références

Crystal, D. (1997). *English as a Global Language*. Cambridge: Cambridge University Press.

Dataprovider.com (2023). "What languages does the web speak?" *Dataprovider.com Blog*. Disponible à : https://www.dataprovider.com/blog/domains/what-languages-does-the-web-speak/ (accès le 26 juin 2026).

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (Eds.). (2025). *Ethnologue: Languages of the World*. 28th Edn. Dallas, TX: SIL International Disponible à : http://www.ethnologue.com (accès le 30 juin 2026).

Field, M., Pimienta, D. (2026). *MECILDI Language Detector (Version 1.8)*. Zenodo. https://doi.org/10.5281/zenodo.21212299

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," *Future Internet* 12(4), 76. doi: 10.3390/fi12040076

Google Inc. (2016). *Compact Language Detector v3 (CLD3)*. GitHub. Disponible à : https://github.com/google/cld3 (accès le 30 juin 2026).

Grefenstette, G., and Nioche, J. (2000). "Estimation of English and non-English language use on the WWW," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)* (Athens, Greece), 443–452. Disponible à : https://www.researchgate.net/publication/200111058_Estimation_of_English_and_non-English_language_use_on_the_WWW (accès le 26 juin 2026).

Hodges, J. (2025). *gocld3: Go implementation of Google's Compact Language Detector v3*. GitHub. Disponible à : https://github.com/jmhodges/gocld3 (accès le 30 juin 2026).

Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., and Joosen, W. (2019). "Tranco: A research-oriented top sites ranking hardened against manipulation," in *Proceedings of the 26th Annual Network and Distributed System Security (NDSS) Symposium* (San Diego, CA). doi: 10.14722/ndss.2019.23386

Netsweeper (2023). "Top languages commonly used on the Internet," *Netsweeper Blog*. Disponible à : https://www.netsweeper.com/government/top-languages-commonly-used-Internet (accès le 26 juin 2026).

Pimienta, D., Prado, D., and Blanco, A. (2009). "Douze années de mesure de la diversité linguistique sur l'Internet : bilan et perspectives." *UNESCO Publications for the World Summit on the Information Society*, CI.2009/WS/1. Disponible à : https://unesdoc.unesco.org/ark:/48223/pf0000187016_fre (accès le 26 juin 2026)

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 (SIGUL 2022)* (Marseille, France: European Language Resources Association), 83–91. Available at: https://aclanthology.org/2022.sigul-1.11
Version française : https://funredes.org/lc2022/Res.Ind.lang.Internet.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.* 8:1149347. doi: 10.3389/frma.2023.1149347
Version française : https://www.obdilci.org/wp-content/uploads/2024/01/METHOBS.fr_.pdf

Pimienta, D. (2024). “Is it True that More Than Half of Web Contents are in English? Not if Multilingualism is Paid Due Attention!” *Forum for Linguistic Studies* 6(5), 201–212. doi: 10.30564/fls.v6i5.7144 - Version française : https://obdilci.org/lc2022/English@www.en.fr.docx

Pimienta, D. (2025). “Inventory and comparisons of all methods for the measure of languages online: Implications for the Internet’s lingua franca,” in *Proceedings from 2nd International Conference on Language Technologies for All (LT4ALL 2025)* (Paris, France: UNESCO Headquarters). Disponible à :
https://obdilci.org/wp-content/uploads/2025/01/LangOnlineReport.docx
Version française actualisée :
https://obdilci.org/wp-content/uploads/2025/07/LangOnlineReport.en_.fr_.pdf

Pimienta, D. (2026). La diversité linguistique dans l’Internet « *De Babel à l’intelligence artificielle – Le plurilinguisme de Dante à nos jours* », Tremblay C., Herreras J.C. (coord.). Collection plurilinguisme, édité par l’Observatoire européen du plurilinguisme. ISBN : 9791042488130. Disponible à : https://obdilci.org/wp-content/uploads/2026/01/BabelIA-DLI.pdf

OBDILCI (2025a). *Rapports sur le multilinguismo de la Toile : Exploration de la présence des langues et du multilinguisme sur la Toile avec les données de Dataprovider.com* https://www.obdilci.org/projets/autres/mlreports-2/

OBDILCI (2025b) *Web Multilingualism Reports #1: Approximation of the Rate of multilingualism of the WWW*
https://www.obdilci.org/wp-content/uploads/2025/02/WebMultilingualismReport1.pdf

OBDILCI (2025c). *Web Multilingualism report #5 (Version 2): Web multilingualism analyzed by ccTLD, languages, gTLDs and much more* https://obdilci.org/wp-content/uploads/2025/08/TECHNICAL-REPORT-DATAP5.pdf

Ruth, K., Kumar, D., Wang, B., Valenta, L., & Durumeric, Z. (2022). Toppling top lists: Evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference* (pp. 374–387). https://doi.org/10.1145/3517745.3561444

Schur, P. (2025). *Language Detection: A language detection library for PHP*. GitHub. Disponible à : https://github.com/patrickschur/language-detection (accès le 30 juin 2026])

W3Techs (2026). *Usage statistics of content languages for websites*. Q-Success. Disponible à : https://w3techs.com/technologies/overview/content_language (accès le 26 juin 2026).