



Observatorio de la diversidad lingüística y cultural en la Internet
Observatoire de la diversité linguistique et culturelle dans l'Internet
Observatory of the Linguistic and Cultural Diversity in the Internet

<https://OBDILCI.ORG/>

SÉRIE : LEÇONS TIRÉES DE MÉCILDI

RAPPORT TECHNIQUE N° 2

Révision des chiffres concernant la présence des langues sur le Web.

OBDILCI, septembre 2026

Après avoir appliqué la plateforme multilingue MECILDI pour mesurer la distribution linguistique des contenus web dans une trentaine de séries web différentes (gTLD, ccTLD ou Tranco), il est temps de tirer les enseignements de cette expérience. Le présent rapport technique est le second d'une nouvelle série de rapports techniques élaborés par l'Observatoire de la diversité linguistique et culturelle sur Internet (OBDILCI). MECILDI a été financé par l'OIF et la DGLFLF.

La plateforme développée par OBDILCI, qui mesure la répartition des langues et le multilinguisme sur un ensemble de sites web, a été appliquée à la série Tranco (un million de sites web parmi les plus visités), à 8 gTLD associés aux langues de France, à 10 ccTLD de pays membres de l'Organisation de la Francophonie (OIF) et à une vingtaine d'autres TLD, sélectionnés pour la présence de langues locales avec un nombre de locuteurs importants. Ces mesures ont permis de confirmer des données antérieures sur la répartition des langues sur le web et de calculer de nouveaux indicateurs de multilinguisme.

La première étude, axée sur les langues **autochtones, vernaculaires, locales ou minoritaires**, dresse un tableau particulièrement inquiétant de leur faible présence sur le Web, avec des données bien inférieures aux prévisions les plus pessimistes. Voir : « [Le Web désertique des langues](#) »

autochtones, vernaculaires, locales et minoritaires », premier rapport de la série « Leçons apprises » du projet MECILDI, septembre 2026.

La deuxième étude se concentrera sur les chiffres réels de la présence des langues sur le web et sur la manière dont ils sont affectés par les résultats MECILDI et sur ce que cela signifie pour le modèle OBDILCI (Pimienta et al., 2024).

1 – ANGLAIS@WEB : LA CONFIRMATION PAR LA MESURE

L'OBDILCI a déclaré à maintes reprises, et ce depuis plusieurs années, que la principale source, W3Techs, citée par de nombreux rapports, y compris des sources réputées comme Wikipédia et Statista, pour affirmer une domination totale de l'anglais sur Internet, était fortement biaisée. Alors que W3Techs affirme, depuis 2011, que le pourcentage de contenu web en anglais se situe entre 50 % et 60 %, l'OBDILCI avance un chiffre compris entre 20 % et 26 %, soit une différence non négligeable.

L'OBDILCI a avancé que le principal biais résulte du fait de ne pas tenir compte de l'existence d'une proportion de sites web multilingues. Dans un article à comité de lecture, Pimienta a démontré ce biais et a même proposé une correction mathématique (Pimienta, 2024).

MECILDI a d'abord été chargé d'analyser TRANCO, la série des millions de sites Web les plus visités utilisée par W3Techs pour sa mesure, avec une méthode qui diffère de celle de W3Techs, en accordant une attention particulière aux sites Web multilingues (Pimienta D., Field M., 2026).

Les mesures indiquent un pourcentage de page web de 20 %, chiffre confirmé à la fois par le modèle OBDILCI et par la démonstration mathématique. Cette confirmation factuelle devrait lever tout doute quant à la fiabilité des données pour les personnes ayant une approche scientifique.

Deux leçons additionnelles ont été tirées de cette mesure ;

- 1) L'extrapolation des résultats de la répartition linguistique de Tranco comme approximation valable pour l'ensemble du web est erronée. L'échantillon Tranco est biaisé par définition (les sites web les plus visités ne reflètent pas la répartition linguistique de l'ensemble du web) mais aussi par construction (la sélection des sites web pour la série Tranco est fortement biaisée en faveur des langues européennes car les pays ne sont pas équitablement représentés dans les sources, les méthodes favorisant les pays occidentaux et, par conséquent, les principales langues européennes).
- 2) La mesure du second indicateur, au-delà du pourcentage de pages web dans une langue donnée, correspond au pourcentage de sites web ayant une configuration linguistique dans cette langue (il est important de noter que la somme pour toutes les langues n'est pas égale à 100 %, mais représente un chiffre que nous avons appelé le taux de multilinguisme de la série). On pourrait s'attendre à ce que le chiffre fourni par W3Techs soit égal à cet

indicateur. Or, en réalité, le chiffre calculé par MECILDI est sensiblement plus élevé (67,5 % contre 49,5 %). Il semble que W3Techs ait défini une notion de « langue principale d'un site web » et que ce soit cette notion qu'il mesure. Si les concepts de langue maternelle et de langues secondes s'appliquent aux êtres humains, il est difficilement transposable aux sites web. La notion de « langue principale d'un site web » est en fait définie par le visiteur, et non par le concepteur. Les sites web ont des configurations linguistiques et il n'existe pas de hiérarchie entre elles, sauf lorsqu'ils sont conçus pour sélectionner automatiquement la langue du navigateur du visiteur. Cette ambiguïté quant à ce que mesure réellement W3Techs s'ajoute au biais déjà identifié, rendant son indicateur difficile à appréhender.

Concernant la première leçon, il est difficile de répondre avec certitude à la question du pourcentage réel de pages web en anglais sur l'ensemble du web tant qu'un système de mesure n'est pas établi, une entreprise extrêmement complexe compte tenu de l'immensité du web. Je reproduis ci-dessous un extrait de (Pimienta, Field, 2026) :

4.7 Prévalence de l'anglais : vers une estimation à l'échelle du Web

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

Deux forces antagonistes sont à l'œuvre.. 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 bien 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 un seuil inférieur de 14,4 % pour l'ensemble du web, contre 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 s'approche du taux de multilinguisme de l'humanité, où le ratio du nombre total de locuteurs aux locuteurs natifs de langue maternelle s'élève à 1,44 (d'après les données de population de locuteurs d'Ethnologue de 2024). Un MultiR plus faible indiquant moins de variantes linguistiques non anglaises par site web, la part relative des pages en anglais augmente naturellement. Cela suggère que la part réelle des pages en anglais sur l'ensemble du web (%PAGES) est probablement légèrement supérieure aux 20 % observés dans le corpus Tranco.

En revanche, Tranco surreprésente les langues européennes et sous-représente le chinois, qui est la langue comptant le plus grand nombre de locuteurs connectés au monde. Un corpus moins

biaisé inclurait une plus grande proportion de contenu en chinois et dans d'autres langues non européennes, ce qui réduirait sensiblement la part proportionnelle de l'anglais.

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 existantes. Ce chiffre est probablement surestimé en faveur de l'anglais, étant donné que l'infrastructure d'exploration de Netsweeper est principalement orientée vers les pays industrialisés. DataProvider.com (2023), bien que ne tenant pas compte du multilinguisme, a trouvé que l'anglais représentait 51,3 %, contre 57,7 % pour W3Techs sur la même période. L'écart de 6,4 points de pourcentage entre ces deux études monolingues est attribuable aux différences de corpus (Tranco contre un échantillon plus large couvrant l'ensemble du web). L'application d'une correction de corpus proportionnellement équivalente au seuil de 20 % de MECILDI donne une estimation pour l'ensemble du web d'environ 18,7 %.

L'ensemble de ces comparaisons confirme une estimation révisée de la prévalence des pages web en anglais sur l'ensemble du web, qui se situe entre 19 % et 26 %, la valeur la plus probable étant plus proche de 20 % que de 26 %. Cette fourchette est cohérente à la fois avec le modèle OBDILCI, qui prévoit une prévalence de 20 % pour les pages web en anglais, et avec l'estimation empirique la plus récente publiée par Pimienta (2024).

2 – SURESTIMATION ET SOUS-ESTIMATION DANS LE MODÈLE OBDILCI

Les autres enseignements tirés de la première série de mesures MECILDI concernent la vision de la distribution des langues sur l'ensemble du web.

2.1 Langues locales

Le premier rapport a démontré, par des exemples convergents, que la présence des langues locales sur le web est extrêmement faible, souvent nulle, et en tout cas bien inférieure à l'estimation du modèle OBDILCI (voir <https://www.obdilci.org/projets/principal/>). Ce modèle doit être réajusté pour refléter cette réalité.

Le résultat final est obtenu à partir de la moyenne pondérée de 5 indicateurs, calculés par un processus complexe de pondération des indicateurs de base :

- Utilisation : recueil des indicateurs d'utilisation d'Internet par langue, en se concentrant notamment sur la présence des locuteurs des différentes langues sur les réseaux sociaux.
- Trafic : collecte, à partir d'un large éventail équilibré de sites web, la répartition du trafic par pays à l'aide d'applications de métriques telles que SimilarWeb.
- Interfaces : un indicateur de support technologique existant pour chaque langue, pondéré indirectement par la présence dans les services de traduction et les principales applications

et, depuis la V4.0, mai 2023, complété par un indicateur de degré de préparation des langues pour le numérique fourni par Ethnologue.

- Indices : collecte d'informations sur les indicateurs de la société de l'information par pays.
- Nombre de personnes connectées à l'Internet par pays : ce dernier indicateur joue un rôle particulier dans le modèle pour aider à pondérer les autres indicateurs en fonction de la présence de personnes connectées.

Avant la version 4.0, l'indicateur « Interfaces » était très sévère envers les langues locales, dont la grande majorité est totalement absente des sources analysées et se voyait donc attribuer la valeur 0. La version 4.0 a permis de rétablir un certain équilibre mais semble offrir un avantage à ces langues qui ne correspond pas à la réalité. Nous reviendrons probablement à la version précédente et, si cela ne suffit pas pour refléter la situation précaire de la plupart des langues locales, nous ajusterons la pondération des différents indicateurs afin de donner plus d'importance aux interfaces. Actuellement, la pondération est équilibrée entre les indicateurs, mais le modèle est conçu pour permettre des modifications de ces valeurs.

L'arabe peut être considéré comme une « langue locale » dans les pays arabophones ; sa présence sur les sites web n'est pas comparable à celle des autres langues locales analysées. Cependant, lors des premières mesures effectuées, le pourcentage de sites web en arabe s'est avéré inférieur aux prévisions. Par conséquent, le modèle OBDILCI présente une certaine surestimation, puisque l'anglais ou le français, selon les pays, occupent une place plus importante sur les sites web que l'arabe.

Vous trouverez ci-dessous les résultats pour le Koweït, le Liban, la Mauritanie et le Maroc, pour l'indicateur %SITES (pourcentage de sites web avec présence de la langue).

PAYS	langue maternelle	deuxième langue	3e langue
Koweït	Anglais (94%)	Arabe (77%)	Français (7%)
Liban	Anglais (81%)	Arabe (34%)	Français (12%)
Mauritanie	Français (75%)	Arabe (61%)	Anglais (36%)
Maroc	Français (81%)	Anglais (34%)	Arabe (23%)

2.2 Langues européennes dominantes

Les mesures TRANCO ont montré des résultats intéressants concernant les sites multilingues, en ce qui concerne les langues les plus privilégiées comme options linguistiques :

LANGUE	PRÉSENCE
Anglais	91,03%
Allemand	51,67%
Français	50,88%
Espagnol	47,95%
italien	34,88%

portugais	30,23%
Néerlandais	26,86%
polonais	21,93%
Russe	21,87%
japonais	20,98%
Chinois	19,84%

Cette hiérarchie de présence, avec l'anglais en première option dans 90 % des cas, puis l'allemand, le français et l'espagnol en deuxième option dans environ 50 % des cas, suivis de l'italien, du portugais, du néerlandais et du polonais, de 1/3 à 1/4 des cas, semble être un modèle reproductible à travers les différents TLD analysés jusqu'à présent.

Ce constat implique que le poids respectif de l'allemand, de l'italien, du néerlandais et du polonais en termes de locuteurs connectés ne reflète peut-être pas la présence réelle de ces langues dans le contenu web, et que le modèle nécessite des corrections. Il est possible de mieux refléter la prédominance de ces langues dans les résultats du modèle en ajoutant un nouvel indicateur : le nombre de domaines situés sous le ccTLD de chaque pays, car cette hiérarchie est plus ou moins liée à cet indicateur manquant.

Cet ajout engendrera toutefois certains défis, en raison de la combinaison de plusieurs facteurs :

- Il existe un nombre limité de sources concernant le nombre de domaines associés à chaque TLD : la plus fiable, CENTR, se limite à un sous-ensemble très restreint ; deux bases de données commerciales, [DomainMetaData](#) et [Zonestats.io](#), couvrent l'ensemble des TLD, mais leurs chiffres diffèrent, parfois considérablement. D'autres sources ([Statista](#)¹, [TLD](#)) ont une couverture très limitée. À ce stade, la meilleure estimation consiste à retenir les chiffres les plus élevés pour chaque domaine parmi toutes les propositions. Quoi qu'il en soit, il apparaît clairement que le .de, le ccTLD de l'Allemagne, est l'un des domaines les plus fournis et que les .nl et .it présentent des chiffres élevés par rapport à leur population respective (voir les données CENTR ci-dessous).

TLD	Domaines	Date de mise à jour
.de	18 150 432	Septembre 2026
.uk	10 657 724	Septembre 2026
.nl	6 135 231	Juillet 2026
.br	5 915 998	Septembre 2026
.fr	4 568 249	Septembre 2026
.au	4 416 817	Août 2026
.ue	3 670 879	Septembre 2026

¹Notez que STATISTA n'est pas la véritable source, mais elle conserve l'habitude contraire à l'éthique de cacher la source première à moins que vous ne payiez pour le service, une pratique que nous et d'autres avons dénoncée (<https://bernoff.com/blog/my-big-problem-with-statista>).

.il	3 624 600	Septembre 2026
.ca	3 589 744	Juillet 2026
.ch	2 654 560	Septembre 2026
.es	2 237 085	Septembre 2026
.pt	2 179 311	Septembre 2026
.be	1 708 285	Septembre 2026
.cz	1 593 675	Septembre 2026

- Il existe toujours un écart important entre le nombre de domaines enregistrés et le nombre de sites web actifs associés. Cet écart peut varier selon les TLD.
- Les États-Unis ont la fâcheuse habitude, pour la majorité des domaines, de ne pas utiliser leur extension nationale de premier niveau (ccTLD .us), préférant les extensions génériques (gTLD) telles que .com ou .edu. Cela pose problème quant au traitement de ce pays censé posséder le plus grand nombre de sites web actifs. Il convient de noter que .com est le domaine majoritaire en termes de nombre de sous-domaines, regroupant plus de la moitié du total des domaines existants.

Quoi qu'il en soit, la gestion de ce type de défis a fait partie intégrante de la construction du modèle et nous sommes convaincus que nous trouverons un moyen de transformer le résultat final du modèle :

	% INTERNET L1+L2	% POP. MOND. L1+L2	% LOC. connectés L1+L2	% USAGE L1+L2	% TRAFIC L1+L2	% INTERF. L1+L2	% INDEX L1+L2	% CONTENU L1+L2
Anglais	15,48%	14,13%	70,66%	25,96%	21,18%	20,17%	17,60%	20,08%
Chinois	17,57%	14,48%	78,23%	12,88%	22,42%	22,88%	19,48%	19,04%
Espagnol	6,63%	5,22%	81,90%	9,42%	6,82%	8,48%	7,13%	7,70%
Hindi	4,27%	5,68%	48,48%	2,52%	3,92%	4,44%	3,71%	3,77%
Russe	3,31%	2,38%	89,53%	4,37%	3,74%	3,75%	3,59%	3,75%
Arabe	4,36%	4,08%	68,79%	3,01%	2,39%	4,94%	3,55%	3,65%
Français	3,05%	2,91%	67,75%	4,11%	2,86%	3,98%	3,11%	3,42%
Portugais	2,97%	2,46%	77,76%	3,36%	3,04%	3,58%	2,83%	3,16%
Japonais	1,55%	1,15%	87,02%	3,30%	2,05%	1,97%	2,26%	2,23%
Allemand	1,80%	1,25%	93,00%	2,76%	1,39%	2,24%	2,58%	2,15%
Malais	2,55%	2,29%	71,73%	1,77%	1,77%	1,74%	2,30%	2,02%
Bengali	1,90%	2,59%	47,11%	0,97%	1,52%	1,57%	1,42%	1,48%
Turc	1,13%	0,84%	86,61%	1,26%	0,86%	1,30%	1,13%	1,14%
Italien	0,85%	0,62%	87,59%	1,28%	0,62%	1,07%	1,12%	0,99%
Vietnamien	0,98%	0,80%	78,47%	0,87%	1,11%	0,94%	0,86%	0,95%
Coréen	0,77%	0,76%	66,00%	1,18%	0,70%	0,91%	0,78%	0,87%
Persan	1,00%	1,03%	62,89%	0,62%	0,99%	0,90%	0,79%	0,86%

Ourdou	1,17%	2,22%	33,88%	0,59%	0,71%	0,90%	0,77%	0,83%
Tagalog	1,02%	0,78%	84,65%	0,78%	0,49%	0,56%	0,95%	0,76%
Thaïlandais	0,79%	0,57%	89,51%	0,64%	0,37%	0,79%	0,74%	0,67%
Marathi	0,69%	0,93%	48,19%	0,40%	0,64%	0,56%	0,60%	0,58%
Telugu	0,68%	0,89%	49,00%	0,42%	0,63%	0,53%	0,59%	0,57%
Polonais	0,53%	0,39%	87,01%	0,63%	0,39%	0,58%	0,67%	0,56%
Tamoul	0,63%	0,81%	50,59%	0,42%	0,56%	0,52%	0,56%	0,54%
Javanais	0,68%	0,64%	69,26%	0,45%	0,50%	0,35%	0,60%	0,52%
Ukrainien	0,47%	0,36%	83,93%	0,47%	0,32%	0,45%	0,52%	0,45%
Néerlandais	0,35%	0,24%	96,00%	0,46%	0,21%	0,40%	0,53%	0,39%
Gujarati	0,44%	0,58%	48,42%	0,28%	0,41%	0,35%	0,38%	0,37%
Pidgin nig.	0,68%	1,12%	39,21%	0,21%	0,27%	0,14%	0,50%	0,36%
Kannada	0,41%	0,55%	48,24%	0,24%	0,38%	0,32%	0,36%	0,34%

Les nouveaux résultats du modèle verraient probablement les langues marquées en rouge baisser dans le classement et celles en bleu monter, afin de mieux refléter la réalité mise en évidence par les mesures MECILDI.

RÉFÉRENCES

- Pages MECILDI donnant accès à tous les résultats publiés : <https://obdilci.org/projets/mecildi/>
- Pimienta D., Field M., 2026, MECILDI : un cadre informatique multilingue pour mesurer la distribution linguistique du contenu web., <https://www.obdilci.org/wp-content/uploads/2026/08/MECILDI-V1-FR.pdf>. Version anglaise en cours d'évaluation par les pairs dans Plos One.
- Études sur le multilinguisme en ligne. <https://obdilci.org/projets/autres/mlreports-2/>
- Pimienta D., 2024, « Est-il vrai que plus de la moitié du contenu web est en anglais ? Pas si le multilinguisme est pris en compte ! », <https://obdilci.org/lc2022/English@www.en.fr.docx>
Publié en anglais dans Forum for Linguistic Studies, 6(5), 201–212 .
<https://doi.org/10.30564/fls.v6i5.7144>
- Pimienta D., Blanco A., Muller de Oliveira G., 2023, La méthode à l'origine de la production sans précédent d'indicateurs de la présence des langues dans l'Internet . <https://www.obdilci.org/wp-content/uploads/2024/01/METHOBS.fr .pdf>
Publié en anglais dans Frontiers Research Metrics & Analytics, Volume 8 – 2023.
<https://doi.org/10.3389/frma.2023.1149347>