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/

**LEÇONS TIRÉES DE LA SÉRIE D'EXPÉRIENCES MÉCILDI**

**RAPPORT TECHNIQUE N° 4**

# Les langues privilégiées par les sites web multilingues

**OBDILCI, octobre 2026**

Après avoir appliqué la plateforme multilingue MECILDI pour mesurer la distribution linguistique du contenu web à près de 50 séries web différentes (gTLD, ccTLD ou Tranco), il est temps de tirer les enseignements de cette expérience. Le présent rapport technique est le quatrième d'une nouvelle série de rapports techniques élaborés par l'Observatoire de la diversité linguistique et culturelle dans l'Internet (OBDILCI).

La plateforme MECILDI développée par l'OBDILCI, avec le soutien du ministère français de la Culture (DGLFLF) et de l'Organisation de la Francophonie (OIF), mesure la répartition des langues et le multilinguisme sur un ensemble de sites web. Il a d'abord été appliqué à la série Tranco, qui recense le million de sites web les plus visités, à huit gTLD associés aux langues de France, à dix ccTLD de pays membres de l'OIF et à une trentaine d'autres ccTLD, sélectionnés pour la présence de communautés linguistiques importantes. 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 (OBDILCI, 1/2026) s'est concentrée sur les résultats relatifs aux langues **autochtones, vernaculaires, locales ou minoritaires** et dresse un tableau particulièrement préoccupant de leur faible présence sur le Web, avec des données bien inférieures aux prévisions les plus pessimistes.

La deuxième étude (OBDILCI, 2026-2) s'est concentrée sur **le tableau global** des indicateurs de répartition linguistique et des indicateurs de multilinguisme, ainsi que sur les ajustements nécessaires au modèle OBDILCI pour refléter les leçons tirées de l'utilisation de MECILDI dans 45 séries de sites web différentes (Tranco, ccTLD ou gTLD).

La troisième étude proposait une analyse du multilinguisme sur le web et, à partir de la caractérisation des différentes situations rencontrées, proposait **une typologie du web multilingue** basée sur certaines caractéristiques clés.

Cette quatrième étude porte sur les langues qui semblent être privilégiées comme choix pour des configurations linguistiques supplémentaires.

## 1-Le million de sites les plus visités

Le cadre MECILDI a permis d'établir des indicateurs de multilinguisme et d'analyser la répartition des langues parmi le million de sites web les plus visités, désignés sous le nom de TRANCO par le groupe de recherche qui gère cette liste (https://tranco-list.eu/). Au-delà de ces indicateurs, des études complémentaires ont été menées sur cet ensemble de sites, notamment l'analyse d'un sous-ensemble présentant les caractéristiques multilingues les plus fréquentes. Cette analyse repose sur l'utilisation de l'attribut HTML Hreflang =, qui spécifie la liste des langues prises en charge par le site (avec possibilité de précision par pays comme dans l'exemple hreflang = en, en-us, en-uk, fr-fr, fr-ca).

L'une de ces études a consisté à séparer, parmi les 100 000 sites web aléatoires utilisés pour calculer la distribution des langues, ceux utilisant cette instruction Hreflang et à analyser l' ensemble des langues.

Les résultats sont présentés dans le tableau suivant, où %pages(l) représente le pourcentage de pages en langue l parmi l'ensemble des pages correspondant à la série de sites web analysés, et %sites(l) le pourcentage de sites web configurés en langue l. La somme des %pages pour toutes les langues est égale à 100% par définition, et la somme des %sites est supérieure à 1 et représente le taux de multilinguisme de la série de sites web. Les langues les plus représentées et leurs pourcentages respectifs sont présentés ci-dessous, jusqu'à une valeur de %sites inférieure à 10 %.

Il se lira ainsi :

- Le taux de multilinguisme de cette série est de 7,47.
- L'anglais est présent sur 12 % du total des pages web.
- Plus de 90 % des sites web de la série incluent une version anglaise
- Plus de la moitié d'entre eux existent en versions allemande et française.

Ces résultats présentent la liste des langues les plus susceptibles d'être utilisées sur les sites web les plus visités, ainsi que les probabilités d'utilisation correspondantes, et définissent clairement 4 catégories principales :

1) L'anglais seul avec 0,9
2) Allemand, français et espagnol, environ 1/2
3) Italien, portugais, néerlandais entre 1/3 et 1/4
4) Polonais, Russes, Japonais et Chinois environ 1/5
5) Suédois, arabe, coréen et turc environ 0,15

*Tableau 1: Répartition linguistique des sites multilingues dans TRANCO*

| | %PAGES | %SITES |
|---|---|---|
| **TOTAL** | **100,00%** | **746,65%** |
| **Anglais** | 12,19% | **91,03%** |
| **Allemand** | 6,92% | **51,67%** |
| **Français** | 6,82% | **50,88%** |
| **Espagnol** | 6,42% | **47,95%** |
| **italien** | 4,67% | **34,88%** |
| **portugais** | 4,05% | **30,23%** |
| **Néerlandais** | 3,60% | **26,86%** |
| **polonais** | 2,94% | **21,93%** |
| **russe** | 2,93% | **21,87%** |
| **japonais** | 2,81% | **20,98%** |
| **Chinois** | 2,66% | **19,84%** |
| **suédois** | 2,12% | **15,82%** |
| **arabe** | 2,07% | **15,43%** |
| **coréen** | 1,99% | **14,88%** |
| **turc** | 1,93% | **14,42%** |
| **tchèque** | 1,78% | **13,28%** |
| **indonésien** | 1,73% | **12,93%** |
| **danois** | 1,66% | **12,36%** |
| **finlandais** | 1,59% | **11,89%** |
| **roumain** | 1,52% | **11,38%** |
| **hongrois** | 1,46% | **10,88%** |
| **ukrainien** | 1,44% | **10,72%** |
| **Grec moderne** | 1,43% | **10,64%** |

## 2-Tout le web

La question suivante se pose naturellement : ce modèle de présence linguistique sur les sites multilingues est-il applicable à l'ensemble du web ? L'objectif de cette étude est d'évaluer cette hypothèse.

Pour ce faire, nous nous appuyons sur les mesures réalisées par MECILDI dans plus de 40 ccTLD ou gTLD différents et nous élaborons des métriques pour extraire de ces différentes mesures la distribution standard des principales langues des sites web multilingues, sachant que la réponse unique dépend évidemment des pays et des langues principales, mais qu'il est possible d'observer une tendance dans une perspective globale.

## 2.1 La méthode

Nous avons répertorié dans un fichier Excel tous les résultats obtenus en limitant, pour chaque série, les langues les plus représentées à 10 %. Ensuite, nous avons testé différentes métriques pour fusionner tous les résultats.

- Nous avons d'abord constitué deux groupes. Le premier comprend toutes les premières mesures effectuées pour les gTLD de langue française et pour un groupe de 10 pays francophones , un groupe clairement orienté vers le français. Le second groupe, non centré sur un critère particulier, inclut 22 ccTLD de pays et régions divers. Nous avons sélectionné les 22 langues les plus représentées dans TRANCO, comme indiqué dans le tableau 1.
- Pour chaque groupe, nous avons collecté, pour chaque langue sélectionnée, le pourcentage pour chaque domaine et avons fixé la valeur à 0 % si la langue est absente.
- Nous avons refait le même travail, après avoir fixé à 0 le pourcentage correspondant à la langue officielle du pays.
- À partir des deux matrices (langue x domaine) où chaque cellule représente le %Sites collecté, nous avons appliqué différentes métriques :
    - moyenne simple par langue
    - Moyenne pondérée en fonction du nombre de sites web multilingues
    - Ensuite, nous avons effectué un mix incluant le million de visites les plus fréquentes et obtenu une moyenne pondérée finale en fusionnant les deux groupes.
- Nous avons construit une autre matrice classant chaque langue dans chaque domaine, en laissant une case vide en cas d'absence. À partir des données recueillies, nous avons calculé la moyenne des classements en ajoutant une pénalité de 1 pour chaque absence.

Finalement, nous obtenons le résultat final après fusion des résultats en faisant la moyenne des résultats des deux matrices, l'une conservant la langue officielle et l'autre la mettant à zéro.

## 2.2 Les résultats

Les résultats finaux sont présentés dans le tableau 2 ci-dessous, où les résultats TRANCO sont présentés avec les indicateurs de classement, l'indicateur de méthode pondérée appliqué à chaque catégorie et mélangé (W-MIX2) et les indicateurs de moyenne simple mélangés (A-MIX2).

L'objectif est désormais de vérifier si la hiérarchie Tranco est maintenue dans les différentes métriques. Si tel est le cas, nous pourrions conclure que, pour l'ensemble des 40 séries que nous avons gérées, la hiérarchie Tranco est confirmée. Bien entendu, cela ne garantit pas que cette conclusion soit valable pour l'ensemble du web, mais il pourrait s'agir d'une hypothèse approximative nécessitant une confirmation lors de son extension à d'autres domaines et à l'ensemble du web.

*Tableau 2: Résultats finaux des différentes métriques*

| | TRANCO | RANG | W-MIX2 | A-MIX2 |
|---|---|---|---|---|
| Anglais | 91,03% | 1,58 | **53,7%** | 87,3% |
| Allemand | 51,67% | 4,58 | **51,4%** | 54,5% |
| Français | 50,88% | 2,64 | **46,7%** | 47,8% |
| Espagnol | 47,95% | 4.26 | **22,2%** | 23,5% |
| Italien | 34,88% | 6,62 | **15,2%** | 16,4% |
| Portugais | 30,23% | 7.24 | **13,0%** | 14,2% |
| Néerlandais | 26,86% | 9,96 | **11,1%** | 12,3% |
| Polonais | 21,93% | 13.38 | **8,9%** | 9,6% |
| Russe | 21,87% | 14,63 | **7,8%** | 8,7% |
| Japonais | 20,98% | 14,94 | **7,2%** | 8,5% |
| Chinois | 19,84% | 13.45 | **6,7%** | 9,0% |
| Suédois | 15,82% | 22.19 | **5,5%** | 6,5% |
| Arabe | 15,43% | 15.41 | **5,3%** | 8,7% |
| Coréen | 14,88% | 20,00 | **3,5%** | 7,3% |
| Turc | 14,42% | 24h30 | **3,1%** | 6,0% |
| Tchèque | 13,28% | 23,58 | **2,9%** | 6,2% |
| Indonésien | 12,93% | 20,94 | **2,7%** | 6,3% |
| Danois | 12,36% | 27,33 | **2,6%** | 5,4% |
| Finlandais | 11,89% | 23.45 | **2,3%** | 6,1% |
| Roumain | 11,38% | 27,38 | **1,9%** | 7,2% |
| Hongrois | 10,88% | 29,00 | **1,7%** | 7,7% |
| Ukrainien | 10,72% | 31,00 | **1,6%** | 7,1% |
| Grec moderne | 10,64% | 27,25 | **1,6%** | 7,6% |

Les résultats montrent très peu d'exceptions à la constance du classement Tranco selon les différents indicateurs ; elles sont indiquées en rouge dans le tableau. L'indicateur le plus équitable est clairement le W-MIX2, qui présente un classement identique à celui de Tranco, à la différence près que l'anglais est relégué au même niveau que l'allemand et le français, et que l'espagnol se retrouve en tête du segment suivant.

Le français apparaît en rouge dans le classement, probablement en raison d'un biais évident d'une partie de l'échantillon en sa faveur. Le chinois et le japonais sont les deux langues affichant des résultats en rouge dans deux indicateurs. Nous estimons que l'échantillonnage Tranco est lui-même biaisé et sous-représente le chinois, ne couvrant pas les sites chinois très fréquentés en Chine (le pays comptant le plus grand nombre de personnes connectées). L'échantillonnage MECIDI est également biaisé en faveur de l'Afrique et, dans une moindre mesure, de l'Amérique latine, et présente une faible couverture de l'Asie. Les résultats du vaste domaine .asia (près d'un million de sites web) sont présentés dans le tableau 3 ci-dessous. Ils révèlent une nette domination du chinois, et il ne serait pas surprenant qu'en réalité, le chinois appartienne au troisième groupe avec l'italien, le portugais et le néerlandais.

*Tableau 3: Résultats MECILDI pour l'Asie (tous les sites web)*

| | %PAGES | %SITES |
|---|---|---|
| **.ASIE** | **100,00%** | **153,42%** |
| **Chinois** | **36,74%** | 54,74% |
| **Anglais** | **21,65%** | 32,51% |
| **Allemand** | **4,40%** | 6,82% |
| **Français** | **3,25%** | 5,09% |
| **vietnamien** | **3,24%** | 5,01% |
| **thaïlandais** | **3,03%** | 4,62% |
| **Espagnol** | **2,75%** | 4,35% |
| **japonais** | **2,25%** | 3,55% |
| **indonésien** | **2,16%** | 3,37% |
| **portugais** | **1,96%** | 3,12% |
| **italien** | **1,94%** | 3,11% |
| **russe** | **1,49%** | 2,31% |
| **bengali** | **1,37%** | 2,03% |
| **Néerlandais** | **1,31%** | 2,08% |
| **coréen** | **1,24%** | 2,00% |

L'autre langue signalée en rouge est l'arabe ; il est possible qu'elle ait été sous-représentée dans l'échantillon MECILDI et qu'elle appartienne donc au troisième groupe avec le polonais, le russe et le japonais, bien que les résultats MECILDI que nous avons recueillis jusqu'à présent pour les pays arabes soient nettement inférieurs aux attentes (voir le premier rapport d'étude). L'indonésien, d'après nos observations, apparaît assez fréquemment dans les ccTLD où sa présence était inattendue. Cela s'explique probablement par le fait que l'Indonésie héberge de nombreux sites web multilingues de jeux qui ont pour stratégie d'être accessibles par redirection depuis des sites web utilisant des domaines placés sous différents ccTLD.

# 3-Conclusions

Cette étude métrique suggère que la tendance observée dans la liste des millions de sites web les plus visités, concernant la hiérarchie des langues sur les sites multilingues, est applicable à l'ensemble du web. Le pourcentage élevé d'anglais pourrait être propre à TRANCO et, au niveau mondial, il pourrait toujours s'agir de la langue de prédilection pour les sites multilingues, mais probablement pas à 90 %, plutôt à un pourcentage plus proche des 50 % observés pour l'allemand, le français et l'espagnol.

Cela laisse subsister une réserve concernant le chinois qui pourrait se trouver, sinon au niveau de l'allemand, du français et de l'espagnol, du moins au même niveau que l'italien, le portugais et le néerlandais.

Le modèle final, assez courant sur les sites web multilingues du monde entier, pourrait finalement être le suivant :

| PRÉSENCE SUR DES SITES MULTILINGUES | PROBABILITÉ |
|---|---|
| Anglais | > 0,5 |
| Allemand, français, espagnol | ~ 0,5 |
| Italien, portugais, néerlandais, chinois | ~ 0,2 |
| Polonais, russe, japonais, arabe | ~ 0,15 |

Aujourd'hui, l'IA a réduit le coût marginal de l'ajout et de la maintenance de versions linguistiques supplémentaires sur un site web à des niveaux qui permettent d'obtenir plus facilement un retour sur investissement rapide.

Dans ce contexte, la tendance observée reflète une certaine paresse intellectuelle de la part des concepteurs web et des fournisseurs de contenu, qui pourraient obtenir un meilleur retour sur investissement en intégrant les langues locales. Notre premier rapport a révélé la rareté générale des langues locales. Dans de nombreux pays, l'ajout de langues aux sites web existants permettrait d'ouvrir considérablement le marché national, comme le montrent ces exemples révélateurs :

En Haïti, le créole est la seule langue comprise par 80 % de la population. Le taux d'accès à Internet est de 40 %. Aujourd'hui, moins de 10 % des sites web proposent une version créole. L'ajout d'une version créole à un site web permettrait d'atteindre plus de 30 % de la population haïtienne en tant que nouveaux clients.

Au Burundi , presque toute la population comprend le rundi, alors que moins de 1 % des sites web proposent une version en rundi et que 11 % de la population est connectée à Internet.

Au Paraguay, 88 % de la population parle le guarani, langue officielle du pays, et près de 80 % a accès à Internet. Pourtant, le domaine de premier niveau national .py ne compte que moins de 1 % de configurations linguistiques en guarani. Soixante-quatorze langues devancent le guarani sur les sites web multilingues. Dans ce contexte, n'est-il pas évident qu'un site de commerce électronique local en guarani pourrait générer des revenus importants ?

**RÉFÉRENCES :**


- Pages MECILDI donnant accès à tous les résultats publiés : https://obdilci.org/projets/mecildi/

- Pimienta D., Field M., 2026, *MECILDI : une plateforme informatique multilingue pour mesurer la distribution linguistique du contenu web* , version anglaise en cours d'évaluation par les pairs dans Plos One. Version française disponible à : https://www.obdilci.org/wp-content/uploads/2026/08/MECILDI-V1-FR.pdf

- OBDILCI, 2025, Études sur le multilinguisme en ligne. https://www.obdilci.org/projets/autres/mlreports-2/

- OBDILCI, 2026, Premier rapport sur les enseignements tirés de MECILDI : Le réseau désertique des langues autochtones, vernaculaires, locales et minoritaires, https://www.obdilci.org/wp-content/uploads/2026/09/WebLocal-fr.pdf

- OBDILCI 2026, Deuxième rapport sur les enseignements tirés de MECILDI : Révision des chiffres réels concernant la présence des langues sur le Web. https://www.obdilci.org/wp-content/uploads/2026/09/WebAfterMecildi.pdf

- OBDILCI 2026, Troisième rapport sur les enseignements tirés de MECILDI : Multilinguisme Web : analyse et proposition de typologie illustrées par 40 exemples (en anglais) https://www.obdilci.org/wp-content/uploads/2026/10/WebMultiL.pdf