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/

LESSONS LEARNT FROM MECILDI EXPERIENCE SERIES

TECHNICAL REPORT #4

The preferred languages of multilingual web

OBDILCI, October 2026

After having applied the MECILDI Multilingual-Aware Computational Framework for Measuring the Language Distribution of Web Content in close to 50 different web series (gTLD, ccTLD or Tranco), times has come to report on the lessons learnt from that experience. The following technical report is the fourth of this new series of technical reports elaborated by the Observatory of Linguistic and Cultural Diversity on the Internet (OBDILCI).

The computational framework developed by OBDILCI, with the support of France Ministry of Culture (DGLFLF) and Organisation de la francophonie (OIF), measures language distribution and multilingualism across any collection of websites. It has been applied first to the Tranco series of the million most visited web sites, to 8 gTLDs associated with languages of France, to 10 ccTLDs of member countries of OIF and to some 30 additional ccTLDs, selected as places with local languages counting important speakers' communities. Those measurements have allowed to confirm some previous figures about language distribution in the Web and to compute new indicators of multilingualism.

The first study (OBDILCI, 2026-1) focused the results for **indigenous, vernacular, local or minority** languages and offers a panorama particularly worrying of the tiny prevalence of those languages in the Web, with data much lower than the most pessimistic forecast.

The second study (OBDILCI, 2026-2) focused **the global picture** of indicators for language distribution and multilingualism indicators and the adjustments required to the OBDILCI model to reflect the lessons learned from the use of MECILDI in some 45 different series of websites (Tranco, ccTLD or gTLDs).

The third study offered the analysis of web multilingualism and from the characterization of the different situations encountered proposed **a typology of multilingual web** based on some key characteristics.

This fourth study focus the languages which appear to be preferred as a choice for additional linguistic configurations.

1-The million most visited sites

The MECILDI framework has allowed to produce the multilingualism indicators and the language distribution within the million most visited websites, referred to as TRANCO from the research group which maintain this list of websites (https://tranco-list.eu/). Further the language distribution and web multilingualism indicators, additional studies have been conducted around that specific series of websites, including the analysis of the subset of websites presenting the most frequent multilingual characteristics, from the use of the HTML instruction Hreflang= which specifies the list of languages supported by the website (which can possibly be declined also by country like for example Hreflang =en, en-us, en-uk, fr-fr, fr-ca).

One of those studies has been to separate from the 100 000 series of random websites used to compute the language distribution those using this Hreflang instruction and analyze the whole set of languages.

The results are presented in the following table, where %pages(l) is the percentage of pages in language l in the whole set of pages corresponding to the series of websites analyzed and %sites(l) is the percentage of websites having a configuration in language l. The sum of %pages for all languages is exactly 100% by definition and the sum of %sites is a number superior at 1 representing the rate of multilingualism of the web series. The top languages and their corresponding percentages are presented here below, for the top languages down to a value of 10% for %sites.

It shall be read this way:

- The rate of multilingualism of this series is 7.47
- English is present in 12% of the total webpages
- More than 90% of the websites in the series include an English version
- More than half of them have German and French versions.

Those results offer the list of languages the most prone to be used in the most visited websites and the corresponding probabilities of being used and define clearly 4 top categories:

1) English alone with 0.9
2) German, French and Spanish, around 1/2
3) Italian, Portuguese, Dutch between 1/3 and 1/4
4) Polish, Russian, Japanese and Chinese around 1/5
5) Swedish, Arabic, Korean and Turkish around 0.15

Table 1: Linguistic distribution of the multilingual sites in TRANCO

	%PAGES	%SITES
TOTAL	**100.00%**	**746.65%**
English	12.19%	**91.03%**
German	6.92%	**51.67%**
French	6.82%	**50.88%**
Spanish	6.42%	**47.95%**
Italian	4.67%	**34.88%**
Portuguese	4.05%	**30.23%**
Dutch	3.60%	**26.86%**
Polish	2.94%	**21.93%**
Russian	2.93%	**21.87%**
Japanese	2.81%	**20.98%**
Chinese	2.66%	**19.84%**
Swedish	2.12%	**15.82%**
Arabic	2.07%	**15.43%**
Korean	1.99%	**14.88%**
Turkish	1.93%	**14.42%**
Czech	1.78%	**13.28%**
Indonesian	1.73%	**12.93%**
Danish	1.66%	**12.36%**
Finnish	1.59%	**11.89%**
Romanian	1.52%	**11.38%**
Hungarian	1.46%	**10.88%**
Ukrainian	1.44%	**10.72%**
Modern Greek	1.43%	**10.64%**

2-The whole web

The following question arises naturally: is this pattern of language presence in multilingual sites applicable to the whole web? The objective of that study is to evaluate that hypothesis.

To do so we dispose of the measurement realized by MECILDI in more than 40 different ccTLDs or gTLDs and we elaborate some metrics to extract from those different measurements the standard top languages distribution of multilingual websites, knowing that obviously the unique answer depends on countries and main languages but it is possible to see a global pattern for the global perspective.

2.1 The method

We have listed in an excel file all the results obtained limiting for each series the top languages down to 10%. Then we have tried different metrics to merge all the results:

- First, we have separated two groups. The first group includes all the first measurements made for gTLDs of language of France and for a group of 10 countries of francophonie, a group clearly biased towards French. The second group is not centered to any criteria and include 22 ccTLDs from diverse countries or region. We have selected the 22 top languages in TRANCO as shown in table 1.
- For each group, we have collected for each selected language the percentage for each domain and set 0% if the language is absent.
- We have done the same work again, after setting to 0 the percentage corresponding to the official language of the country.
- From the two matrices (language x domain) where each cell is the collected %Sites we have applied different metrics:
    - Simple average per language
    - Weighted average with the number of multilingual websites
    - After that we have made a mix including the million most visited and obtain a final weighted average merging the 2 groups.
- We have built another matrix with the ranking of each language in each domain leaving a blank in case of absence. From the collected data, we have computed the average of the ranking adding a penalty of 1 for each absence.

Finally, we obtain the final result following merge results by averaging the results for the two matrix one leaving the official language one zeroing the official language.

2.2 The results

The final results is the table 2 below where the TRANCO results are presented together with the ranking metrics, the Weighted method metric applied to each category and mixed (W-MIX2) and the simple average metrics mixed (A-MIX2).

The objective is now to check if the Tranco hierarchy is maintained in the different metrics. If so we could conclude that, for the set of the 40 series we have managed the Tranco hierarchy is confirmed. Obviously, his does not warrant that this conclusion would be confirmed for the whole web, but it could be an approximative assumption needing confirmation while extending to more domains and reaching the whole web.

Table 2: Final results from the different metrics

	TRANCO	RANK	W-MIX2	A-MIX2
English	91.03%	1.58	**53.7%**	87.3%
German	51.67%	4.58	**51.4%**	54.5%
French	50.88%	2.64	**46.7%**	47.8%
Spanish	47.95%	4.26	**22.2%**	23.5%
Italian	34.88%	6.62	**15.2%**	16.4%
Portuguese	30.23%	7.24	**13.0%**	14.2%
Dutch	26.86%	9.96	**11.1%**	12.3%
Polish	21.93%	13.38	**8.9%**	9.6%
Russian	21.87%	14.63	**7.8%**	8.7%
Japanese	20.98%	14.94	**7.2%**	8.5%
Chinese	19.84%	13.45	**6.7%**	9.0%
Swedish	15.82%	22.19	**5.5%**	6.5%
Arabic	15.43%	15.41	**5.3%**	8.7%
Korean	14.88%	20.00	**3.5%**	7.3%
Turkish	14.42%	24.30	**3.1%**	6.0%
Czech	13.28%	23.58	**2.9%**	6.2%
Indonesian	12.93%	20.94	**2.7%**	6.3%
Danish	12.36%	27.33	**2.6%**	5.4%
Finnish	11.89%	23.45	**2.3%**	6.1%
Romanian	11.38%	27.38	**1.9%**	7.2%
Hungarian	10.88%	29.00	**1.7%**	7.7%
Ukrainian	10.72%	31.00	**1.6%**	7.1%
Modern Greek	10.64%	27.25	**1.6%**	7.6%

The result shows very few exceptions to the permanence of the Tranco ranking across the different metrics; they are shown in red in the table. The best metric in terms of fairness is clearly the W-MIX2 and this metric presents exactly the same ranking than Tranco, with the difference to downsize English almost at the same level as German and French and downsize Spanish to be the head of the next segment.

French is in red in Rank, probably a reflect of the obvious bias of part of the sample towards French. Chinese and Japanese are the two languages with red results in two metrics. We consider that the Tranco sampling is itself biased underrepresenting Chinese, lacking coverage of the Chinese sites heavily visited in China (the country with the highest population of connected persons). The actual MECIDI sampling is also biased in favor of Africa and in less extent Latin America and have low coverage of Asia. The results of the huge .asia domain (close to one million websites) are presented in table 3 below. They show a strong leadership for Chinese and it would not be surprising that in reality Chinese belongs to the third group with Italian, Portuguese and Dutch.

Table 3: MECILDI Results for Asia (all websites)

	%PAGES	%SITES
.ASIA	**100.00%**	**153.42%**
Chinese	**36.74%**	54.74%
English	**21.65%**	32.51%
German	**4.40%**	6.82%
French	**3.25%**	5.09%
Vietnamese	**3.24%**	5.01%
Thai	**3.03%**	4.62%
Spanish	**2.75%**	4.35%
Japanese	**2.25%**	3.55%
Indonesian	**2.16%**	3.37%
Portuguese	**1.96%**	3.12%
Italian	**1.94%**	3.11%
Russian	**1.49%**	2.31%
Bengali	**1.37%**	2.03%
Dutch	**1.31%**	2.08%
Korean	**1.24%**	2.00%

The other flagged in red language is Arabic; it could have been underrepresented in the MECILDI sampling and could therefore belong to the third group with Polish, Russian and Japanese, although the MECILDI results we have collected for Arabic countries so far are way below expectation (see the first study report). Indonesian, from what we have seen so far, appears quite often in ccTLD where it was not expected to be so high, and the reason is that it seems that Indonesia hosts many gaming websites which are multilingual and have the strategy to be reached by redirection from websites using domains placed under various ccTLD.

3-Conclusions

From this metrics study, it is reasonable to consider that the pattern discovered in the list of million most visited websites in terms of the hierarchy of languages in multilingual websites is applicable to the whole web. The huge percentage for English could be proper to TRANCO and worldwide percentage could still show English as the first choice for multilingual websites but probably not at the 90% level, but rather at a percentage closer to the 50% shown by German, French and Spanish.

It remains the reservation about Chinese which could be, if not at the level of German, French and Spanish, certainly together with Italian, Portuguese and Dutch.

The final pattern quite common in multilingual websites world-wide could finally be :

PRESENCE IN MULTILINGUAL SITES	PROBABILIDAD
English	> 0.5
German, French, Spanish	~ 0.5
Italian, Portuguese, Dutch, Chinese	~ 0.2
Polish, Russian, Japanese, Arabic	~ 0.15

Today, IA have dropped the marginal cost of adding and maintaining additional linguistic version in a website to levels which could more easily make the investment obtain fast return.

In this context, the observed pattern is the reflect of intellectual laziness from the web designers and content providers who could obtain better return on investment considering local languages. Our first report has revealed the general scarcity of local languages. In many countries, the selection of languages for addition in existing websites would open the national market in huge extent, as shown in those symptomatic examples:

In Haiti, creole is the only language understood by 80% of the population. The percentage of persons connected is 40%. Today the percentage of websites having a creole version is less than 10%. Adding a creole version to a website is opening access to more than 30% of the Haitian population as new customers.

In Burundi, almost all the population understand Rundi while there is less than 1% of websites having a Rundi version while 11% of the population is connected to the Internet.

In Paraguay, 88% of population speaks Guarani, official language of the country, close to 80% of the population have Internet access, however the .py ccTLD present less than 1% of linguistic configuration in Guarani. There are 74 languages with more presence than Guarani in multilingual websites. In this context, is not it obvious that a local ecommerce site in Guarani could make good business?

REFERENCES:

- MECILDI pages including access to all published results: https://obdilci.org/projects/mecildi/

- Pimienta D., Field M., 2026, *MECILDI: A Multilingual-Aware Computational Framework for Measuring the Language Distribution of Web Content*, in peer review stage in Plos One. A preliminary version is still accessible in: https://doi.org/10.13140/RG.2.2.33908.10886

- OBDILCI, 2025, Web multilingualism studies. https://www.obdilci.org/projects/other/mlreports/

- OBDILCI, 2026, First report on the lessons learned by MECILDI : The Desert Web of Indigenous, Vernacular, Local, and Minority Languages, https://www.obdilci.org/wp-content/uploads/2026/09/WebLocal.pdf

- OBDILCI 2026, Second report on the lessons learned by MECILDI : Revising the actual figures for languages presence in the Web. https://www.obdilci.org/wp-content/uploads/2026/09/WebAfterMecildi.pdf|

- OBDILCI 2026, Third report on the lessons learned by MECILDI : Web Multilingualism : analysis and typology proposal illustrated by 40 examples, https://www.obdilci.org/wp-content/uploads/2026/10/WebMultiL.pdf