



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 #2

Revising the actual figures for languages presence in the Web.

OBDILCI, September 2026

After having applied the MECILDI Multilingual-Aware Computational Framework for Measuring the Language Distribution of Web Content in some 30 different web series (gTLD, ccTLD or Tranco), times has come to draw the lessons learnt from that experience. The following technical report is the second of this new series of technical reports elaborated by the Observatory of Linguistic and Cultural Diversity on the Internet (OBDILCI). MECILDI has been funded by OIF and DGLFLF.

The computational framework developed by OBDILCI, that measures language distribution and multilingualism across any collection of websites, has been applied so far 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 Organisation de la francophonie (OIF) and to some 20 additional TLDs, 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.

This first study 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. See : [“The Desert Web of Indigenous, Vernacular, Local, and Minority Languages”](#), first report on MECILDI lessons learned series, Sep. 2026.

The second study will focus the actual figures of the presence of languages in the web and how they are impacted by the MECILDI results and what does that mean for the OBDILCI model (Pimienta et al., 2024).

1 – ENGLISH@WEB: THE CONFIRMATION BY THE FACTS

OBDILCI have stated in many opportunities, and since many years, that the main source, W3Techs, which is referenced by multiple reports, including sources considered serious such as Wikipedia and Statista, to claim a total domination of English in Internet content was biased in huge proportion. While W3Techs claim, since 2011, that the percentage of web content in English is between 50% and 60%, OBDILCI offers a figure between 20% and 26%, not a small difference.

OBDILCI have argued that the main bias is the consequence of not counting with the existence of a proportion of websites which are multilingual. In a peer reviewed article, Pimienta demonstrated the bias and even offer a mathematical correction (Pimienta, 2024).

MECILDI was first set to analyze TRANCO, the series of the million most visited websites which is used by W3Techs for its measurement, with a method which differ from W3Techs by paying due attention to the multilingual websites (Pimienta D., Field M., 2026).

The results of the measurements in terms of percentage of webpage is 20%, the figure which has been offered both by the OBDILCI model and by the mathematical demonstration. This is a powerful confirmation by the facts which should leave science minded persons totally clear on that debate.

Two side learned lessons emerged from this measurement;

- 1) The extrapolation of the results of language distribution in Tranco as a valid approximation for the whole web is not correct. The Tranco sample is biased by definition (the most visited websites would not reflect the language distribution of the whole web) but also by construction (the websites selection to belong to Tranco series is extremely biased in favor of European languages because the countries are not fairly represented, the methods favoring western countries, and therefore main European languages).
- 2) The measurement of the second indicator, beyond the percentage of webpages in a given language, is the percentage of websites having a linguistic configuration in a given language (note that the sum for all languages is not 100% but a figure we have called the rate of multilingualism of the series). One would expect that the figure given by W3Techs is equal to this indicator. As a matter of fact, the figure computed by MECILDI is sensibly

higher (67.5% versus 49,5%). It seems that W3Techs has defined somehow a concept of “main language of a website” and this is the object of its measuring. If the concept of first language (or mother tongue) and second languages applies to Humans, it hardly could be replicated in websites. The concept of “main language of a website” is in fact defined by the visitor of the website not by the designer. Websites have linguistic configurations and there is no hierarchy between them, except when it is designed to automatically select the language of the visitor’s browser. This ambiguity of what is really measured by W3Techs add to the bias already identified to make its indicator difficult to address.

As for the first lesson, the question of what is really the percentage of webpages in English in the whole web is difficult to answer firmly until some type of measurement is established, a venture highly complex due to the size of the web. I copy below an extract from (Pimienta, Field, 2026):

4.7 English Prevalence: Toward a Whole-Web Estimate

The Tranco corpus reflects the one million most visited domains globally, which are not a neutral sample of the web as a whole. It is therefore essential to ask whether the 20% English page share observed in Tranco is likely to be higher or lower than the true proportion of English across all approximately 200 million active websites globally.

Two competing forces shape these distributions. High-traffic websites tend to invest more heavily in multilingual infrastructure than the average website; accordingly, the rate of multilingualism observed in the Tranco sample (MultiR = 3.41) is much higher than would be expected for the web as a whole. Parallel research conducted by OBDILCI utilizing DataProvider.com database samples (OBDILCI, 2025b) established a lower baseline of %Multi = 14.4% for the entire web compared to the figure of 37% for Tranco.

While the average number of languages per multilingual site (AvgL) for the entire web remains unknown, a conservative estimate of 4.5—distinctly lower than the high-density Tranco average of 7.5—yields a plausible global MultiR of approximately 1.5. This closely mirrors the multilingual index of humanity itself, where the ratio of total speakers to native L1 speakers stands at 1.44 (based on 2024 Ethnologue speaker population data). Because a lower MultiR indicates fewer non-English linguistic variants per website, the relative proportional share of English pages naturally expands. This suggests that the true, whole-web English page share (%PAGES) is likely somewhat higher than the 20% observed in the Tranco corpus.

On the other hand, Tranco overrepresents Western European languages and underrepresents Chinese, which is the language with the largest connected speaker population globally. A less biased corpus would include a larger share of Chinese-language content and other non-European languages which would notably reduce English's proportional share.

External comparisons offer useful corroboration. Netsweeper (2023) measured language distribution directly at the webpage level across 12 billion pages—avoiding the multilingualism bias entirely—and found English at 26.3%, covering approximately 30% of the existing webpages.

This figure likely carries a positive bias toward English, given that Netsweeper's crawling infrastructure is primarily oriented toward industrialized countries. DataProvider.com (2023), despite not accounting for multilingualism, found English at 51.3%, compared to W3Techs' 57.7% at the same period. The 6.4 percentage point gap between these two single-language studies is attributable to corpus differences (Tranco versus a broader, whole-web sample). Applying a proportionally equivalent corpus correction to MECILDI's 20% baseline yields a whole-web estimate of approximately 18.7%.

Taken together, these comparisons support a revised estimate of English webpage prevalence across the entire web of approximately 19% to 26%, with the most probable value closer to 20% than to 26%. This range is consistent with both the OBDILCI model, which projects an English webpage prevalence of 20%, and the most recent empirical estimate reported by Pimienta (2024).

2 – OVERESTIMATION AND UNDERESTIMATION IN OBDILCI MODEL

The other learned lessons from the first series of MECILDI measurements concern the vision of language distribution in the whole web.

2.1 Local Languages

The first report has shown by convergent examples that the proportion of local language presence in the web is extremely low, and often null, in any case much lower than the estimation made by the OBDILCI model (see <https://www.obdilci.org/projects/main/>). The model needs to be readjusted to reflect this reality.

The final result is made from the weighted average of 5 indicators built by a complex process of weighting basic indicators:

- Usage: gathers indicators of usage by language of the Internet, focusing in particular social networks presence of speakers of the different languages.
- Traffic: collect from a large and balanced series of websites the distribution of traffic per country using metrics applications such as SimilarWeb.
- Interfaces: an indicator of technological support existing for each language, weighted indirectly by presence in translation service and main applications and, since V4.0, May 2023, complemented by an indicator of Digital Language Scale provided by Ethnologue.
- Indexes: gathering information society indicators pr country.
- Persons connected to the Internet per country: this last indicator playing additionally a special role in the model to help weight the other indicators depending of the presence of connected speakers.

The Interfaces indicator before V4.0 was very harsh on local languages, the great majority of which are totally absent from the sources analyzed and received therefore a value of 0. V4.0 allowed to reestablish some balance, and apparently offer a new advantage to those languages which is not

reflected in the reality. We will probably come back to the previous version, and if not enough to reflect the terrible reality of most local languages, play with the weighting associated with the different indicators to give more preeminence to Interfaces. At this stage the weighting is set to equal balance between the indicators but the model is prepared to have those value changed.

Arabic may be considered as a “local language” in Arabic speaking countries; its presence on websites is not comparable to the other local languages analyzed; however, in the first measurements made, the percentage of Arabic websites appears below the expectations. Therefore, some type of overestimation exists on the OBDILCI model, since English or French, depending on the country, takes over more websites than Arabic’s.

See below the results for Kuwait, Lebanon, Mauritania and Morrocco, in terms of the indicator %SITES (percentage of websites with presence of language).

COUNTRY	1st language	2nd language	3rd language
Kuwait	English (94%)	Arabic (77%)	French (7%)
Lebanon	English (81%)	Arabic (34%)	French (12%)
Mauritania	French (75%)	Arabic (61%)	English (36%)
Morocco	French (81%)	English (34%)	Arabic (23%)

2.2 Dominant European languages

The TRANCO measurements showed an interesting results corresponding to multilingual sites, in terms of the languages most privileged as linguistic options:

LANGUAGE	PRESENCE
English	91.03%
German	51.67%
French	50.88%
Spanish	47.95%
Italian	34.88%
Portuguese	30.23%
Dutch	26.86%
Polish	21.93%
Russian	21.87%
Japanese	20.98%
Chinese	19.84%

This hierarchy of presence, with English, first option in 90% of cases, then German, French and Spanish as second option in around 50% of the case, followed by Italian, Portuguese, Dutch and Polish, from 1/3 to ¼ of the cases, seems to be a replicable pattern across the different TLD analyzed so far.

This fact implies that the respective weight of German, Italian, Dutch and Polish in terms of connected speakers maybe not reflecting the real presence of those languages in web content and some corrections are required in the model. There is a way to better reflect those languages dominance in the model results, with the addition of a new indicator : the number of domains below the ccTLD of each country, as this hierarchy is roughly linked to this missing indicator.

This addition, however, will create some challenges, due to the combination of several factors:

- There are a limited number of sources for the figure of the number of domains below each TLD : the most reliable, CENTR, is limited to a very small subset, two commercial ones do cover the complete list of TLD, [DomainMetaData](#) and [Zonestats.io](#), but they propose figures which differ, sometimes largely. Other sources ([Statista](#)¹, [TLD](#)) have very limited coverage. The best guess at this stage, would be to take the higher figures for each domain from all the proposals. In any case, it appears clearly that .de, the ccTLD of Germany is one of the most numerous domains and that .nl and .it have high figures compared to their respective populations, see below the CENTR data.

TLD	registered	last_updated
.de	18 150 432	Sep-2026
.uk	10 657 724	Sep-2026
.nl	6 135 231	Jul-2026
.br	5 915 998	Sep-2026
.fr	4 568 249	Sep-2026
.au	4 416 817	Aug-2026
.eu	3 670 879	Sep-2026
.it	3 624 600	Sep-2026
.ca	3 589 744	Jul-2026
.ch	2 654 560	Sep-2026
.es	2 237 085	Sep-2026
.pt	2 179 311	Sep-2026
.be	1 708 285	Sep-2026
.cz	1 593 675	Sep-2026

- There is always a huge gap between the number of registered domains and the number of active websites associated. This gap may not be homogenous across the TLDs.
- The US have an old habit not to use their ccTLD (.us) for the majority of domains, preferring the use of gTLDs such as .com or .edu. This poses a problem of how to treat this country supposed to have the highest number of active websites. Note that .com is the

¹ Note that STATISTA is not the real source but it keeps the unethical habit to hide the very source unless you pay for the service, a practice we and others have denounced (<https://bernoff.com/blog/my-big-problem-with-statista>).

major domain in terms of number of subdomains, gathering more than half the total of existing domains.

In any case, dealing with those types of challenges has been the routine in the building of the model and we trust we will find a way to transform the last result of the model :

	% INTERN. L1+L2	% WORLD POP L1+L2	% CONN. Speakers	% USAGE L1+L2	% TRAFIC L1+L2	% INTERF. L1+L2	% INDEXES L1+L2	% CONTENT L1+L2
English	15.48%	14.13%	70.66%	25.96%	21.18%	20.17%	17.60%	20.08%
Chinese Macro	17.57%	14.48%	78.23%	12.88%	22.42%	22.88%	19.48%	19.04%
Spanish	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%
Russian	3.31%	2.38%	89.53%	4.37%	3.74%	3.75%	3.59%	3.75%
Arabic Macro	4.36%	4.08%	68.79%	3.01%	2.39%	4.94%	3.55%	3.65%
French	3.05%	2.91%	67.75%	4.11%	2.86%	3.98%	3.11%	3.42%
Portuguese	2.97%	2.46%	77.76%	3.36%	3.04%	3.58%	2.83%	3.16%
Japanese	1.55%	1.15%	87.02%	3.30%	2.05%	1.97%	2.26%	2.23%
German	1.80%	1.25%	93.00%	2.76%	1.39%	2.24%	2.58%	2.15%
Malay Macro	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%
Turkish	1.13%	0.84%	86.61%	1.26%	0.86%	1.30%	1.13%	1.14%
Italian	0.85%	0.62%	87.59%	1.28%	0.62%	1.07%	1.12%	0.99%
Vietnamese	0.98%	0.80%	78.47%	0.87%	1.11%	0.94%	0.86%	0.95%
Korean	0.77%	0.76%	66.00%	1.18%	0.70%	0.91%	0.78%	0.87%
Persian Macro	1.00%	1.03%	62.89%	0.62%	0.99%	0.90%	0.79%	0.86%
Urdu	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%
Thai	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%
Polish	0.53%	0.39%	87.01%	0.63%	0.39%	0.58%	0.67%	0.56%
Tamil	0.63%	0.81%	50.59%	0.42%	0.56%	0.52%	0.56%	0.54%
Javanese	0.68%	0.64%	69.26%	0.45%	0.50%	0.35%	0.60%	0.52%
Ukrainian	0.47%	0.36%	83.93%	0.47%	0.32%	0.45%	0.52%	0.45%
Dutch	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 Niger.	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%

The new results of the model would probably see the languages marked in red getting down in the ranking and those in blue getting higher to better reflect the reality which has been discovered by the MECILDI measurements.

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>
- Web multilingualism studies. <https://www.obdilci.org/projects/other/mlreports/>
- Pimienta D., 2024, “Is it true that more than half the web contents are in English? Not If Multilingualism Is Paid Due Attention!”, 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, The method behind the unprecedented production of indicators of the presence of languages in the Internet. Frontiers Research Metrics & Analytics, Volume 8 – 2023. <https://doi.org/10.3389/frma.2023.1149347>