



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 #1

The desertic web of indigenous, vernacular, local or minority languages

OBDILCI, September 2026

After having applied the MECILDI Multilingual-Aware Computational Framework for Measuring the Language Distribution of Web Content in 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 first 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, 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 11 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.

This first study focuses 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. In the following, we will use term “local” for any of the mentioned categories of language.

The following table offers the compilation of data obtained from the MECILDI run, completed with appropriate indicators so to establish fair comparisons and sorting.

The column of the table has the following definitions:

LANGUAGES : the English name of the local language

ISO: The ISO-639-3 code associated with this language

TLD: The top-level domain measured by MECILDI.

PLACE: The place (country or region) corresponding to this TLD

%SP: Percentage of L1+L2 speakers of this language in the place. (Source: Ethnologue)

L1+L2: Number of speakers on this language in the place or, otherwise indicated, in the World of a specific country. (Source: Ethnologue)

%P: % of webpages in this language among the total webpages in the TLD zone. (MECILDI)

%S: % of websites having a linguistic configuration in that language. (MECILDI)

WRate: The ratio between %SP over %S, indicator of prevalence, 1 being strict proportionality.

The table is sorted by decreasing WRate.

LANGUAGES	ISO	TLD	PLACE	%SP	L1+L2	%P	%S	WRate	N
Basque	eus	.eus	Pays Basque	36%	1M	36.37%	94.02%	2.612	1
Catalan	cat	.ad	Andorra	37%	30900	20.75%	61.68%	1.672	2
Breton	bre	.bzh	Bretagne	5%	200K (W)	2.91%	2.19%	0.438	
Papiamentu	pap	.cw	Curacao	92%	136K	11.52%	18.47%	0.201	
Tahitian	tah	.pf	French Polynesia	19%	56K	1.69%	2.67%	0.144	
Malgache	mlg	.mg	Madagascar	58%	18M	4.61%	7.58%	0.131	
Corsican	cos	.corsica	Corsica	35%	120K (W)	3.55%	3.68%	0.105	
Haitian Creole	hat	.ht	Haiti	95%	13.8M (W)	3.41%	7.89%	0.083	3
Bislama	bis	.vu	Vanuatu	34%	115K (W)	0.99%	1.60%	0.047	
Swahili	swc	.cd	Dem. Rep. of Congo	25%	32M	-	0.53%	0.021	4
Yiddish	yid	.il	Israel	5%	443K (W)	0.02%	0.03%	0.006	
Rundi	run	.bi	Burundi	98%	13M (W)	-	0.56%	0.006	
Guarani	grn	.py	Paraguay	88%	6M	0.13%	0.26%	0.003	5
Quechua	que	.pe	Peru	21%	7.3M	0.04%	0.06%	0.003	
Guadeloupean Creole	gcf	.gp	Guadeloupe	90%	400K (W)	0.07%	0.18%	0.002	6
Tamazight	zgh	.ma	Maroc	25%	9M	-	0.04%	0.002	
Kabuverdianu	kea	.cv	Cabo Verde	82%	492K	ε	ε	0.000	7
Ladino	lad	.il	Israel	0.5%	43K	ε	ε	0.000	8
Alsatian	gsw	.alsace	Alsace	35%	900K (FR)	0	0	0	
Aymara	aym	.bo	Bolivia	14%	1.7M	0	0	0	
Jula	dyu	.ci	Côte d'Ivoire	27%	8.5M	0	0	0	
Kanak languages		.nc	Nouvelle Calédonie	36%	71K	0	0	0	9
Kimbundu +Various		.ao	Angola	90%	33M	0	0	0	10

NOTES

- 1: Almost perfect bilingualism with Spanish
- 2: Somehow balanced share between Catalan, English, Spanish and French.
- 3: 80% of population only speaks creole
- 4: Among many other indigenous such as Lingala (lin) 40M, all absent
- 5: Guarani is official language of Paraguay
- 6: What was detected is Haitian creole
- 7: A unique site was found over 1100 processed
- 8: Four websites having ladino (lad) where found
- 9: Drehu (12K), Wallisian (18K) & 32 other Kanak languages.
- 10: Kimbundu 3.7M, Umbundu 5.9M, Chokwe 2.4M, Kikongo 2.4M, Nhaneka 1.4M, Nganguela 700K, Kwanhama 1M, Fiote 391K, Muhumbi 253K, Luvale 180K

What does those data are telling?

Except for the cases of Basque and Catalan, which show impressive vitality in the web ecosystem, all the local languages studied indicate extremely low presence and vitality in the web. The WRate, taken as indicator of the prevalence of the local language within the corresponding TLD shows null or almost null values in one third of the local languages, critical values, lower than 0.01, in more than half of the cases, and very low values (between 0.1 and 0.5) for 20% of the local languages.

Some thirty analyzed places are obviously not enough to draw statistical definitive conclusions, however, it is sufficient to raise legitimate preoccupation.

OBDILCI, from its previous studies, was expecting to find most results within the category “very low values”, not in the critical and null values.

Examples such as the following establish the size of the problem which is faced and the urgent need for appropriate public policies :

- **Guarani** is official language of Paraguay, receiving growing public policies attentions, it is spoken by 88% of Paraguay’s population, yet there is only 0.26% of websites having a Guarani configuration, e.g. few units.
- In Haiti, 80% of the population only understand **Creole**, that huge proportion of Haitian population have linguistic access to less than 8% of the nation-wide web ecosystem.
- A Bolivian scientist have created and developed, since the late 80`s, a remarkable and quite original translation algorithm using **Aymara** as pivot language¹, yet there is absolutely no web site in that language, however spoken by 14% of the population.

¹ <https://www.obdilci.org/blog/ivan-guzman-de-rojas-and-the-aymara-language/>

- **Papiamentu** is the dominant language of notably multilingual island of Curacao; more than 90% of population use it daily, yet the Curacao web holds just a little more than 10% of pages in that language.
- **Kabuverdianu** is the dominant language of Cabo Verde, spoken by more than 80% of population, versus 60% for Portuguese, yet the unique website identified is from the Institute of Cultural Patrimonial.
- **Rundi** is the main language of Burundi, understood by virtually all inhabitants of the country, yet it is present in less than 1% of websites in the .bi domain.

We are close to the middle of the *Indigenous language decade* promoted by UNESCO; many organizations, such as OIF, are organizing experts' workshops to face the challenges posed by IA to fully integrate local languages; a growing number of languages (maybe 750) have received attention from UNICODE to receive the codification required for digital existence, but what about **contents** in those languages?

METHOD

We have applied the MECILDI framework to the complete list of identified websites below the TLD of the mentioned countries or regions. In some cases, we have used the official zone of the TLD, when shared by the network information center, thru ICANN/CZDS; in the remaining cases we have used a commercial provider (zonestats.io). In the second case, the list of domains may have lacked some recent websites with local languages, but the statistic result would have remained anyway very close.

MECILDI explores systematically all the domains below the TLD, again a significant part is excluded from measurement for many possible reasons, including the will of the web owner not to be processed by robots, but again this situation would not alter the statistical results in a significative manner.

All detailed results are publicly consultable, see references, except for the last series of ccTLDs, which have not been published yet, the example of Cabo Verde (.cv) is provided in annex to allow the understanding of the process and the type of indicator obtained.

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>

ANNEX: The full results for Cabo Verde (.cv)

INDICATORS	Indicator definition
%SITES (L)	Websites having a linguistic version in language L
%PAGES (L)	Percentage of web pages in L (number of sites in L divided by number of linguistic versions)
MultiR	Rate of Web multilingualism (number of linguistic versions divided by number of web sites)
%Multi	Percentage of multilingual websites
AvgL	Average number of languages per multilingual web sites
ERR-TOT	Total percentage of web sites non processed
ERR-DNS	Domain not accessible
ERR-BLK	Valid sites but not allowing process from program
ERR-CFG	Parked or under construction site
ERR-HTTP	Web site error
ERR-LANG	Valid site but language detection failure
MonoL	Percentage of monolingual web sites
GooT	Percentage of web sites using Google Translate widget
HDEF	Heuristic Derived Extrapolation Factor.
Lang=	Percentage of websites using the Lang= HTML attribute
HrefLang=	Percentage of websites using the HrefLang= HTML attribute
IDN	Percentage of sites with Internationalized Domain Names
NON-ID	Language non identified by the detection algorithm

Number of analyzed sites : **2 429**

Number of sites validated for the process: **1 212**

MultiR	1.79
%Multi	23.98%
AvgL	4.30
MonoL	76.02%
GooT	1.97%
HDEF	2.62
Lang=	79.51%
Lang=Err	11.39%
IDN	0.00%
ERR-DNS	18.37%
ERR-BLK	2.87%
ERR-CFG	1.76%
ERR-HTTP	6.03%
ERR-LANG	53.10%
ERR-TOT	82.13%

LANGUAGES DISTRIBUTION	%PAGES	%SITES
TOTAL	100.00%	178.48%
English	33.16%	59.18%
Portuguese	25.12%	44.83%
French	5.86%	10.45%
Indonesian	4.44%	7.93%
Vietnamese	3.91%	6.97%
Spanish	2.10%	3.76%
German	1.80%	3.21%
Chinese	1.68%	3.01%
Dutch	1.19%	2.12%
Japanese	1.15%	2.05%
Italian	1.03%	1.84%
Russian	0.96%	1.71%
Arabic	0.84%	1.50%
Korean	0.80%	1.43%
Turkish	0.73%	1.30%
Polish	0.69%	1.23%
Thai	0.61%	1.09%
Bulgarian	0.50%	0.89%
Czech	0.46%	0.82%
Malay	0.46%	0.82%
Romanian	0.46%	0.82%
Swedish	0.46%	0.82%
Danish	0.34%	0.61%
Modern Greek	0.34%	0.61%
Finnish	0.34%	0.61%
Ukrainian	0.34%	0.61%
Latin	0.27%	0.48%
Persian	0.27%	0.48%
Bengali	0.23%	0.41%
Estonian	0.23%	0.41%
Hindi	0.23%	0.41%
Hungarian	0.23%	0.41%
Javanese	0.23%	0.41%
Latvian	0.23%	0.41%
Lithuanian	0.23%	0.41%
Mongolian	0.23%	0.41%
Burmese	0.23%	0.41%
Norwegian	0.23%	0.41%

Serbian	0.23%	0.41%
Tagalog	0.23%	0.41%
NON-ID	0.19%	0.34%
Basque	0.15%	0.27%
Irish	0.15%	0.27%
54+ languages	0.11%	0.20%

Notes:

- ISO639-3 code for Kabuverdianu (kea) is not present in Lang of Hreflang HTML fields.
- The country code (cv) may have been used as a language code in rare opportunities, such as <https://ipc.cv/cv/>.
- Indonesian sites appear in redirections, mostly for gaming websites.
- The .cv is sold worldwide for people to store their professional curriculum with the format JohnSmith.cv. The manager of those sites has marked them as non-accepting robot browsing and therefore have not affected the results, except for a huge ERR-LANG percentage as it was impossible for MECILDI to address their content.
- Those sites represent more than half of total websites.