WebMultilingualism Reports #1: Approximation of the Rate of multilingualism of the WWW

CONTEXT

Dataprovider.com maintains a database gathering information related to a series of websites close to the total WWW (80% if compared with the figure given by Netcraft: 872 920 299 over 1 101 431 853). Among the many information kept for each web site, there is the unique main language of the site (often taken from the home page) and the presence of hreflang tags, which a proportion of sites uses to list the linguistic versions of the site with the format hreflang= ll; ll-cc;… where ll is the ISO639-2 and cc a country code with 2 digits[1].

Dataprovider.com has offered a courtesy access of the database to allow OBDILCI to develop useful statistics on the subject of languages online, in coherence with its non-for-profit mission. This report is the first of a series of report of the findings of OBDILCI obtained thanks to Dataprovider.com database.

The data base of Dataprovider.com is extremely impressive, both by its uncommon extraordinary reach, a large proportion of the whole WWW, close to completion, and the ease of use, friendliness and powerfulness of the interface which allow searches using close to 100 different parameters with unlimited combinations and immediate response.

SCOPE OF THE STUDY

A key, yet unknown, indicator of the Internet, in terms of linguistic diversity, is the **rate of multilingualism** (**MRate**) defined as the ratio between the total number of linguistic versions and the total number of web sites (in other term the average number of languages per web site). The two following indicators allow to compute MRate in the following formula:

MRate = 1+MULTI% x AVG-ML

- The percentage of websites having more than one linguistic version (MULTI%)
- The average number of linguistic versions per multilingual web sites (AVG-ML)

[1] See https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes

BACKGROUND

The Ionan University Visual Arts Department study about the prevalence of English in websites of ccTLD of European Union countries (Giannakoulopoulos, A and al., 2020) produced the following results for the specific space targeted and for 2019:

MULTI% = **10.99%** (figure computed precisely for each ccTLD and weighted with the total number of sites per country to obtain the global figure)
AVG-ML = **4.91** (same)
MRate = **1.136** (same)

The new study using Dataprovider.com presents new figures:
MULTI% = **17.00%** (figure computed precisely for each ccTLD and weighted with the total number of sites per country to obtain the global figure)
AVG-ML = **4.91** (unknown figure taken by default from the previous study as a conservative figure)
MRate = **1.834** (same as MULTI%).

OBDILCI used the results of the Ionan university study, including the raw calculations, to support a study about the bias resulting of non-considering the fact that many web sites have several linguistic versions (Pimienta D., 2024). The sample of web sites used to convey that study encompasses 102 149 sites. The main results of the Ionan university study, as related to languages online, were that, in 2019, **28.4%** of web sites have and English version, the rate of multilingualism (MRate) of the ccTLD of EU countries (before Brexit) was 1.14 with an average 4.9 different languages per multilingual website.

This report will compute the same data in 2025 using a much wider reach of web sites.

Warning: as a consequence of the change of method, the comparison is to be taken with caution, being undecidable whether the changes are totally due to the evolution of the web realm over the 6 elapsed years or if they also reflect the unavoidable **selection bias** in both studies.

MAIN RESULTS:

English proportion is now reduced at **19.76%** (30% decrease), the MRate of the sample increases 50% to **1.83**. The average number of linguistic versions is not currently decidable and has been lest identical to allow computation of the MRate. The new sample includes 15 779 643 web sites.

Table 1: Main results

	IONAN UNIV.	DATAPROVIDER.COM /OBDILCI
DATE	2019	2025
SAMPLE (number of sites)	102 149	15 779 643
MRate	1.136	1.834
AVG-ML	4.91	?
ENGLISH	28.4%	19.76%
MULTI%	10.96%	17.00%

Notes:

For Ionan Univ., the average results vary depending on the method of counting: simple average, weighting by population, weighting by number of sites explored in each country and the one which has been chosen, weighting by total number of sites existing in the country, which results in the lowest figures for MULTI% and AVG-ML and the highest for English %.

For Dataprovider.com, this is an elaboration made by OBDILCI from raw data plus some computation assumptions described hereafter.

METHOD

The sample of data collected from Dataprovider.com database is made by:

1) selecting Response = available (this reduces the total of sites from 872 920 299 to 166 058 954, by eliminating websites which are not in use[2])
2) Selecting sites which are "developed", avoiding therefore sites with error and site which are parked (e.g. hold for future use). The figure declines to 78 847 894 which will represent the working sample.

The process is :
Complete the existing table created for the (Pimienta, 2024) study and insert new data for each corresponding ccTLD, after collecting the following information:

- From Dataprovider.com, the total number of sites (T), the number of sites in English and the percentage of sites having Hreflang tag present in the HTML (Nh)
- From https://domainnamestat.com/statistics/tld/others the number of sites registered, for cross checking purpose.

Hreflang is used to a proportion of multilingual sites say Pm, the indicators are computed with the following operations: MULTI% = Nh x 1/(1-Pm) and MRate = 1+ MULTI% x AVG-ML

From manual experience made in (Pimienta, 2024), confirmed by queries to ChatGPT Pm has been set to 40%. AVG-ML cannot be computed at this stage and its value has been retaken from the Ionan study, as a conservative data. In order to understand the sensitivity of the data, some runs with other values of Nh and Pm are shown:

Table 2: Sensitivity of AVG-ML factor

AVG-ML	Pm	MRate
5	35%	1.97
6	30%	2.36
4	40%	1.68
4.9	**40%**	**1.834**
6	40%	2.02
4	50%	1.54
4	60%	1.45
3	60%	1.34

[2] Note that the equivalent Netcraft figure is 192 282 517, the covering would be then of 86%.

Both parameters are quite sensitive on the MRate and we will look forward to get, if possible, some data to strengthen precision and confidence on them. The MRate is a key factor of the evolution of multilingualism in the Internet. Note that the trend of the use of GoogleTranslate imbedded in websites could easily boost that figures, given that only 1% of such usage would add 2.5 to MRate (because it adds 250 languages). In any case, it is a parameter required to un-bias the measure of English made by W3techs, as demonstrated in (Pimienta, 2024).

Hereafter, is the table reduced with the main data. The totals are computed by weighting the data with the number of domains in order to maintain the relative importance of each country. The columns in green are the data from Ionan university and the columns in brown the data collected from Dataprovider.com. The figures in red in DomainStat appeared wrong or missing.

Table 3: Measurement comparison

TLD	Country	Sites	AVG-ML	MRate	English	DomainStat	Domain	English	HREF
at	Austria	4 063	3.50	1.131	12.0%	**2 197 489**	424 985	24 780	32 655
be	Belgium	4 063	3.34	1.227	**20.7%**	**2 454 152**	470 801	60 604	60 244
bg	Bulgaria	3 178	3.70	1.346	**25.7%**	**277 449**	38 899	4 206	6 822
hr	Croatia	3 489	3.85	1.363	**25.0%**	**167 347**	66 460	9 623	9 044
cy	Cyprus	828	5.11	1.402	59.9%	**16 076**	6 771	4 628	1 387
cz	Czechia	4 084	3.76	1.156	12.5%	**2 011 885**	565 083	18 629	29 803
dk	Denmark	4 067	3.19	1.158	**17.4%**	**1 827 738**	347 506	32 247	18 418
ee	Estonia	3 556	3.50	1.471	**21.9%**	215 250	75 525	8 875	13 361
fi	Finland	3 992	3.70	1.220	**17.0%**	**716 183**	201 999	19 906	22 854
fr	France	4 125	6.42	1.127	8.9%	**8 667 628**	1 379 860	58 341	82 105
de	Germany	4 150	4.10	1.098	9.7%	**27 083 237**	4 184 798	178 166	222 110
gr	Greece	3 953	3.70	1.357	30.2%	**658 538**	237 111	51 295	50 686
hu	Hungary	3 993	3.30	1.172	**14.1%**	**1 194 260**	322 815	17 655	19 225
ie	Ireland	3 825	3.50	1.014	98.3%	**479 441**	110 428	105 853	3 539
it	Italy	4 123	4.10	1.220	**15.3%**	**5 859 250**	1 172 271	88 050	123 552
lv	Latvia	3 406	3.25	1.527	**25.7%**	**189 691**	44 085	6 031	9 922
lt	Lithuania	3 773	3.38	1.339	**21.1%**	**319 107**	75 028	6 330	11 405
lu	Luxemburg	2 876	3.72	1.362	**25.9%**	**149 793**	22 128	5 400	5 009
mt	Malta	444	6.60	1.095	89.5%	**19 637**	4 117	3 780	290
nl	Netherlands	4 133	3.39	1.140	**14.2%**	**8 667 628**	1 475 505	168 840	117 035
pl	Poland	4 110	3.64	1.120	10.2%	**4 359 431**	788 235	36 006	51 792
pt	Portugal	4 084	3.80	1.163	**18.3%**	**545 400**	145 896	15 823	20 296
ro	Romania	3 975	3.57	1.314	**27.9%**	**1 042 223**	272 838	38 508	23 093
sk	Slovakia	3 943	3.34	1.190	**14.3%**	**740 532**	203 334	7 429	14 148
si	Slovenia	3 619	3.60	1.286	**20.6%**	49 558	65 413	5 497	9 198
es	Spain	4 088	3.44	1.162	11.9%	**2 958 675**	515 874	39 221	53 953
se	Sweden	4 084	3.00	1.146	**16.7%**	**3 036 275**	487 922	53 437	21 728
uk	UK	4 125	7.83	1.013	98.4%	**23 764 012**	2 073 956	2 049 102	39 566
	Total	**102 149**	**4.91**	**1.136**	**28.4%**	**99 667 885**	**15 779 643**		

The countries with the higher multilingualism rate in the web were, in 2019, by order of importance: Latvia, Estonia, Cyprus, Croatia, Luxemburg, Greece, Bulgaria, Lithuania and Romania. Those with lower rate were: UK, Ireland, Malta, Germany, Poland and France.

In 2025, the highest**: Luxemburg, Latvia, Cyprus and Estonia, Bulgaria, Slovenia and Portugal** and the lowest**: UK, Ireland, Malta, Sweden, Germany, Denmark, Czechia and France.**

The highest rate of English in 2019 were: UK, Ireland, Cyprus and Greece; the lowest: France, Germany, Poland, Spain, Austria and Czechia.

The highest rate of English in 2025: **UK, Ireland, Malta, Cyprus and Greece; the lowest: Czechia, Slovakia, France, Germany, Hungary, Poland and Austria**.

Table 4: Results comparison

ccTLD	English %	HREF%	COVER	ENG2019 Vs ENG2025	MRate	MULTI%
at	5.83%	7.68%	19.34%	-51.30%	1,94	19,21%
be	12.87%	12.80%	19.18%	-37.71%	2,57	31,99%
bg	10.81%	17.54%	14.02%	-58.01%	3,15	43,84%
hr	14.48%	13.61%	39.71%	-41.98%	2,67	34,02%
cy	68.35%	20.48%	42.12%	14.18%	3,51	51,21%
cz	3.30%	5.27%	28.09%	-73.56%	1,65	13,19%
dk	9.28%	5.30%	19.01%	-46.76%	1,65	13,25%
ee	11.75%	17.69%	35.09%	-46.27%	3,17	44,23%
fi	9.85%	11.31%	28.20%	-42.09%	2,39	28,28%
fr	4.23%	5.95%	15.92%	-52.30%	1,73	14,88%
de	4.26%	5.31%	15.45%	-56.10%	1,65	13,27%
gr	21.63%	21.38%	36.01%	-28.38%	3,62	53,44%
hu	5.47%	5.96%	27.03%	-61.29%	1,73	14,89%
ie	95.86%	3.20%	23.03%	-2.51%	1,39	8,01%
it	7.51%	10.54%	20.01%	-51.05%	2,29	26,35%
lv	13.68%	22.51%	23.24%	-46.73%	3,76	56,27%
lt	8.44%	15.20%	23.51%	-59.95%	2,86	38,00%
lu	24.40%	22.64%	14.77%	-5.92%	3,78	56,59%
mt	91.81%	7.04%	20.97%	2.58%	1,86	17,61%
nl	11.44%	7.93%	17.02%	-19.40%	1,97	19,83%
pl	4.57%	6.57%	18.08%	-55.07%	1,81	16,43%
pt	10.85%	13.91%	26.75%	-40.88%	2,71	34,78%
ro	14.11%	8.46%	26.18%	-49.49%	2,04	21,16%
sk	3.65%	6.96%	27.46%	-74.53%	1,85	17,40%
si	8.40%	14.06%	131.99%	-59.29%	2,73	35,15%
es	7.60%	10.46%	17.44%	-35.96%	2,28	26,15%
se	10.95%	4.45%	16.07%	-34.43%	1,55	11,13%
uk	98.80%	1.93%	8.62%	0.39%	1,23	4,77%
TOTAL	**19.76%**	**6.80%**		**-30.46%**	**1,83**	**17,00%**

The HREF% is an indicator of the percentage of multilingual sites (Multi%). If the assumption of Pm = 40% is the most probable, then the Multi% in EU ccTLD would be now of 17%.

Based on Dataprovider.com database, the ccTLDs of European Union countries, plus .uk, show a proportion of English of **19.8%** and an average proportion of multilingual sites of **17%**. If the average number of languages per multilingual site remains **4.9**, then the rate of multilingualism (number of linguistic versions divided by number of site) is **1.83**, higher than the Human multilingual rate (1.44 from Ethnologue). Those figures show a decrease for English and a growth for multilingualism parameters compared to a 2019 study, however the comparison has to be observed with caution as it is undecidable if the difference is due to the selection bias or to the evolution over time.

RATE OF MULTILINGUALISM OF THE WWW

Is it possible to compute the same parameters for the whole WWW using Dataprovider.com hreflang data collection?

The Hreflang parameter allow to infer the Multi% parameter with the rule of extending it on the assumption that 40% only of multilingual sites use that approach. In the current state of the data base it is possible to approximate the percentage of multilingual sites worldwide, but it is difficult to determine the average number of languages per multilingual site, which prevents to offer a close figure for MRate.

Following the assumption that Hreflang is used by 40% of multilingual sites, the percentage of multilingual site for the whole WWW is a figure around **12%**. If the average number of languages per website is confirmed for the whole Web at 5, the MRate for the whole web would then be of the value of **1.6**. At this stage we do not have a way to compute precisely this figure for the whole WWW so this estimate is to be taken with caution as tentative.

PROPERTY OF WEB MULTILINGUALISM

Using the same href data collection from Dataprovider.com and the same method of inferring Multi%, and crossing in the data base with other parameters it is possible to correlate the parameters with the level of multilingualism, targeting the criteria propense for multilingualism (those with Multi% higher than the global average of 12%), and those which are at the contrary more propense to monolingualism.

Table 5: Variability of Multi% depending on context

SAMPLE	%Hreflang	Multi%
70<Economic footprint[3]<79	28.3%	**70.75%**
60<Economic footprint<69	27.4%	**68.50%**
Economic footprint > 89	24.5%	**61.25%**

[3] Citation from Dataprovider.com dictionary: Economic Footprint is a proprietary metric that gives an estimate of the economic impact of the website ranging from 0 (low) to 100 (high). This score is calculated based on several key factors, each offering insights into different aspects of the website's commercial presence and activity. It takes into account the overall size of the website such as the number of subdomains or forwarding domains. Economic activity is further approximated by checking for a variety of e-commerce features.

50<Economic footprint<59	19.1%	**47.75%**
International shipping	16.6%	**41.50%**
40<Economic footprint<49	16.4%	**41.00%**
Headquarters	12.8%	**32.00%**
Ecommerce max	11.6%	**28.00%**
30<Economic footprint<39	10.3%	**25.75%**
Online store	9.8%	**24.50%**
WebType= E.commerce	9.8%	**24.50%**
Ecommerce high	9.1%	**22.75%**
Ecommerce Trust grade A	8.2%	**20.50%**
All sTLD	8.1%	**20.25%**
EU ccTLD	6.8%	**17.00%**
Geo. gTLD	5.8%	**14.50%**
All ccTLD	5.6%	**14.00%**
WebType=Business	5.5%	**13.75%**
20<Economic footprint<29	6.2%	**12.50%**
All sites	**4.9%**	**12.25%**
All continents (see below)	**4.7%**	**11.75%**
.com	4.3%	**10.75%**
All gTLD	4.2%	**10.50%**
.org	3.7%	**9.25%**
.net	3.5%	**8.75%**
10<Economic footprint<19	3.5%	**8.75%**
All new gTLD	3.1%	**7.75%**
WebType=Content	1.7%	**4.25%**
Economic footprint <11	0.9%	**2.25%**
WebType=Blog	0.8%	**2.00%**
WebType=Forum	0.3%	**0.75%**

This first table shows a strong positive correlation for multilingualism with :

- High economic footprint sites
- Ecommerce sites
- Company's headquarters site
- Sponsored Top Level Domains[4]
- European Union ccTLD

A positive correlation for :

- Geo gTLD and all ccTLD

[4] https://en.wikipedia.org/wiki/Sponsored_top-level_domain

- Web of Business type.

On the other hand, it shows a strong negative correlation with multilingualism with:

- Webs of type Forum or Blog
- Very low economic footprint sites
- New gTLDs

And a relatively weak level of multilingualism in average for .net, .org, gTLDs and .com.

The following table reports the checking of correlation of multilingualism with the main language of the web sites.

Table 6: Variability of Multi/ depending on main language of the site

SAMPLE	%Hreflang	Multi%
Ukrainian	23.6%	59.0%
Catalan	22.2%	55.5%
Estonian	21.6%	54.0%
Greek	20.3%	50.8%
Italian	10.9%	27.3%
Hindi	10.0%	25.0%
Dutch	9.6%	24.0%
Finnish	9.6%	24.0%
French	8.6%	21.5%
Hebrew	8.1%	20.3%
Spanish	7.6%	19.0%
Turkish	7.6%	19.0%
Vietnamese	7.5%	18.8%
German	6.8%	17.0%
Persian	6.8%	17.0%
Polish	6.5%	16.3%
Urdu	6.2%	15.5%
Swahili	6.0%	15.0%
Thai	5.4%	13.5%
Swedish	5.1%	12.8%
Telugu	5.0%	12.5%
AVERAGE	**4.8%**	**12.0%**
Afrikaans	4.5%	11.3%
Portuguese	4.2%	10.5%
English	3.8%	9.5%
Malay/Indonesian	3.4%	8.5%
Russian	3.0%	7.5%

Marathi	1.8%	4.5%
Japanese	1.7%	4.3%
Korean	1.5%	3.8%
Chinese	1.0%	2.5%

As for the main language of the sites, more than 50% of sites in Ukrainian, Catalan, Estonian and Greek are multilingual. In the other side of the table, Chinese, Korean or Japanese sites are multilingual in less than 5% of the case[5] and Portuguese, at difference from Italian, French, Spanish and other European languages is below average.

Table 7: Continent analysis

CONTINENT	Total Count	%	HREF Count	HrefLang Tag occurrences	Href%
North America	81 700 171	42,85	30 585 750	775 311	2,53%
Europe	48 926 074	31,19	29 284 084	2 294 511	7,84%
Asia	24 338 209	19,35	16 402 470	613 182	3,74%
South America	3 956 843	2,74	3 103 861	119 150	3,84%
Oceania	3 154 042	1,95	1 884 996	35 682	1,89%
Africa	1 477 744	1,91	933 273	31 704	3,40%
Antarctica	2 418	0,00	588	15	2,55%
TOTAL	**163 555 501**	**100**	**82 195 022**	**3 869 555**	**4,71%**

European websites are top leader in multilingualism and North America has a long way to go[6].

If the assumption of 5 languages in average per multilingual site is correct, then the global MRate would be 1.6, higher than the Humanity equivalent rate, but not higher than 2 as we expected from manual measurements (Pimienta, 2024). However, the figure of the average number of languages per multilingual site is still difficult to approximate, the only reliable data being the one of the sampling of the Ionan University on EU ccTLD study with a value of 4.91 and a variance relatively low (see Nm column in the first table).

It is interesting to discover, with those tables, that even if those figures are still tentative as based on an assumption to be verified, the variance is high, depending on country, type of site or main language. This result confirms however two of OBDILCI's past assumptions:

- The rate of multilingualism of ecommerce sites is much higher than the average.
- As for the Tranco sample, it is not possible yet to have figures, but if the global MRate is globally 1.6 there will be no surprise that the figure for the Tranco sample get close to 2, as the top more visited sites are more likely to be multilingual than the rest of the web.

[5] Note however that there is a potential bias: it is possible that Asian web designers have a reluctance to use the hreflang tag which force them to specify the language using Latin alphabet.

[6] And Canada, with 5.5%, above average of 4.8% is not to be blamed! Clearly the country with today less multilingualism is USA with 1.9% hreflang tags corresponding to an estimated less than 5% multilingual sites.

ANALYSIS OF THE VARIABILITY OF MULTILINGUALISM OF THE WEB

The crosschecking realized by categories and consigned in the 3 previous tables provides insightful inputs on the multilingualism of the Web.

- ✓ The variance is extremely high, percentages go from 1% to 100% for Multi% depending on category and from 2.5% to 60% depending on the main language of the site. Yet it is probable than the average is today around 12% of web sites being multilingual with an expected trend to grow thanks to the pervasiveness and boost of multilingual tools based on IA.
- ✓ Europe is much more multilingual than the other continents, the lowest being North America, with USA as one of the lowest places for multilingualism, although Australia and Japan show still lower data (respectively 1.8% and 1.5%).
- ✓ There is a clear and strong correlation between ecommerce activities and high level of multilingualism.
- ✓ Some languages are more propense to multilingualism (Ukrainian, Catalan, Estonian, Greek) while some other have yet a long way to go (Chinese, Korean, Japanese, Russian).
- ✓ The leaders in the digital economy are highly multilingual, the correlation between economic leadership and multilingualism is spectacular with the parameter "economic footprint".

IV REFERENCES

Giannakoulopoulos, A., Pergantis, M., Konstantinou, N., Lamprogeorgos, A., Limniati, L., and Varlamis, I. (2020). Exploring the dominance of the English language on the websites of EU countries. Fut. Int. 12, 76.
https://doi.org/10.3390/fi12040076

Pimienta, D. (2024). Is it true that more than half the Web contents are in English? If Web multilingualism is paid due attention then no! Forum for Linguistic studies, Vol. 6 , Iss. 5
https://doi.org/10.30564/fls.v6i5.7144