BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Is It True That More Than Half of Web Contents Are in English? Not If Multilingualism Is Paid Due Attention!

Daniel Pimienta

Observatory of Linguistic and Cultural Diversity, 06000 Nice, France

ABSTRACT

The belief that English is and will remain largely dominant as the first language of the Internet in terms of content and is the natural lingua franca in cyberspace plays against the mobilization of human and funding resources to incorporate minority languages. We sustain that this belief stands on biased data and that multilingualism is more and more the nature of the Internet and translation its lingua franca. We challenge the validity of a source widely used, since 2011, to state that English represents a steady percentage of web contents over 50%. This business source, W3Techs, is well-famed and considered reliable for its surveys on web technologies, exploring a large sample of the Web. However, languages differ from other web technologies, in the fact than more than one language could be used on a website. Not taking into account the multilingual nature of the Web is a serious bias that leads to major errors. The study of the rate of multilingualism of the sample of websites used by W3Techs concludes that the percentage of English contents on the Web is within a 20%–30% range, a value coherent with the results from three referenced alternative methods. We plan for 2025 to create a tool for measurement of languages and rate of multilingualism in a series of websites, with thorough attention to list all the languages used within a website, a complex matter. This tool will be applied to the same sampling and should close definitively this matter.

Keywords: Biases; Languages in the Web; Multilingualism; English; Lingua Franca

*CORRESPONDING AUTHOR:

Daniel Pimienta, Observatory of Linguistic and Cultural Diversity, 06000 Nice, France; Email: pimienta@funredes.org

ARTICLE INFO

Received: 24 August 2024 | Revised: 4 September 2024 | Accepted: 14 September 2024 | Published Online: 6 November 2024
DOI: https://doi.org/10.30564/fls.v6i5.7144

CITATION

Pimienta, D., 2024. Is It True That More Than Half of Web Contents Are in English? Not If Multilingualism Is Paid Due Attention! Forum for Linguistic Studies. 6(5): 201–212. DOI: https://doi.org/10.30564/fls.v6i5.7144

1. Introduction

The unbearable heaviness of being English contents on the Internet is taken as a solid fact by most media and many researchers, including those working in language technology. This situation insidiously tears apart policies and technical decisions around the concept of English as the *lingua franca* of cyberspace, that is to say the unique language which is systematically used to make communication possible between people who do not share the same language.

There is, however, a contradiction on the will to connect everybody to the Internet (today, 2/3 of the world population is already connected according to ITU) and the reality of the limited percentage of people in the world being either first speakers (noted L1) or second speakers (noted L2) of English, with L1 + L2 meaning the set of speakers using English either as a first or second language. Different sources offer values around 1.5 billion for L1 + L2 and no source crosses the line of 2 billion. This implies that less than 19% of world population understands English.

"Yes but, more than 50% of web contents are in English!" This paper will show that this common claim, supported by an often-cited source, is wrong. In reality, the correct figure is between 20% and 30%, and this paper will correct the bias of this source and point to alternative sources to make the demonstration.

Why is that important? Why is that a concern for researchers dedicated to incorporating less-resourced, under-resourced, endangered, minority, and minoritized languages?

Certainly not for some type of hard feeling or politically minded struggle against the English language! It is important because that belief has long played and still plays as a demotivator for localization and content creation in other languages, especially those that are minority or less-resourced.

Is that figure a concern only for researchers on minority languages? Not so! It is also a concern for business, as e-commerce share of total commerce crossed the 20% threshold in 2020 and it keeps growing[1]. Many sources[2] converge in the diagnosis that it is essential for e-commerce applications to speak the mother tongue of their customers. Therefore, getting wrong figures about the reality of language spread on the Internet could turn into bad business.

Finally, this overestimated figure acts as a screen hiding the reality of *multilingualism* in the Internet, already the realm with the utmost linguistic diversity and it is just the beginning. Every day, Internauts make more use of the complete variety of their languages (L1 and L2), either in communication or information retrieval, and this use is exactly the definition of multilingualism. The approximate number of languages existing in the digital sphere (they are said to be "localized") has grown from less than 100 before 2000, to 500 before 2020, and reaches 750 today[3]. This growth will not stop, although the target of some 8000 existing languages indicates it is a long way to go.

In February 2022, Statista[4] claimed that "English is the universal language of the Internet"[5]. Statista supported that statement with data from W3Techs'[6] measures on web contents claiming, on the same date, that 63.7% of websites were in English[7].

The browsing of yearly historical data from W3Techs[8] shows figures always above 50% since 2011 and none of the other languages ever reaching 10%. The analysis from the share of languages on the Web based on those figures is countless and it is probable that many public policies have been designed on the same grounds.

Between 1998 and 2007, other alternatives have existed about the language of web contents [1]. However, since 2011 and until 2017, W3Techs has been the unique source for such data and has logically become a universal reference for linguistic data about web contents, promoted not only by Statista but also by Wikipedia[9]. The fame of that source reaches out the research community, despite the fact that its methodology has not been published and much less peer

[1] https://www.emarketer.com/content/worldwide-ecommerce-sales-break-6-trillion

[2] Such as https://motsdici.be/wp-content/uploads/2019/04/Article-cant-read-wont-buy.pdf or https://www.t-works.eu/en/e-commerce-growth-and-the-importance-of-language/

[3] Figures derived from the work of https://unicode.org

[4] https://statista.com, a company specialized in providing statistics.

[5] https://www.statista.com/chart/26884/languages-on-the-internet/

[6] https://w3techs.com, a company specialized in surveys about web technologies

[7] https://w3techs.com/technologies/overview/content_language

[8] https://w3techs.com/technologies/history_overview/content_language/ms/y

[9] https://en.wikipedia.org/wiki/Languages_used_on_the_Internet

reviewed.

A scrutiny of the methodology used by W3Techs[10] reveals a lack of consideration of the multilingual nature of the Web. A legitimate question arises then: what would be its produced figures if the multilingual property of many websites was receiving due attention by the method driving its algorithm?

This paper addresses that point by analysing, on one hand, the method used by W3Techs and the bias resulting from the assumption that all websites are monolingual. On the other hand, by computing the "rate of multilingualism" of the sample of the Web used by W3Techs to produce its figures. The rate of multilingualism of a set of websites is defined by the ratio of the sum of all linguistic versions of the websites over the total number of websites. Based on that analysis, an attempt is made to un-bias, with a simple equation, the value of the percentage of English contents produced by W3Techs.

The result of this attempt to un-bias the W3Techs figure for English leads to the conclusion that the corrected percentage of English contents on the Web could be in the range between 20% and 30%, instead of over 50%. This is the same range provided by another source, the Observatory of Linguistic and Cultural Diversity on the Internet (OBDILCI[11]). OBDILCI produces indicators of languages on the Internet, since 2017, using a different approach[12]. This 20%–30% window range is also the one deduced from a recent study focusing on the websites of European Union national domains[2]. A third and last source for data on the presence of languages on the web[13] offers figures extremely close to OBDILCI's, with English at 26.3%. It seems (and needs to be confirmed) that the method of Netsweeper is to measure webpages instead of websites, which implies full consideration of multilingualism. The size of the sampling of Netsweeper is also way superior to W3Techs's with an announced 12 billion pages (versus the 1 million websites of the sampling used by W3Techs).

Although quantitative studies about the multilingual nature of the Web are scarce and none so far has addressed specifically the rate of multilingualism of the Web, qualitative studies on the subject exist. As early as 2007,[3] addressed the "multilingual Internet". Studies focusing a particular segment of the Web have followed, such as[4], for university websites, or[2], for European Union top level domain websites. The transversal idea across those studies is consistent in that the Internet is, has been, and will be still more, increasingly multilingual. One author, even argued, in 2019, that the web has passed the multilingualism step and evolved into *hyperlingualism*[5]. More recently, reference[6] have advocated for a better understanding of the multilingual use on the Internet as it has implications for issues in applied linguistics such as the study of heteroglossia, language learning, language education, and language policy.

The present article pretends to assess the size of the W3Techs bias, made by ignoring the fact that the Web is by nature, and every day more, a multilingual realm. As a matter of fact, other biases exist which inflate the English percentage and have been analysed[14].

2. Materials and Methods

2.1. First and Second Languages (L1, L2)

The total number of L1 speakers is generally computed as the world population, assuming that all humans have only one mother tongue and attributing it to babies as soon as they are born. Following the Ethnologue global data set #26 of March 2023[15], there are 7,404 million L1 speakers in the world, of which 380 million are L1 English speakers, e.g., 5.13% of the world population. Ethnologue is generally considered the world's most comprehensive catalogue of languages[16]. However, demo-linguistic figures are extremely difficult to gather, and absolutely no source is considered exempt from errors in that field.

A proportion of humans speaks more than one language. The majority of persons are monolingual, and their L2 lan-

[10] https://w3techs.com/technologies

[11] https://obdilci.org

[12] https://www.obdilci.org/projects/main/

[13] https://www.netsweeper.com/government/top-languages-commonly-used-interneto

[14] https://www.obdilci.org/projects/main/englishweb/

[15] https://www.ethnologue.com/Ethnologue-26-Global-Dataset-Doc.pdf

[16] https://en.wikipedia.org/wiki/Ethnologue

[17] https://en.wikipedia.org/wiki/List_of_polyglots

guages count is null. However, many persons are bilingual, and others speak 3 or 4 languages. The number of persons speaking more than 4 languages is quite low[17]. The number of persons with more L2 languages decreases rapidly with the number of languages, due to obvious limitations, although the literature mentions that speakers of over 60 languages exist and over 200 languages have existed[18].

The total number of L1 + L2 speakers computed by Ethnologue in 2023 is 10,599 million, which implies that the rate of multilingualism of humanity would be 10,599/7,404 = 1.432. Hence, around 40% of the world population speaks more than one language. It should remain clear that the figure of 10,599 million includes the same person as many times as this person speaks different languages. While there are some differences between sources on L1 figures, the differences are much higher for L2 figures. On one hand, the definition of the level of control of languages to be accounted as L2 is not precise. On the other hand, computations are cumbersome because L2 speakers are spread across a larger number of countries. Among the total L2 speakers, 1,078 million are English speakers according to the Ethnologue dataset #26, then the world total of L1 + L2 English speakers would be 380 + 1,078 = 1,458 million. Other sources propose the figures of 1,180 million[19] or 1 500 million[20] and Crystal expressed the possibility of that figure tending in the long term to 2,000 million[7].

It is common to see the percentage of L1 + L2 speakers in one language computed by dividing the number of L1 + L2 speakers by the world population. Doing so would require a warning that the total of percentages for all languages will then exceed 100%. When the total is forced at 100%, this provokes an error hidden in the figures for the rest of the languages. To maintain a real percentage, the division should be made over the total number of L1 + L2 speakers. With that rule, the world percentage of English L1 + L2 speakers would be, based on the Ethnologue source: 1,456/10,599 = 13.74%. Even using the boundary of 2 billion English speakers set by ([7], the total percentage of people understanding English as a first or second language could not cross the 20% line. In other words, a little more than 4 persons out of 5 do not understand English.

If the trend is that the Internet will be accessed by almost every human, then this figure of 20% should represent a hard frontier. English cannot be the lingua franca of the Internet if it bridges the communication for less than 20% of the internauts. It was so on the first stages of the Internet, when a high percentage of internauts, academics, researchers and business persons, have English as L1 or L2. Today the bridge is made by translation, strongly assisted by applications, and tomorrow artificial intelligence will do the bridging.

2.2. Computing Languages on the Web

In theory, computing the repartition of languages on the Web is based on web pages, not websites. The formal definition of the percentage of presence of a language on the Web is: total number of web pages in this language divided by the total number of web pages. Some web pages are multilingual, and the sum of percentages for all languages will then be higher than 100%.

According to Netcraft[21], there are today over 1.2 billion websites, of which 200 million are active. One source[22] evaluates the total number of web pages around 50 billion, of which less than 10% would be indexed by search engines.

Computing the presence of languages on the Web by counting websites instead of web pages is therefore an understandable simplification, but caution is required. This method implies a process quite similar to the one previously described for humans. Websites have an L1 language and some have L2 languages. Making an analogy with the percentages of L1 + L2 speakers per language, correct computations of the percentage of languages in a sampling of websites should be made over the total number of linguistic versions of the sample (not over the total of websites).

Experience shows that websites suffer fewer limitations than humans and can cross the boundary of 4 more easily. Wikipedia.org is available today in 331 languages, Facebook.com in 112, YouTube.com in 85 and Google.com in 87. As a matter of fact, many well-known websites exceed

[18] https://lawlinguists.com/fr/record-languages-spoken-one-person/

[19] https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

[20] https://www.statista.com/chart/26884/languages-on-the-internet/

[21] https://news.netcraft.com/archives/category/web-server-survey

[22] https://www.worldwidewebsize.com

[23] https://translate.google.com/?op=websites

100 languages while many other less-known sites, especially within the e-commerce realm, offer tens of linguistic versions. Google Translate[23] allows, since June 2024, to dynamically translate web pages into 244 languages, and it is used by some websites to open, widely and at no cost, their linguistic coverage.

2.3. The Rate of Multilingualism of the Web

The question about the rate of multilingualism of the whole Web is open and not easy to answer. Intuition and experience say that this indicator could be higher for the Web than for humanity (it would then be higher than 1.43) because the limitation factor is clearly much lower. Obviously, many humans with different languages can contribute contents to the same website. The incentives to reach out more widely, using different languages on the website exist, especially in e-commerce. So, the rate of multilingualism is a key indicator for the Web. Some targeted studies have been conducted at the European Union level [2] but no worldwide data exists as of today, and the automation of such analysis for the whole Web is not easy.

However, for the sake of the objective to assess W3Techs methodology, it is sufficient to focus, instead of the whole Web, on the same sampling that W3Techs is using for its measurement.

W3Techs has used for many years the list of 10 million most visited websites proposed by alexa.com (a commercial provider), but this service ended in May 2022. W3Techs has then switched to the use of the list of one million most visited websites proposed by Tranco[24], a non-profit organization. Tranco presents itself as "a Research-Oriented Top Sites Ranking Hardened Against Manipulation". The switch from Alexa to Tranco could, by the way, explain why W3Techs results for 2023 are sensibly different from 2022, with a notable decrease of English and a strong growth of many of the close followers of English in the ranking…

Having in mind the W3Techs assessment, we have therefore developed a non-automatic approximation of the Tranco data.

We have analysed manually (browsing and looking for the website's language options) 7 series of 100 websites extracted from the Tranco list and computed the results. Note that the Tranco list is sorted from the most visited to the least.

The counting was made by browsing each website, one by one, searching for linguistic options and taking note in the same Excel file, having one website per line, in order to make a final count. During this process, the following situations were identified:

- In the majority of the cases, the linguistic options are at the top of the homepage and the number of linguistic versions is easily counted.
- In some cases, they appear at the bottom, with the same facility to count.
- In less frequent cases, the linguistic options are implicit as country options and we have verified, for each country, if each presented language was effectively translated before making the count. Also, we have avoided counting twice if the same language is used in different countries.
- In some cases, the scope of languages allowed by the websites is explicitly exhibited after specifying the country.
- Some very large and famous websites (like Facebook and Google) prefer to deduce automatically from the user's device the language to be used. Changing this is not an obvious matter as it implies searching for the configuration page. In those cases we have searched the configuration page and counted the number of languages offered.

The computations were made by creating an Excel file for each sampling with the websites in rows and the parameters counted in columns. Seven files were defined and filled, corresponding to:

- The set of websites in the first 100 positions of the Tranco sample
- The set of websites in the last 100 positions of the Tranco sample
- Five series of 100 websites obtained using the random function of Excel applied to different ranges of the Tranco sampling as mentioned in **Table 1**.

The results are summarized in **Table 1** below, where:

- M.Rate: is the rate of multilingualism, e.g., the total number of linguistic versions discovered over the total number of websites analyzed;
- Invalid: is the percentage of websites found invalid during the process (different situations are discussed

[24] https://tranco-list.eu

hereafter)

- Mono: is the percentage of websites with a unique language
- Bi: is the percentage of websites with 2 linguistic versions
- Tri: is the percentage of websites with 3 linguistic versions
- Multi: is the percentage of websites with more than 3 linguistic versions
- M.Avg: is the average number of linguistic versions of the Multi websites.

We have always taken a conservative approach, preferring to let the count of M.Rate be 1 or not selecting a higher figure, in case of doubt. In some cases, the linguistic versions are automatically generated using the Google Translate dynamic page. When that happens, in most cases, all the 132 linguistic potential options are proposed. This approach strongly affects positively the M.Rate average and could be considered legitimate in terms of multilingualism. However, in keeping with the conservative approach, we have still counted those websites as monolingual.

Table 1. Results of random websites analysis.

	FIRST 100	1000–10000	100000–1000000	1–1000000	1–1000000	1–1000000	LAST 100	MEAN
M.Rate	44.10	3.07	2.09	1.94	2.31	2.97	1.81	2.23
Invalid	23.0%	19.0%	25.0%	33.0%	25.0%	26.0%	25.0%	27%
Mono	18.2%	67.9%	72.0%	82.1%	70.7%	78.4%	84.0%	77%
Bi	6.5%	11.1%	17.3%	7.5%	16.0%	5.4%	5.3%	10%
Tri	2.6%	2.5%	2.7%	3.0%	4.0%	4.1%	4.0%	4%
Multi	74.0%	19.8%	9.3%	9.0%	10.7%	13.5%	8.0%	10%
M.AVG	59	11	10	10	11	15	9	11

The analysis shows that the first two positions, for the most visited websites of the sample, have an extremely high rate of multilingualism: over 44 for the first hundred, and over 3 for the random sample between 1,000 and 10,000. In keeping with the conservative approach, we discarded the first columns for the computation of the mean and kept only the 6 columns on the right. We did include the measurement of the last 100 less visited sites of the ranking, where logically the M.Rate is the lowest.

These measurements are obviously not sufficient to apply statistical laws on the distribution of what would be considered random variables (the 7 elements measured and shown in the first column). They only represent a first level of approximation of the data, to be taken with caution and within a large confidence interval.

The most stable result appears to be the high number of invalid websites, around 25%, witnessing that websites are born, can get sick for some period of time, and eventually die.

The average rate of multilingualism of that websites sample appears to be higher than 2, which makes it higher than the Human rate, which is not a surprise. Based on the results in **Table 1**, the typical repartition of valid websites seems close to:

- 75% of websites are monolingual
- 10% of websites are bilingual
- 4% of websites are tri-lingual
- 10% of websites have more than 3 languages, with an average around 11.

These results, although very approximative, will help anyway to compute the bias of the W3Techs figure for English since W3Techs results are based on the same Tranco sample. The question about what happens for the many websites which are outside the list of one million most visited remains open and is related to the bias of working with the more visited websites.

It is meaningful to compare these results with those of[2], which are based on a larger sample of websites (over 100,000). Those websites are restricted to European Union ccTLD[25] (such as .fr or .uk) and not correlated to the highest number of visits. The authors have allowed public access to the data resulting from their collection[26]. There is a file for each country, showing all the websites analysed and the different counts. Another file summarizes all the collected

[25] Country Code Top Level Domain

[26] https://zenodo.org/record/3698008

data by country and the totals.

It is possible to compute in each country file the average number of linguistic versions for websites with over 2 languages and report the value in the summary file. Computations of the rate of multilingualism as well as the percentage of English websites are then made. The results are presented in **Tables 2** and **3**, where the grey data are the result of our computations while the rest of the data is taken directly from the source[2].

- Population: number of residents in the country
- Websites: number of websites analyzed in the sample
- % Mono: percentage of websites found to be monolingual
- % Bi: percentage of websites found to be bilingual
- % Multi: percentage of websites with more than 2 linguistic versions
- M.Avg: average number of versions of sites with more than 2 linguistic versions
- MRate: Rate of multilingualism
- Domain count: number of domains below the ccTLD[27] for all countries, except Latvia which was missing and was completed by another source[28].

Table 2. Repartition of EU websites by type of linguistic versions analyzed by[2].

Country	Population	Websites	% Mono	% Bi	% Multi	M.Avg	MRate	Domains
Austria	8,999,973	4,063	87.96	11.35	0.69	3.50	1.131	1,289,274
Belgium	11,579,502	4,063	81.20	15.92	2.88	3.34	1.227	532,005
Bulgaria	6,964,301	3,178	69.26	28.51	2.23	3.70	1.346	228,272
Croatia	4,112,131	3,489	69.30	27.66	3.04	3.85	1.363	137,901
Cyprus	1,186,194	828	66.91	30.80	2.29	5.11	1.402	220,947
Czechia	10,705,712	4,084	86.46	12.34	1.20	3.76	1.156	1,510,721
Denmark	5,785,741	4,067	84.98	14.36	0.66	3.19	1.158	1,472,212
Estonia	1,327,561	3,556	67.32	23.06	9.62	3.50	1.471	215,250
Finland	5,538,872	3,992	81.96	15.68	2.35	3.70	1.220	742,867
France	65,227,357	4,125	90.86	8.34	0.80	6.42	1.127	7,447,877
Germany	83,792,987	4,150	90.67	9.08	0.24	4.10	1.098	6,604,705
Greece	10,439,436	3,953	68.13	29.65	2.23	3.70	1.357	174,018
Hungary	9,668,737	3,993	84.22	14.70	1.08	3.30	1.172	289,796
Ireland	4,927,661	3,825	98.72	1.23	0.05	3.50	1.014	265,024
Italy	60,496,082	4,123	82.61	15.18	2.21	4.10	1.220	3,817,443
Latvia	1,891,687	3,406	59.54	30.65	9.81	3.25	1.527	136,718
Lithuania	1,954,244	3,773	73.63	20.94	5.43	3.38	1.339	40,430
Luxemburg	623,897	2,876	72.95	21.73	5.32	3.72	1.362	192,571
Malta	441,161	444	95.72	3.15	1.13	6.60	1.095	66,264
Netherlands	17,123,478	4,133	87.10	12.10	0.80	3.39	1.140	4,340,730
Poland	37,864,109	4,110	89.44	9.68	0.88	3.64	1.120	573,641
Portugal	10,205,235	4,084	74.34	16.67	2.25	3.80	1.163	299,126
Romania	19,266,079	3,975	70.36	28.48	1.16	3.57	1.314	330,703
Slovakia	5,459,814	3,943	83.11	15.32	1.57	3.34	1.190	444,701
Slovenia	2,079,226	3,619	76.02	21.08	2.90	3.60	1.286	49,558
Spain	46,767,543	4,088	86.89	10.96	2.15	3.44	1.162	2,172,046
Sweden	10,081,948	4,084	85.70	13.96	0.34	3.00	1.146	961,089
UK	67,803,450	4,125	99.59	0.27	0.15	7.83	1.013	7,148,183

Note: The study was made before the Brexit.

From those two tables the average results for all European Union nations are computed and summarized in **Table 4**. Note that there are different ways to compute the average: simple average, weighted by country population, weighted by number of sites explored, and the last one, probably the most appropriate to get a global figure for the European Union,

[27] https://domainnamestat.com/statistics/country/others

[28] https://www.nic.lv/en/look-back-at-the-lv-in-2021

Table 3. Repartition of EU websites in English analyzed by [2].

Country	Websites	Mono	Bi	Multi	Mono Eng.	Bi Eng.	Multi Eng.	Total English	Total Versions	% English
Austria	4,063	3,574	461	28	68	454	28	550	4,594	12.0%
Belgium	4,063	3,299	647	117	365	553	112	1,030	4,984	20.7%
Bulgaria	3,178	2,201	906	71	139	891	71	1,101	4,276	25.7%
Croatia	3,489	2,418	965	106	143	940	104	1,187	4,756	25.0%
Cyprus	828	554	255	19	421	255	19	695	1,161	59.9%
Czechia	4,084	3,531	504	49	69	474	46	589	4,723	12.5%
Denmark	4,067	3,456	584	27	218	577	26	821	4,710	17.4%
Estonia	3,556	2,394	820	342	158	656	330	1,144	5,231	21.9%
Finland	3,992	3,272	626	94	157	578	94	829	4,872	17.0%
France	4,125	3,748	344	33	46	335	31	412	4,648	8.9%
Germany	4,150	3,763	377	10	62	370	10	442	4,547	9.7%
Greece	3,953	2,693	1,172	88	369	1,165	86	1,620	5,301	30.6%
Hungary	3,993	3,363	587	43	55	564	42	661	4,679	14.1%
Ireland	3,825	3,776	47	2	3,763	47	2	3,812	3,876	98.3%
Italy	4,123	3,406	626	91	88	595	89	772	5,031	15.3%
Latvia	3,406	2,028	1,044	334	188	817	331	1,336	5,202	25.7%
Lithuania	3,773	2,778	790	205	121	747	196	1,064	5,051	21.1%
Luxemburg	2,876	2,098	625	153	375	497	144	1,016	3,917	25.9%
Malta	444	425	14	5	416	14	5	435	486	89.5%
Netherlands	4,133	3,600	500	33	143	493	33	669	4,712	14.2%
Poland	4,110	3,676	398	36	47	387	34	468	4,603	10.2%
Portugal	4,084	3,036	681	92	136	646	89	871	4,748	18.3%
Romania	3,975	2,797	1,132	46	311	1,104	45	1,460	5199	28.1%
Slovakia	3,943	3,277	604	62	70	544	59	673	4,671	14.4%
Slovenia	3,619	2,751	763	105	143	722	96	961	4,655	20.6%
Spain	4,088	3,552	448	88	91	392	81	564	4,751	11.9%
Sweden	4,084	3,500	570	14	230	540	12	782	4,682	16.7%
UK	4,125	4,108	11	6	4094	11	6	4,111	4,177	98.4%

Table 4. Results averaged for the whole European Union.

	% Mono	% Bi	% Multi	M. Avg	MRate	% English
Simple average	80.9	16.5	2.3	3.9	1.225	24.21%
Weighted avg. by country population	87.3	11.4	1.2	4.4	1.147	26.04%
Weighted avg. by number of sites explored	81.3	16.2	2.3	3.7	1.216	24.81%
Weighted avg. by number of domains	86.3	9.6	1.0	4.6	1.095	28.42%

weighted by the number of domains below the ccTLD.

The last line of the results in **Table 4** allows us to claim that, based on the sample established by [2], 86% of European Union ccTLD sites are monolingual, 10% are bilingual and 1% have more than 2 linguistic versions, with an average of 4.6. Following this approach, 28.4% of European Union linguistic versions of websites in 2020 were in English (this percentage is obtained by dividing the total number of linguistic versions of websites in English, being a monolingual English website or an English version of a multilingual website, divided by the total number of linguistic versions).

The rather large differences in the rate of multilingualism (MRate) between the four modes of averaging shown in **Table 4** could be an indicator that, outside the Tranco list of the one million most visited, the rate of multilingualism of the Web is not so high. It could be also, and much more probably, that the rate of multilingualism of ccTLD websites is logically lower than that of websites with global purposes (which are often in non-national domains like .com). It could also be expected that the presence of English linguistic versions will increase strongly in non ccTLD sites at the same time as the rate of multilingualism. One thing compensating for the other, the final percentage of English could remain in the 20–30% range.

As for the global rate of multilingualism of the Web, it is notable that ccTLD registrations represent together only 38% of the total domain registrations. However, often, albeit not in the European region, they are diverted from their original definition, associated with one country, towards worldwide business proposes. The best example is that of .tk, the Tokelau islands ccTLD, which is the top in terms of registrations, even more than .cn for China. Obviously, most .tk websites do not contain national information about Tokelau (see sources[29]). As a matter of fact, a look at the table of the ccTLD in the corresponding Wikipedia article[30] reveals that a large proportion are now used as "domain hacks"[31], a usage perverted from the original country specific definition. Domain hacks are generally linked to some e-commerce activity, much more prone than national websites to multilingualism.

In any case, the data from [2] do not alter the discussion about W3Techs biases which are related to a different sampling, the Tranco list. The considerations about the ccTLD of European Union, which show a relatively low rate of multilingualism, cannot neither be generalized to the whole Web, as ccTLDs by nature are less prone to multilingualism than generic domains.

2.4. The W3Techs Methodology

According to the W3Techs methodology[32], a language recognition algorithm is applied every day to the homepage of the websites on the Tranco list. This algorithm identifies a unique language for each website and procceds to count. This method implies that multilingual websites with an English version are probably counted only as English. Finally, the percentages per language are computed over the total of measured websites and are applied for the 40 first languages (the rest of the languages are mentioned as having less than 0.1% of the total).

Clearly, the method is considering the Web as a monolingual space, and this could represent a serious bias given the figures computed above for the rate of multilingualism of the Web. Is it possible to un-bias this figure, at least for the result of English?

3. Results

There are two elements to consider in order to un-bias the W3Techs outputs:

- The percentage of non-English websites computed erroneously as English. This can happen because there are a number of English words on the homepage alongside the other main language. It can also happen because it is an invalid website not detected by the algorithm as such and treated as English. This could occur because the message resulting from the invalidity is in English.
- The impact of a rate of multilingualism higher than one on the computation of the percentage.

The correcting equation is: P' = (P − Err)/MRate, where:

- P' is the un-biased percentage for English contents;
- P is the percentage output for English contents provided by W3Techs;
- Err is the percentage of websites erroneously computed as English;
- MRate is the rate of multilingualism of the sampling.

The manual analysis of the 700 websites has shown that some percentage (around 10%) are non-English websites with many English words on the homepage. From the 27% of websites found invalid, less than 50% were identifiable clearly as such (returned codes 404, not found, or 403, forbidden access). Most of the time, the invalid message came from the host server or from the domain name manager, and it is in English. In more than 10% of the cases, there is just a short sentence in English or a "*site under construction*" message. Not knowing the details of the algorithm makes it impossible to determine if these situations are treated as non-counting or counted as English. For the sake of the equation we will use a conservative figure between 5% and 15% (e.g. between ¼ and ½ of that 27% figure) as the percentage of websites counted erroneously as English.

The average value of MRate is 2.23 in our sampling (**Table 1**), setting a value between 1.5 and 2.5 seems a reasonable. **Table 5** offers then three scenarios for the MRate and Err values and use R = 56.1%, the 23/3/2023 value for English contents in the W3Techs site:

Table 5. W3Techs results un-biased.

	Low	**Medium**	**High**
MRate	1.5	2	2.5
Err	5%	10%	15%
R'	34.1%	23.1%	16.4%

[29] https://www.verisign.com/assets/domain-name-report-Q42019.pdf and https://www.verisign.com/en_US/domain-names/dnib/index.xhtml?section=additional-information

[30] https://en.wikipedia.org/wiki/Country_code_top-level_domain

[31] https://en.wikipedia.org/wiki/Domain_hack

[32] https://w3techs.com/technologies

It is quite interesting to note that the medium figure for R' is very close to the results from OBILCI[33], one of the field players in the period before 2009. OBDILCI offered a new method in 2017 and published its results in 2022 for 342 languages, giving English contents around 20%, together with Chinese contents, and making a claim quite different from Statista's: "The transition of the Internet between the domination of European languages, English in the lead, towards Asian languages and Arabic, Chinese in the lead, is well advanced and the winner is multilingualism, but African languages are slow to take their place."

At difference with W3Techs, the methodology used for OBDILCI results has been revealed in all details and peer reviewed [8] as well as the produced results [9]. The indicators produced by OBDILCI are now openly (cc-by-sa 4.0) available to the public in the form of a database accessed by ISO Code 639-3[34]. OBDILCI's biases do exist and are thoroughly discussed in [8]. Their extent explains why the OBDILCI results are presented with a large confidence interval of –20% to +20%.

As for the other 39 languages which received percentages from W3Techs, the lack of consideration of web multilingualism prevents correcting the W3Techs figures. They should be considered unreliable. As a matter of fact, the percentage for non-English languages presented by W3Techs indicates the percentage of websites of the mentioned language without an English version.

4. Discussion

The study published in [2] is a welcomed first academic interest, for a long time, on the subject of languages on the Web. This subject has been left for too long to very few commercial entities and non-profit organizations. This situation has limited academic discussions, on a matter of growing general interest, with implications much beyond linguistics, such as e-commerce, public policies, geopolitics and cybergeography [10].

The programmed algorithm from [2] pays due attention to multilingualism by checking the language of all the internal links of the analysed website. It could be reused, modified, and applied to the Tranco list of websites (instead of the European Union list of websites). This would provide even stronger evidence for our conclusion that if the multilingual nature of the Web were taken into consideration, the W3Techs outputs for English would slide from the 55–65% range into the 20–30% range.

Some attempts have been made to partner in that perspective but it seems that this Ionian university department of audio and visual arts has logically other priorities. In that context, we have decided to plan for the realisation of a project for the direct measurement of languages in a sampling of websites, using language detection. In order to avoid the heavy investment associated with the crawling of millions of sites, we plan to apply a traditional statistical approach by randomly selecting, say, a hundred sets of 1,000 websites and deriving output from the statistical distribution of results.

We are in a preparation phase, testing various language detection algorithms, thinking about the best way to manage the challenge of multilingualism detection and looking for funding. As part of this preparation, we have set up a new sampling of 5 sets of 100 randomly selected websites from Tranco which are used for testing. This provides another set of results that we leave in open access[35] and replaces the previous sets which have been lost in a computer crash. The new results, presented in **Table 6**, remain in the same range as the first set, with an M.Rate of 1.93 (was 2.23), and a measured percentage of English of 29% for that sampling.

The project, targeted for 2025, will be used to complete the tools available in OBDILCI. It will allow conducting studies specific to some subset of the Web (like for instance the ccTLD of Portuguese speaking countries) and obviously it will also be applied to the Tranco sample in order to have a definitive correction of the W3Techs results.

5. Conclusions

There is more and more evidence of the fast growing and already high level of multilingualism of the Web, especially in its e-commerce component. It is increasingly evident that the lingua franca of the Web is not English but translation, a process that is becoming more and more assisted by computation, with promising perspectives offered by the new generation of AI tools.

[33] https://obdilci.org

[34] https://obdilci.org/Base

[35] https://www.obdilci.org/wp-content/uploads/2024/08/TRANCO-Sampling.zip

Table 6. New set of measurement.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Mean
M.Rate	2.51	1.73	2.07	1.65	1.69	1.93
Invalid	31%	30%	28%	28%	28%	29%
Multi	28%	16%	25%	19%	18%	21%
M.AVG	5.47	5.10	4.53	3.36	3.85	4.46
English %	23%	30%	29%	30%	34%	29%

Researchers should apply caution when using commercial sources which are not supported by scientific publication and peer reviews, especially in areas at the crossroads between demo-linguistics and web figures, where confidence intervals are quite large.

As for the multilingual historically unique property of the Web, evidence is around the corner of the virtual street and will impose itself progressively in the coming years, with artificial intelligence becoming an obvious enabler and amplifier of a booming multilingualism in the digital realm.

The planned activity from OBDILCI to provide an alternative method for crawling a set of websites for language detection and counting, with due consideration to website multilingualism, could represent a breakthrough in the studies of languages on the Internet. Beyond enabling a large scope of possibilities for studying the language distribution in targeted segments of the Web, as a complement to its general method targeting the complete Web, it could provide a definitive answer to the controversial question about the reality of the dominance of English in the digital world.

Funding

This research was funded by Direction Générale pour le français et les langues de France du ministère de la culture de la France.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The Excel files of the last sampling of 5 set of 100 websites selected randomly from the Tranco list are available at: https://www.obdilci.org/wp-content/uploads/2024/08/TRANCO-Sampling.zip.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Callahan, E., Herring, S.C., 2012. Language choice on university websites: Longitudinal trends. International Journal of Communication. 6, 322–355. Available from: http://ijoc.org/ojs/index.php/ijoc/article/view/1451/703

[2] Crystal, D., 2008. Two thousand million? English Today. 24(1), 3–6. Available from: https://www.cambridge.org/core/journals/english-today/article/two-thousand-million/68BFD87E5C867F7C3C47FD0749C7D417

[3] Giannakoulopoulos, A., Pergantis, M., Konstantinou, N., et al., 2020. Exploring the dominance of the English language on the websites of EU countries. Future Internet. 12(4), 76. DOI: https://doi.org/10.3390/fi12040076

[4] Kelly-Holmes, H., 2019. Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism. Annual Review of Applied Linguistics. 39, 24–39. DOI: https://doi.org/10.1017/S0267190519000102

[5] Pimienta, D., Blanco, A., Müller de Oliveira, G., 2023. The method behind the unprecedented production of indicators of the presence of languages in the Internet, Frontiers in Research Metrics and Analytics. 8. DOI: https://doi.org/10.3389/frma.2023.1149347

[6] Pimienta, D., 2022. Resource: Indicators on the presence of languages in Internet. Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages; Marseille, France; 24–25 June 2022. pp. 83–91. Available from: https://aclanthology.org/2022.sigul-1.11/

[7] Pimienta, D., Müller de Oliveira, G., 2022. Cybergeography of languages: Part 1: Method, results and focus on English; Part 2: The demographic factor and the growth of Asian languages and Arabic. The Inter-

national Review of Information Ethics.

[8] Danet, B., Herring, S.C., 2007. The multilingual internet: Language, culture, and communication online. Oxford University Press: New York, NY, USA. Available from: https://academic.oup.com/book/32471

[9] Leppänen, S., Peuronen, S., 2020. Multilingualism and the Internet. Wiley Online Library. DOI: https://doi.org/10.1002/9781405198431.wbeal0805.pub2

[10] Pimienta D., Prado D., Blanco A., 2009. Twelve years of measuring linguistic diversity on the Internet: Balance and perspectives. UNESCO publications for the World Summit on the Information Society. CI-2009/WS/1. Available from: http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016