---------- Forwarded message ---------
From: **Daniel Pimienta** <pimienta@funredes.org>
Date: Sun, April, 3rd 2022 at 3:44
Subject: Is English the Internet's Universal Language?
To: <felix.richter@statista.com>
Cc: <tristan.gaudiaut@statista.com>, Gilvan Müller de Oliveira <gimioliz@gmail.com>, Álvaro Blanco <blancoro@gmail.com>, Alexandre Wolff <alexandre.wolff@francophonie.org>

Mr. Richter,

I am the Head of the Observatory of linguistic and cultural diversity on the Internet, a research group working on the space of languages on the internet **since 1990**; I am also an old customer of STATISTA and I do trust you are professionals of the field of statistics.

While you published few days ago data based on W3Techs statistics, under the headline *English Is the Internet's Universal Language,* at the same time, more or less, we published data based on our model outputs, under headline *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*.

While W3Techs measure English, Chinese and Hindi contents at respectively, 62.9%, 1.4% and 0.1% we offer the same data at respectively, 19.6%, 21.6% and 3.8%...

Obviously, one (at least :-)) of the two data sources is **extremely** wrong!

STATISTA, as a source aggregator and a competent body in statistics, should have the duty to pay some attention and evaluate if it is not broadcasting misinformation on any subject.

Linguapax Review, a serious scientific reference about languages, published, few days ago, in a special edition on *Language Technologies and Language Diversity*, a paper of my authorship under the title "*Cyber-geography of languages*" of which I copy an extract, in annex, focusing the explanation of the main reason of W3Techs huge biases.

During our studies, since their beginning, back in the 90s, we have seen that the ratio of Web contents % per language over L1+L2 Speakers connected % per language stayed consistently within a window 0.5 - 1.5.
Why so? Because there is some kind of economic law which links Internet offer (contents in one language) with Internet demand (speakers connected in that language). All studies and surveys around the theme shows that people connected to the Internet prefer clearly to use their mother tongue, especially for e.commerce (see for instance https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556), and, when available, do use their second languages. When a country starts growing its Internet population, the offer logically starts (Government's websites, company's websites, university's websites, associative websites and personal websites progressively create their web contents and from the new connected persons a percentage get involved in content creation feeding the content curve.

We measure that with some indicators: *virtual presence coefficient* (content over speakers) and *content productivity* (content over connected speakers). When this economic phenomenon reaches very strong in locality (you ask your pizza from the corner shop via Internet), like in Japan, we see Virtual Presence coefficients over 2 and content productivity of 1.3.

Obviously, there are countries which are more prone than others at producing contents and some languages, English first, but also French, Spanish, German, and others, in some countries, benefit of being first options for multilingual websites; however ratios over 10 or below 0.1 of content productivity indicator are just totally **implausible**. You can check W3Tech ratios in https://funredes.org/lc2022/V3.4.htm.

The case of **Hindi** is the most interesting and significant. India have, following Ethnologue, the second English speaking L1+L2 population of the world and, from ITU/World Bank data, it is computed as the second English connected speakers population (close to half US population : 12.4% compared to USA 30.1%), as you can check in https://funredes.org/lc2022/V3.2.htm.

So let's wonder if the 244 million Indian persons connected to the Internet prefer to use English than their native language (for instance because of lack of contents in their mother tongue). An in-depth study conducted by KPMG in 2017 shows exactly the contrary! The trend is extremely strong for the **use of Indian native languages in the Internet**: Indian languages defining Indian Internet (Citation : *Indian language Internet users are expected to account nearly 75% of India's Internet user base by 2021*).

As shown in the annex below the **lack of consideration of multilingualism** leads W3Techs, indeed en excellent and reliable provider of **stats for Web technologies**, to huge mistakes when processing ***web content's languages***, as *language* is some kind of Web technology quite different from *Java Script Libraries* or *Web servers*. The problem of using language recognition algorithms starts with focusing only home pages of websites (if you compute contents you need to compute webpages not websites and home pages in non English websites often include few English words misleading the algorithm... and if you do not pay attention to multilingualism you count sites as facebook.com as English only when they offer tenth of language's option in their interfaces).

Best,
Daniel Pimienta

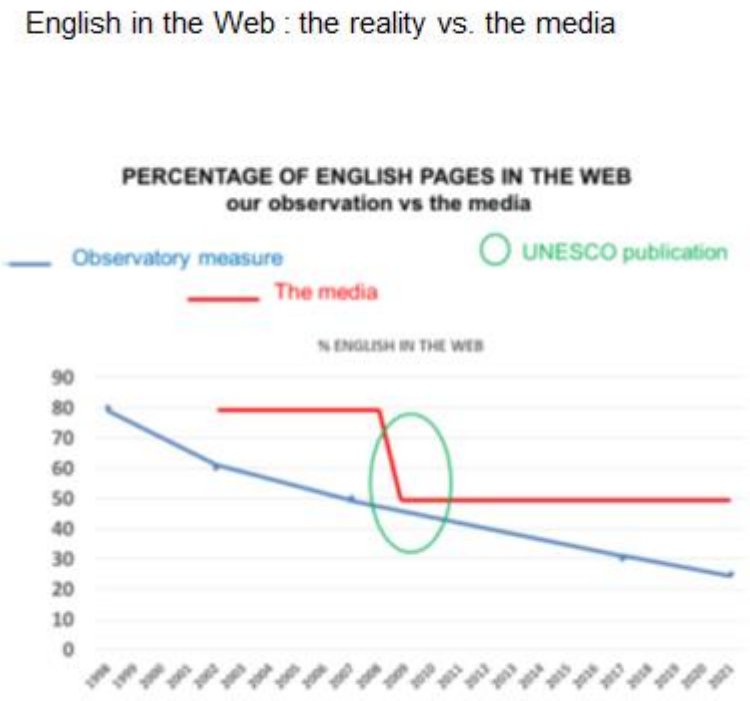
==========

**Extracted from Linguapax Review.**

Let us focus on the first languages of the Internet, starting with **English**. The most used, and for a long time the only source of data regarding the presence of languages on the Web, is W3Techs[1].

How is it possible that this source indicates a content percentage in English of 62.5% in 2021 when the Observatory suggests 30%? By measuring the websites directly with a language recognition algorithm, it should not make mistakes. *Well, that's multilingualism, stupid!* If we are allowed to rehash that famous and provocative expression... and address a historical

problem of misinformation about the space English occupies in the Internet, illustrated with these two curves:




Until 2009, the media published reports on the presence of English on the Web placing it at 80% with no changes during a decade, meanwhile our measurements indicated a progressive decline towards 50%. The media was supported by 3 publications which suggested, with the same methodology, the same results in 1997, 1999 and 2002. The methodology was not really biased but scientifically invalid[2], see Pimienta (2009) for more details. After UNESCO's publications on the matter (Pimienta 2006, 2009), the media[3] progressively adopted this new 50% value. Then W3Techs appeared as the sole source, whose results maintained a value between 50% and 60% since 2011[4].

How does W3Techs work, and which problem related to multilingualism occurs?

W3Techs selects the 10 million most visited sites on the Web, according to the Alexa digital marketing application (http://alexa.com). Let us brush aside the biases in favour of English of language recognition algorithms and the fact that selecting the 10 million most visited sites (out of the close to 1.2 billion existing websites[5], that is, less than 1%) favours websites in English. Today, analysing the entire Web seems an insurmountable task and we do not aim to disregard the commendable work of W3Techs, which provides many useful data. We focus particularly on how it manages multilingualism. W3Techs applies its algorithm to the **homepage** of these 10 million sites daily. The decision to restrict itself to the homepage, without somehow compensating for this, is part of the problem. Of course, the languages on the Web should be counted at page level and not site level, since a website limited to a single page cannot be counted in the same manner as another with thousands of pages. To this we

must add that this website with thousands of pages may also include pages in different languages, even though the homepage may be mainly in English, thus increasing the error threshold. Nowadays, a large part of the most visited sites (such as Facebook, for example) offer scores of linguistic versions from the homepage; counting the homepage in English is brushing aside all those versions. Finally, it is extremely common for the homepage of a site in a language other than English to have some words in English (for example, keywords or copyright); counting it as an English page, which is probably what happens with the W3Techs algorithm, means leaving out scores of pages in other languages.

One does not need to be a statistics expert to understand that the method, due to not considering the reality of multilingualism, can be hugely mistaken... What could the W3Techs algorithm do to improve its products, without abandoning a pragmatic approach, that is, without facing the challenge of analysing all the pages of all websites?

- ✔ Analyse the language options offered on the homepage and count each option as well as the English version.
- ✔ Find a method to obtain an approximate estimate of the number of website pages and multiply each linguistic version by that number to count pages instead of sites.
- ✔ When the algorithm reports more than one language on the homepage, do not count it as English.

Other factors that should alert us to the improbability of the W3Techs data and draw attention to some symptomatic statistical anomaly of a gross error:

- it makes no sense that the amount of content in English has remained stable for the last 10 years while Asian and Arab countries have invaded the Web during the same period and a set of non-European languages[6] now takes up close to a third;
- the presence of English-speaking Internet users (L1+L2) has gone from 32% in 2017 to 13% today;
- showing Chinese with just 1.3% of content and Hindi with 0.1% when both these languages represent 17.5% and 4.2% of connected people, respectively.

To close this chapter, the fact that the number of websites in English decreases in no way means that the presence of English in absolute terms diminishes, nor that it has stopped growing; it just means that new languages are taking up more and more space, which reduces the proportion of English. Of course, English continues to be the leading language in the Internet, whose estimated amount of content (30%) surpasses the number of Internet users (15%) by a factor of 2.

In Pimienta (2017), we have discussed the biases in different projects and how the lack of consideration for multilingualism can lead to blatant errors. The most typical frequent occurrence is calculating the elements based on L1+L2, divided by the world population, which causes magnitude errors, hidden within the values of the rest of the languages. The number of L1+L2 speakers is much higher than the world population, we had estimated the proportion of multilingual people in 2017 at 25%, in that updated version, Ethnologue offers us a more accurate figure of 43%.

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

[2] A language recognition algorithm was applied a single time to the homepage of 3000 randomly chosen websites, based on IP numbers, and the percentages were calculated. The statistical method to validate this approach would be to repeat the method many times and then analyse the random variable with statistical tools (average, variance, etc.). A single arrow shot at a target does not usually report on the archer's abilities!

[3] And also, unfortunately Wikipedia (https://en.wikipedia.org/wiki/Languages_used_on_the_Internet), from whom we expect more caution.

[4] See https://w3techs.com/technologies/history_overview/content_language/ms/y).

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

[6] Chinese, Hindi, Arabic, Turkish, Bengali, Vietnamese, Urdu, Persian and Marathi.

SOME REFERENCES


- Pimienta D.,, "Internet and linguistic diversity: the cyber-geography of languages with the largest number of speakers" in *LinguaPax Review 2021, Language Technologies and Language Diversity*, page 9. Also, in Spanish and Catalan. - https://www.linguapax.org/wp-content/uploads/2022/02/LinguapaxReview9-2021-low.pdf
- Pimienta D., "Is language a technology or a culture?" Imminent Annual Research Report 2021 - The Wealth of Languages, PP 69-70 -https://imminent.translated.com/imminent-annual-report-2021
- Pimienta D., "Indicators of Languages in the Internet ", in Proceedings of International Conference Language Technologies for All (LT4All), 4-6 December 2019, UNESCO, Paris; PP 315-319 - https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.79.pdf
- Pimienta D., Prado D., Blanco A. (2009) Twelve years of measuring linguistic diversity on the Internet: balance and perspectives, UNESCO, Publications for World Summit on the Information Society, CI-2009/WS/1-http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016
- Paolillo J., Pimienta D. and al.(2005) Linguistic Diversity in cyberspace: models for development and measurement", in Measuring Linguistic Diversity on the Internet, UNESCO, Publications for World Summit on the Information Society, 2005- http://unesdoc.unesco.org/images/0014/001421/142186e.pdf