Authors: Daniel Pimienta, Gilvan Müller de Oliveira

Cyber-geography of languages. Part 1: method, results and focus on English

Abstract:

The methodology, sources, biases and results of the recent study of the Observatory of Linguistic and Cultural Diversity on the Internet, for the creation of indicators of the presence on the Internet of the 330 languages with more than one million L1 speakers, are presented. It appears that the languages of Europe, and especially English, are still dominating the Internet but that the languages of Asia and the Arabic world are in a strong progression and will take the lead, first in terms of connected speakers (part 2 develops this point). The case of English is focused to demonstrate that its share of the Internet keeps declining to reach now 25%, despite some kind of mediatic disinformation placing it above 50% by trusting sources that do not pay due attention to multilingualism. The lingua franca of the Internet is translation, the indispensable crutch of multilingualism.

Keywords: Bias, Cultural Diversity, Disinformation, Linguistic Diversity, Multilingualism, Cyber-Geography of languages

Agenda:

Authors:

Daniel Pimienta

- Observatory of Linguistic & Cultural Diversity on the Internet, Email: pimienta@funredes.org, Website: funredes.org/lc

Gilvan Müller de Oliveira

- UNESCO Chair on Language Policies for Multilingualism, Federal University of Santa Catarina (UFSC), Brazil, Email: gimioliz@gmail.com, Website: https://www.unescochairlpm.org/uclpm/

ISSN 1614-1687

Introduction

The Observatory of linguistic and cultural diversity on the Internet[1] has just published, in August 2021, the results of its latest study[2], updating and improving its previous work from 2017, aimed at producing indicators of the presence, on the Internet, of the languages of more than 1 million L1 speakers[3]. This paper analyzes the data obtained in terms of the cyber-geography of languages.

First, an explanation regarding the selected languages: it is not an assumed political decision to focus on the languages with the greatest number of speakers and leave aside the others and in particular the languages classified as indigenous, in that period that UNESCO has marked, in 2019, as the year of indigenous languages[4], followed by the decade of indigenous languages 2022-2032[5]. The selection of the languages with the largest number of speakers is simply a restriction resulting from the methodology adopted whereas the analysis of biases leads to the conclusion that these would be too high for languages with a number of speakers of less than one million.

This takes us, before discussing the results, to briefly present the methodology, the main sources and the biases that can affect the data produced for the languages considered.

Methodology, sources, biases

The demo-linguistic source for this new edition is Ethnologue's Global Dataset #24[6], from March 2021, undoubtedly the most complete source in terms of languages, as well as the most up-to-date and reliable, although it should be clear that perfection does not exist in that field and some professionals in the field may object to some figures. We have also adopted the regrouping of some languages into macro-languages. The first edition focused on the 130 languages with the number of L1 speakers higher than 5 million, which list can be consulted in (Pimienta, 2021). The last edition includes the 330 languages with L1>1M[7], a figure closer to the estimated number of 500 languages present on the Internet.

The detailed methodology, sources, and associated biases, as well as the results, are fully documented in (Pimienta 2017, 2021). It is an indirect approach to measuring indicators of the presence of languages on the Internet, based on a large number of data sources about languages or countries on the Internet. The data by country have been transformed into data by languages by weighting with demo-linguistic data, this being one of the major originalities of the method[8].

The presence of languages is measured with statistical computations made from primary sources, in terms of global percentage over the L1 + L2 population, according to 6 indicators:

---

[1] https://funredes.org/lc

[2] See https://funedes.org/lc2021. This study has especially focused on Portuguese and has been possible thanks to the support of the Department of Culture and Education of the Ministry of Foreign Affairs of Brazil within the framework of the International Institute of Portuguese Language and under the coordination of the UNESCO Chair on Language Policies for Multilingualism, based at the Federal University of Santa Catarina (UFSC), Brazil.

[3] We use the terminology L1 to refer to the mother tongue and L2 for the second language (s).

[4] https://es.iyil2019.org/

[5] https://en.unesco.org/news/unesco-launches-global-task-force-making-decade-action-indigenous-languages

[6] https://www.ethnologue.com/product/ethnologue-global-dataset-0

[7] The new results and list of languages is accessible in https://funredes.org/lc2021/Results1M.xlsx.

[8] Credits to Daniel Prado who originated this concept in 2012.

For each considered language,

Internauts: percentage of connected speakers
Traffic: percentage of traffic
Usages: percentage of participation in platforms or connectivity resources
Contents: percentage of contents
Interfaces: the presence of the language in application interfaces or as an online translation language
Indexes: transformation, in terms of languages, of the countries ranking in various classifications of parameters related to the information society.

From these 6 indicators, 4 macro-indicators are computed:

Power: the average of the 6 indicators (absolute weight of the language on the Internet, which obviously favors languages with more speakers)

Capacity: the power divided by the number of speakers (a relative weight that makes it possible to measure the strength of languages regardless of their number of speakers)

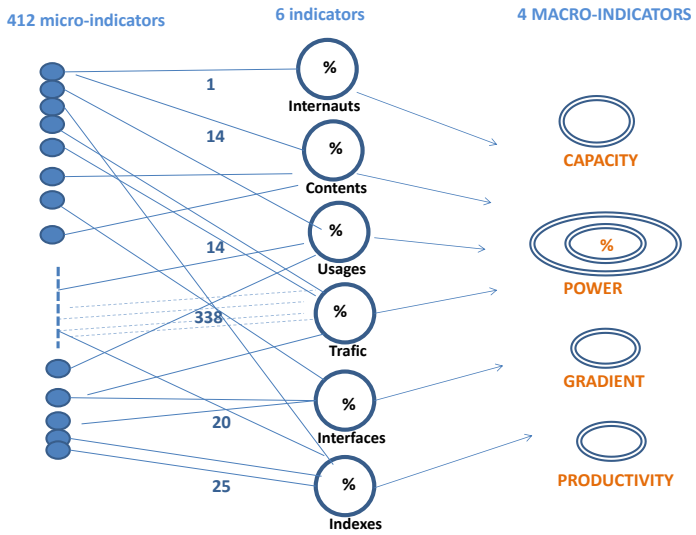
Gradient: the power divided by the number of connected speakers (measures the dynamism of the connected speakers at the time of producing content, generating traffic, subscribing to platforms ...)

Content productivity: Indicator of contents divided by the number of speakers.

All data are processed as percentages of the world number of L1+L2 speakers, a value higher than the world population[9]. According to the last version of Ethnologue:

- ✓ World Population (world total L1 Speakers): 7 231 699 136
- ✓ World Total L1 + L2 Speakers: 10 361 716 756
- ✓ The "world rate of multilingualism" is therefore 10 361 716 756 / 7 231 699 136 = 1.4328
  (In other words, 43% of the world's population is at least bilingual).

Indicator's scheme



[9] Of course, in that count, the same person counts once for their mother tongue and as many times he or she has a second language.

No statistical method is free from biases and it is very important to identify and analyze biases and their effects on the results. The indication of the distribution of L2 speakers by country, now offered by Ethnologue, has made it possible to eliminate the major bias of the method which consisted in extrapolating the results in terms of L1 to L2 (bias that favored languages such as English and French with a high L2 population in countries with low connectivity rate). It remains a notable bias in the method, that of considering that, within a country, the percentage of people connected to the Internet is the same for all existing languages. Thus, the percentage of Hindi speakers connected to the Internet in India is calculated identically to the percentage of English or Bengali connected speakers and the various other languages with more than 1 million L1 or L2 speakers; the reality is probably different, with some languages with rates higher than the national average and others with lower rates, based on cultural, educative or socio-economic factors. This bias has been considered acceptable if it is not intended to compare languages within a country and if languages with low numbers of speakers are not processed; in any case, it is important to know it when interpreting the results.

The other biases result from the selection of sources for the calculations of the indicators and are summarized in the following table, scoring from 0 (unacceptable biases) to 20 (totally free of biases) each indicator and showing the changes between 2017 and 2021.

Bias assessment

Indicators	**Assessment**	**Comments On Biases**
Internauts	19 ->16	The main source is the ITU[10]. In 2017, it was the best-rated source with a 19/20, but in this version, the score drops to 16 because the ITU has stopped providing its estimate when the country does not produce official data (the World Bank[11] is the secondary source in those cases).
Indexes	15 -> 18	This indicator is derived from a combination of 25 micro-indicators (in 2017 there was a single source with 5 parameters). The sources are international organizations, NGOs or universities.
Contents	5 -> 8	There are only 13 micro-indicators to build this indicator and 11 of them come from the excellent Wikimedia statistics. However, Wikimedia does not reflect the true diversity of the Web, being somehow biased towards a western view. A weighting system has been implemented to reduce this dependency a bit. It is quite a sensitive bias.
Traffic	13 -> 11	This indicator is derived from measuring traffic by country using Alexa.com on a selection of 338 Web sites. In 2017, the analysis showed that Alexa was negatively biased towards Asian countries and Brazil. In 2021, new biases are detected that now affect some European countries.
Interfaces	19 ->19	These are objective data. However, it remains a "radical indicator" that omits the vast majority of the world's languages and focuses on a very limited subset.
Usages	12 ->12	This indicator is mainly based on subscription data by country to the most popular social networks (Facebook, Twitter, Linkedin, etc.), which implies a bias against non-Western countries where there are alternative applications.

[10] The International Telecommunication Union (http://itu.int), the United Nations agency that provides statistics on telecommunications, including the percentage of people connected to the Internet by country.

[11] https://data.worldbank.org/indicator/IT.NET.USER.ZS

There is an ongoing study focusing on French, with the support of the OIF[12], which have allowed to extend the scope of languages and aims to reduce the three more important remaining biases by trying alternative ways to reflect the applications of Asian countries that offer services similar to those of Wikimedia or the most popular social networks. Additionally, the coming version will increase sensibly the number of websites of the sampling and therefore allow a finer analysis of thematic differences for given languages (the themes provoking more or less traffic, as for the case of French MOOC sites has been found to create much more traffic than the average). Meanwhile, the referenced documents offer results with correction "by hand" for biases (but applied only to a small part of the results).

We consider an ethical duty to give the elements of appreciation of the biases before offering results to avoid the so common and so nefarious practice of using data found on the Internet as facts, without the precaution to carefully check corresponding methodology and biases.

Results

Let's look at the most powerful languages first using the bias-corrected data from the Observatory.

Power ranking

	POWER
English	25,0%
Chinese	15,0%
Spanish	7,0%
French	3,5%
Hindi	3,5%
Portuguese	3,0%
Russian	3,0%
Arabic	2,5%
German	2,5%
Japanese	2,5%
Malay	1,8%
Italian	1,4%
Turkish	1,2%
Korean	1,2%
Bengali	1,2%
Vietnamese	0,7%
REMAIN	**25,0%**
TOTAL	**100%**

The macro-indicator power, by definition, favors languages with the highest number of speakers. Let us then look at the languages that lead the capacity and gradient macro indicators, as well as the most connected languages to have an indication that is independent of the number of speakers.

[12] https://francophonie.org

Capacity ranking

	CAPACITY
Norwegian	4.65
Hebrew	4.40
Estonian	*3.93*
Finnish	3.49
Serbo-Croatian	3.18
Swedish	2.64
Dutch	2.28
Danish	2.21
Catalan	2.14
Italian	2.11
German	2.11
Macedonian	2.08
Japanese	2.08

The bias resulting from the use of Wikimedia statistics significantly favors the languages that have invested in this space (it is the case of Hebrew or Swedish, for example). It is important to look at the list globally, without relying too much on order, and to find that it points to the national languages of countries and regions recognized for their leadership in the information society field and/or the digital economy.

The most connected languages are the following:

Percentage of connected speakers ranking

	% SPEAKERS CONNECTED
Norwegian	97.87%
Danish	97.82%
Swedish	93.49%
Japanese	92.62%
Dutch	92.03%
Limburgish	91.90%
Swiss German	91.56%
Catalan	90.47%
Western Flamengo	90.43%
Finnish	89.67%
Estonian	89.10%
Latvian	88.95%

And finally, the languages with the highest gradient, meaning with the connected population the most dynamic:

Gradient ranking

	GRADIENT
Hebrew	2.82
Norwegian	2.60
Estonian	2.41
Serbo-Croatian	2.24
Finnish	2.13
Malagasy	2.13
English	1.73
Swedish	1.54
Italian	1.53
Polish	1.27
Spanish	1.25

The presence in this table of Malagasy, a language with an extremely low number of connected people (less than 10%), raises questions about the method. It happens that the Malagasy have a presence, in some of the segments of Wikimedia[13], which is highly disproportionate in relation to its presence on the Internet and, due to the combination of averages, the disproportion is so enormous that it manages to impact the final results. It is one of the symptoms of the biases of the content indicator on which we continue to work.

Compared to the previous edition of 2021, the new languages that appear high in capacity and gradient in the edition with 330 languages, confirm the diagnosis of 2017: Norwegian, Slovenian, Estonian and Catalan are associated with countries or regions with high marks in Information Society considerations.

The case of English

Let us focus now on English, historically the first and more powerful language on the Internet. Its relative place continues to shrink and goes from 30% in 2017 to 25% in 2021, even if the media, relying on strongly biased results, for lack of serious consideration of multilingualism and in contradiction with the evolution of the Net, continue to report figures above 50%.

The Observatory of Linguistic and Cultural Diversity on the Internet has carried out measures on the place of English and other languages since 1998, with a long absence between 2007 and 2017, due to the evolution of search engines[14]. Since 2011, a company specializing in surveys on the use of web technologies, W3Techs[15], includes the language of websites in its list of surveys[16]. It has become the reference for data on the presence of languages on the Web, in the absence of any other choice. Its results, as of October 1, 2021, for the first languages, with more than 2% of the contents, are presented in the following table. The last values produced by the Observatory for Linguistic and Cultural Diversity on the Internet are entered in the last columns for comparison purposes.

[13] Especially the "wiktionnary" (https://en.wiktionary.org/wiki/Wiktionary) where it comes to have 18% of the total entries.

[14] The first method was based on counting the occurrences in the Web of a collection of words chosen to ensure the best semantic and syntactic equivalence between different languages. After 2007, the figures announced by the search engines lost all credibility and the method had to be abandoned.

[15] https://w3techs.com covers more than 20 different types of "technologies", such as operating system of servers, hosting providers or traffic analysis tool.

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

Results comparison W3Techs vs. Observatory

	W3TECHS RANK	W3TECHS SHARE	OBSERV. RANK	OBSERV. SHARE	OBS/W3 RATIO
English	1	62.8%	1	26.5%	0.42
Russian	2	7.3%	7	3.1%	0.42
Turkish	3	3.8%	13	1.2%	0.32
Spanish	4	3.7%	3	8.7%	2.35
Persian	5	3.5%	19	0.7%	0.20
French	6	2.5%	4	3.7%	1.48
German	7	2.1%	9	2.8%	1.33
Chinese	10	1.3%	2	13.9%	10.69
Arabic	11	1.2%	8	3.04%	2.53
Portuguese	13	0.7%	6	3.37%	4.81
Hindi	32	0.1%	5	3.42%	34.20

The differences between the two sources are extremely important and raise questions. The first and major difference is the measure of the presence of English and other significant differences are in Chinese and Hindi, respectively more than 10- and 30-times lower figures for W3Techs.

In addition, W3Techs retains the history of its results, since its inception in 2011, and shows stability in English, and even growth, from 57.6% in 2011 to 63.2% in 2021[17].

Yet the Internet has changed dramatically over the past decade with a massive influx of Asian-speaking and Arabic-speaking Internet users! The global percentage of connected English speakers (L1 + L2) has decreased from 32%[18], in 2007, to 15%, in 2021. Why would the content curve be different? Is it credible, as W3Techs points out, that Chinese and Hindi speakers connected to the Internet, who together represent around 22% of the total number of people connected, only gather 1.4% of the content on the Internet?

Difficult not to paraphrase the famous expression "*It's economy stupid!*" replacing *economy* by **multilingualism**.

The answer to these questions lies in the lack of management of multilingualism in the W3Techs method. W3Techs, undoubtedly an expert company in technology survey management, applies a comparable method for languages which are however a rather different "technology" (see Pimienta, 2021) from those which the company has mastered.

The company's method is to apply a language recognition algorithm to the entry page of the top 10 million most visited websites, without crawling other pages on the site and without worrying about multilingualism. Thus, single webpage sites are counted in the same way as sites of hundreds or thousands of web pages and the many sites which offer several language options for their interface are counted only in the implicit language of the programmed robot, which is most probably English. The presence of a few words in English on the entry page of a site (for example navigation or copyright or a summary, as in the case of scientific articles) is probably enough for it to be counted as an English site, even if all the other pages do not include any English words. Under these conditions, the size of the error can be huge. In addition, the percentages are calculated on the world population whereas it should be computed on the population of L1+L2 speakers, which artificially inflates the percentages and hides the error in the values for the rest of the languages which are not exhibited.

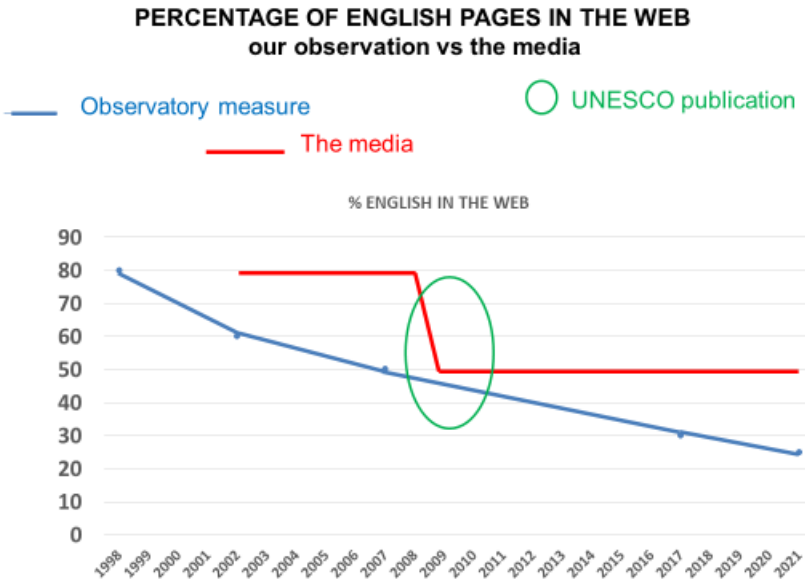
[17] https://w3techs.com/technologies/history_overview/content_language/ms/y

[18] Estimation of the Observatory for Linguistic and Cultural Diversity in the Internet.

You do not have to be an expert in statistics to understand that the method because it does not consider the reality of multilingualism, can be wrong in gigantic proportions. What could the W3Techs algorithm do to improve its figures, without abandoning a pragmatic approach, that is to say without getting into the challenge of analyzing all the pages of all the sites?

- ✓ Analyze the language options offered on the home page and increase the counter for each option just like the English version.
- ✓ Find a method to obtain an estimate, even a rough estimate, of the number of pages on the website and multiply each language version by that number to have an accounting in pages and not in sites;
- ✓ When there is no language option specified and the algorithm reports more than one language on the entry page, as a matter of principle, do not count it as English.

The same phenomenon of the marked difference between observation and measurements had been observed between 1998 and 2007 and had ceased, for a time, when UNESCO published on this subject (Pimienta, 2005) and (Pimienta, Prado, Blanco 2009).



Until 2009, the media published reports on the presence of English on the Web positioning it, unchanged for a decade, at 80%, while our measurements indicated a progressive decline towards 50%. The media relied on 3 publications that proposed, with the same methodology, the same results, in 1997, 1999 and 2002. The methodology was not really biased but scientifically invalid[19], see (Pimienta, 2009) for more details. After

[19] A language recognition algorithm was applied **once** to the entry page of 3000 sites chosen randomly from IP numbers and the percentages were calculated. A legitimate statistical method to validate this approach would repeat the random sampling a large number of times and then analyze the random variable obtained with statistical tools (average, variance, etc.) to approach its distribution pattern and deduce figures with corresponding confidence intervals. A single archery shot on a target does not usually inform on the shooter's ability!

UNESCO publications on the subject (Pimienta, 2006, 2009), the media[20] progressively adopted that new value of 50%. Then W3Techs appeared as the only source, whose results maintained a value between 50% and 60% since 2011[21].

To close this chapter, the fact that the proportion of web pages in English decreases in no way means that the presence in absolute terms of English is decreasing, nor does it mean that it has finished growing; it just means that new languages are taking up more and more space on the Internet, reducing the proportion of English.

The fight against disinformation has become a major issue in this period of health crisis where it can lead to death. Following (Pimienta, Rodríguez, 2020) the need to develop comprehensive **information literacy** programs is an emergency as acute as that of global warming. These programs must include the education of citizens to deal with the data offered on the Web, with a critical mind, and a firm demand on methodological and algorithmic transparency, including an honest presentation of the biases inherent in any approach of constructed data, whether statistical or by other methods. It is clear that advances in artificial intelligence, based on the intensive use of data, will make that requirement still more acute.

References

Pimienta, Daniel, Müller de Oliveira, Gilvan. Cyber-geography of languages. P2: the demographic factor and the growth of Asian languages and Arabic - International Review on Information Ethics, Vol 32 (12/2022)

Pimienta, Daniel. New and improved version of an alternative approach to the production of linguistic indicators on the Internet. Observatory of linguistic and cultural diversity on the Internet – 8/2021.

https://funredes.org/lc2021/ALI%20V2-EN.pdf

Pimienta, Daniel. "Is language a technology or a culture?" Imminent Question of the Year - 2021 https://imminent.translated.com/question-of-the-year

Pimienta, Daniel, Rodríguez Luis German. "Rock the Internet Blues: A critical vision of the evolution of the Internet from civil society", Revista Ibero-Americana de Ciência da Informação, V13 N3, pp. 979-1000 – 2020

https://periodicas.unb.br/index.php/RICI/article/view/33041/27497

https://funredes.org/RockInternetBlues (English version)

Pimienta, Daniel. An alternative approach to produce indicators of the presence of languages on the Internet. Observatory of linguistic and cultural diversity on the Internet, 2017

https://funredes.org/lc2019/Alternativa%20Lengua%20Internet.docx

Pimienta, Daniel, Prado Daniel, Blanco Alvaro. 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-

https://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016

Pimienta, Daniel. 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

https://unesdoc.unesco.org/images/0014/001421/142186e.pdf

[20] And also, unfortunately, Wikipedia (https://en.wikipedia.org/wiki/Languages_used_on_the_Internet) of whom one expects more caution.

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