The method behind the unprecedented production of indicators of the presence of languages in the Internet

Daniel Pimienta
Observatory of Linguistic and Cultural Diversity on the Internet
http://funredes.org/lc

August, 2022

ABSTRACT

A complete description of the methodological elements involved in the production of an unprecedented set of indicators of the presence of 329 languages with more than 1 million L1 speakers in the Web. The paper put a special emphasis in the treatment of the comprehensive set of biases involved in the process, either from the method or the various sources used in the modelling process. The biases related to other sources providing similar data are also discussed and it is shown how the lack of consideration of the high level of multilingualism of the Web leads to a huge overestimation of the presence of English in the Web. The detailed list of sources is presented in the various annexes and the Excel file supporting the model is left in access, with the only limitation of the hiding of the Ethnologue source concerning demo-linguistic data.

KEYWORDS: Languages, Web, Internet, Indicators, Methodology, Bias, Webometrics

ACKNOWLEDGMENT

The work conducing to version 3 (and version 1) of the described model has been realized thanks to the funding of **Organisation de la Francophonie** and with the valuable assistance of **Álvaro Blanco** (without his special ability in writing support programs on demand, this study could not have been concluded) who is to be credited for the new approach for the *traffic* indicator. The foundation of the method is partially based on ideas exposed by **Daniel Prado** in 2012, particularly the idea, applied to the Web, of using country statistics crossed with demo-linguistic figures to create language figures. Thanks to **Gilvan Müller de Oliveira**, from the **UNESCO Chair on Language Policies for Multilingualism**, for his support and the coordination of the **Brazilian government** and the **International Portuguese Language Institute** contribution which allowed for version 2 of the model, an indispensable step towards version 3.

Table of contents

TABLES AND FIGURES

1. INTRODUCTION

The measurement of the space of representation of languages on the Internet does not fascinate the crowds and yet the stakes, on the linguistic, cultural, socio-economic and even geopolitical levels, are far from being neutral. Many languages are threatened or simply in decline and the intensity of their presence on the Internet is a determining indicator of their future. In 2020, e-commerce accounts for 20% of total global retail sales[1] and platforms must speak the language of their customers in order to be in capacity of competing (see Annex 8).

Since 2011, policy makers and linguistic researchers had to rely exclusively on two available sources, both from the business marketing area, for evaluating the impact of their policies or sustaining their theories:

- ✓ W3Techs offers the percentage of Internet contents per language[2], for the 35 top languages, with a daily update, and also keeps the history[3].
- ✓ InternetWorldStats offers the percentages of connected speakers to the Internet for the 10 first positions[4], with a yearly update.

Since March 2022, the Observatory of Linguistic and Cultural Diversity in the Internet (the Observatory hereafter) offers both indicators and some meaningful additional ones (check Pimienta, 2022)[5], for 329 languages[6], with plans for yearly updates. This is the reach of a long process of bias depuration of a method defined in 2017 (Pimienta, 2017), which finally provides outputs with an acceptable threshold of reliability (within a +-20% confidence interval).

The analysis of W3Techs method reveals huge biases as a consequence of not taking into account the important amount of multilingualism prevailing in the Web (see 4.1). The computations of InternetWorldStats rely on the combination of percentage of connected people per country, a trustable figure yearly available from ITU[7] and World Bank and demo-linguistic data for L1 and L2 speakers. The existing sources on demo-linguistic data have large differences, especially for L2 numbers; among them, Ethnologue is considered the most reliable, although not free of charge[8].

The Observatory is not a newcomer in that field: it has been conducting a series of pioneering measurements of Web contents for English, German and Latin Languages[9], between 1997 and 2007 (Pimienta et al., 2009). The method made use of the total of word or expressions occurrences in Webpages, provided by Search Engines exploring a large percentage of the Webspace. The Observatory was obliged to resign, after 2007, when Search Engines stopped providing trustable figures and the proportion of indexed webpages was considerably reduced.

In 2017, the Observatory developed a new method allowing the production of a set of indicators for the 139 languages with more than 5 million L1 speakers (Pimienta, 2017). The method stood

---

[1] Source : https://www.digitalcommerce360.com/article/global-ecommerce-sales/
[2] https://w3techs.com/technologies/overview/content_language
[3] https://w3techs.com/technologies/history_overview/content_language/ms/y
[4] https://www.internetworldstats.com/stats7.htm
[5] http://funredes.org/lc2022
[6] Those holding a population of L1 speakers higher than one million
[7] The International Telecommunications Unit (http://itu.int), the United Nations body that provides telecommunications statistics, including the percentage of people connected to the Internet by country.
[8] https://www.ethnologue.com/data-consulting
[9] French, Italian, Portuguese, Spanish and Romanian.

in a new approach, defined in 2012 and applied for single languages, mainly French (OIF, 2014) and Spanish (Pimienta, Prado, 2016), which focused the management of a set, as large as possible, of disperse sources of figures about languages or countries, having some type of relationship with the Internet. This relation could be direct (e.g. repartition per country of subscribers to a specific social network or languages supported in on-line translation services) or indirect (e.g. ranking in e.commerce or average number of mobile per person in each country). The scarcity of figures related to languages in the Internet[10] was compensated by the use of figures related to countries, more numerous, those being transformed into figures per language, by weighting with demo-linguistic data. The collected figures were organized into different themes: *contents*, *traffic*, *usages*, *indexes*[11] and *interfaces*. In 2017, giving mathematical coherence and using statistics technics, for instance to extrapolate missing data, the method was generalized for many languages, beyond French or Spanish. A model was designed to process the whole set of sources into meaningful indicators for the 139 languages with L1 speakers counted over 5 million.

Starting from there and since 2017, the work was essentially dedicated to the struggle against the various **biases** proper of the method or the data sources. This fight leaded to a Version 2, in 2021, with the same structure, but some important biases controlled, in particular thanks to the use of Ethnologue Global dataset 24 of March 2021. The language coverage was extended then to the 329 languages with L1 speakers higher than one million. The pursuit of the fight against biases continued and leaded, in March 2022, to a final redefinition of the approach and the trust that a reasonable level of control of biases has been reached, with the capacity to produce figures reliable, within a confidence interval of -+20%.

The elaborated model produces, for each language, the following indicators, all figures applied for L1+L2:

A. Share of world L1+L2 speakers
B. Percentage of connected L1+L2 speakers
C. Share of the world L1+L2 connected speakers
D. Share of total Internet contents
E. Virtual presence indicator (defined as the ratio D/A)
F. Content Productivity indicator (defined as the ratio D/C)

More elaborated constructions are made from the aggregation of those indicators, such as the following *Cyber-Geography of Language' Families* table, which gives a global perspective of the situation of the different *language families*[12] and shows that Asian languages are in the way to take the lead over European languages while African languages are in a difficult situation, due to the prevailing digital divide translating into a language divide[13].

[10] With the notable exception of the Wikimedia Foundation offering figures for each of the provided services and for the 327 supported languages (source: https://en.wikipedia.org/wiki/List_of_Wikipedias).
[11] Index refer to rankings in different parameters associated with Information society progress.
[12] The language families include, for each region, the languages which are native of that region. English, French and Spanish are European languages and, following the Ethnologue classification that we use, Russian is classified as European language while Turkish and Hebrew as Asian languages.
[13] Less than 30% of African language's speakers connected to the Internet and very low virtual presence and content productivity are obtained.

Table 1: Cyber Geography of language families

Languages from	Africa	Americas	Arab world	Asia	Europe	Pacific	Not included	TOTAL
Speakers L1+L2	9.21%	0.31%	3.53%	48.24%	30.91%		7.81%	**100%**
Internauts %	29.8%	56.7%	64.0%	49.3%	82.6%		47.06%	**56.91%**
% from Internauts	5.21%	0.32%	3.89%	44.63%	39.51%		6.36%	**100%**
Contents	2.89%	0.22%	3.09%	44.77%	45.39%		3.64%	**100%**
Virtual Presence	0.31	0.71	0.88	0.93	1.47		0.47	**1**
Contents Productivity	0.56	0.69	0.79	1.00	1.15		0.57	**1**
Number of languages	138	8	1	135	47	0		**329**

The products of the established model are available in CC-BY-SA 4.0, from http://funredes.org/lc2022 and presented in (Pimienta, 2022). This paper is focusing **the method** which have allowed the production of those indicators.

2. THE THEORY BEHIND THE MODEL

2.1 OVERVIEW

It is an **indirect approximation** to Web contents per language, based on the experimental observation consistently made, since the beginning of the Observatory, that the ratio between *world percentage of contents* and *world percentage of connected speakers* (defined as *content productivity*) has hardly be measured outside of the window [0.5 --- 2], for languages with full digital existence.

This observation suggests the existence of some kind of natural economic law, which would link, for each language, the **offer** (web contents and applications in the given language) to the **demand** (speakers of that language connected to the Internet). When the number of connected persons increases, the number of webpages naturally increases together, in **more or less** the same proportion. This happens because governments, businesses, educative institutions, etc., and some individuals create contents and applications to respond that demand.

It is important to note, in support to the previous statement, that surveys and studies on Internet user's behavior have been consistently reporting that Internet users prefer to use their mother tongue, when contents are available, especially for e.commerce, and in complement are eager to make use of their second language(s) (see, in Annex 8, a selection of sources to support that claim).

Thus, depending of each language context, there is some kind of modulation of the mentioned ratio, to make it, **more or less**, above or below one. Some languages have a better *content productivity* than others, depending on a set of factors proper of the language or related to the different country's context where some proportion of the speakers of that language connects to the Internet. The following factors has been identified:

Proper of the language:

- Obviously, the relative amount of L2 speakers, as some people produce, for instance for economic reasons, contents in language different from their mother tongue.

- The technological support of the language for cyberspace, reflected somehow in its presence in application's interfaces and translation programs, which would make easier or not the content production.

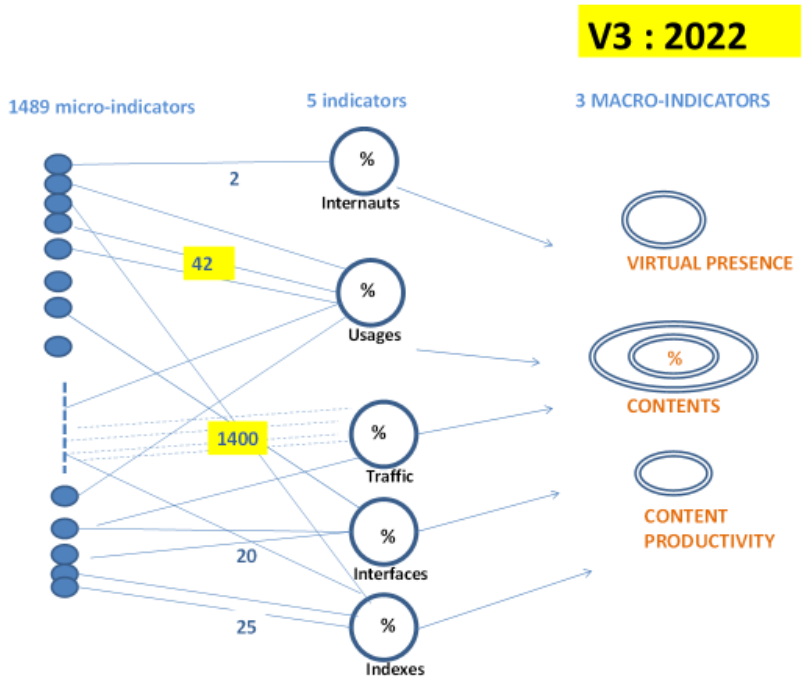
But also, depending on each country where there is L1 or L2 speakers of this language:

- The amount of Internet traffic, depending of country's tariff, cultural or educational context.
- The number of subscriptions to social networks and other Internet applications.
- The level of progress of the country in terms of Information Society services (e-commerce, government applications to pay taxes and so on).

Therefore, if it was possible to collect sufficient and meaningful figures about each of the mentioned factors to create corresponding indicators, one would approximate the value of the *content productivity* ratio and, from the proportion of speakers connected, deduce the contents proportion.

This is the core of the method and it is synthetized in the following diagram which shows all the indicators which are processed for each language and the corresponding quantity of sources the model is using.

Figure 1: From sources to products



2.2 DESCRIPTION OF THE INPUTS OF THE MODEL

The inputs of the model are split into 5 types of sources: *internauts*, *usages*, *traffic*, *interfaces* and *indexes*.

2.2.1 Internauts

This the percentage of L1+L2 speakers connected to the Internet for each language. The transformation of the source' figures, expressed by country, into the required figure, expressed per language, is made by weighting:

CS(j) is the percentage of connected speakers for language j.

$$CS(j) = \sum_{i=1}^{i=P} SP(i, j) \times CC(i) / \sum_{i=1}^{i=P} SP(i, j)$$

Where:

P is the total number of countries
SP (i, j) = The number of L1+L2 speakers of language j in country i.
CC (i) = The percentage of connected persons for country i

The matrix product CS = SP + . x CC in APL[14] notation or = SumProduct (SP;CC) in Excel notation, is a weighting operation which produces from a vector the size of the number of countries, a new vector, this time the size of the number of languages.

The validity of this computation stands on the implicit hypothesis that, within the same country, all language groups share the same figure for the percentage of connected persons. This is one of the founding biases of the method, discussed in the chapter Biases.

The vector CS(j) is a key element of the model which will serve, again in weighting operations with various sources, to compute the modulation of each indicator.

The source for the SP matrix is Ethnologue; the model uses Ethnologue Global Dataset #24 of March 2021. The sources for the CC matrix are the International Telecom Unit (ITU) [15] and the World Bank; the ITU, the historical source of those figures, relies on government reports and when not available in its own estimation. As ITU has stopped in 2017 providing its own estimations, the source is completed by figures[16] from World Bank which is filling that gap in many cases. When no recent figures are available an extrapolation of older figures is kept.

2.2.2 Interfaces

Researchers from the MetaNet network[17] are doing a fair job in analyzing the technological support for European languages but there is not yet some type of metrics to evaluate the technological support for all languages in the world. In order to approximate that parameter, the focus has been put in the presence of each languages in the interfaces of a set of popular Internet applications and as one of the pairs in on-line translation services. Sixteen elements have been identified where the list of supported languages is accessible. The list of measured applications can be read in Annex 3.

2.2.3 Indexes

The theme here is to rate countries in regards to their progress on Information Society criteria's. A further weighting with demo-linguistic data will transform this figure into a rating of languages. In version 1, a list of 4 sources was used. Starting in Version 2, a systematic search

[14]APL, "A Programming Language", which is both a mathematical formalism and its implementation in the form of programming language, designed by Kenneth. Iverson. For more details see https://fr.Wikipédia.org/wiki/APL_ (language).
[15] https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2021/December/PercentIndividualsUsingInternet.xlsx
[16] Source : https://data.worldbank.org/indicator/IT.NET.USER.ZS
[17] http://www.meta-net.eu

was realized and 27 sources were identified making the selection almost exhaustive (see Annex 4).

2.2.4 Usages

Five sub-indicators have been identified and corresponding sources has been used:

- Subscribers to social networks: 36 sources have been used, each related to social networks with more than 100 million subscribers. For the main occidental social networks, direct figures on subscribers per country have been identified; for the remaining social networks, mostly from Asia, partial traffic figures per country has been used, using SimilarWeb[18], extrapolated to the rest of countries, proportionally to the percentage of connected persons per country.
- E.commerce: a single source has been used which does the job perfectly. This is the T-index indicator from the Imminent Translated Research Center[19]. This indicator ranks countries according to their potential for online sales thus estimating the market share proportion of each country in relation to global e-commerce. The set of percentage per country is transformed by weighting with the connected speakers per language into a set of percentage per language[20].
- Video streaming: the model makes use of only two sources at this stage: the percentage of Netflix subscribers per country and YouTube penetration per country.
- Open contents: the model makes use of only one source at this stage: the percentage per country of the sum of 2012/21 % OpenOffice downloads.
- Infrastructure: the model uses three keys World Bank's figures which are merged into two indicators: % Fixed broadband subscribers per country and % Fixed Telephone + mobile subscribers per country.

The final results have been first weighted, to reflect the today trust in the figures[21] and therefore reduce the biases, with the following values:

- Subscribers to social networks: 0.3
- E.commerce: 0.3
- Video streaming: 0.05
- Open contents: 0.05
- Infrastructure: 0.3

and then transformed by weighting into repartition per language.

The detailed list of sources for USAGE is in Annex 1.

2.2.5 Traffic

Tools (such as SimilarWeb already mentioned) exist to obtain an estimate of traffic repartition per country to any specific website; in general, those tools offer data for websites which are

[18] A marketing service providing proportion of traffic per country to a large set of websites: https://www.similarweb.com/

[19] https://imminent.translated.com/t-index

[20] Note that Imminent provides also the set of percentages per language, probably doing a similar operation. There are slight differences between Imminent and our computations, probably because of different demo-linguistic data. The model uses our computations instead of the direct Imminent source because Imminent is limited to 89 languages while the extrapolation technic with our computations allows to reach all the languages of the study.

[21] A simple average without weighting will be used in next release when each element obtains the necessary sources to fil it.

ranked within the first million or ten million more visited. The challenges here are to evaluate those tools and understand their potential bias and to establish a selection of websites with minimum bias, while staying in a workable size (says less than or around 1000). Many changes occurred from version 1 to version 3 to overcome the biases; they are described in 2.4.5. The list of websites used for *traffic* is in Annex 5.

2.2.6 Contents

Contents was an input of the model for the two first versions, as the original methodological vision was to collect a maximum of sources and Wikimedia, which collects, for each one of its applications[22], and for each supported language, reliable and interesting statistics per language, non-withstanding the fact that it is probably the more multilingual application of the Web with its 327 linguistic versions. Version 3 decided to cancel this indicator from the input list. The chapter 2.4.8 discusses the biases and gives the rationale for that decision.

2.3 DESCRIPTION OF THE OUTPUTS OF THE MODEL

The model provides the following outputs, for each language:

Speakers: the share of world L1+L2 speakers

Connected Speakers: the percentage of speakers connected to the Internet

Internauts: the total share of connected speakers expressed in percentage

Contents[23]: the total share of Web contents expressed in percentage

Virtual presence: the ratio of *contents* over *speakers*.

The world value (and average) is 1: a value higher than 1 means a virtual presence higher than the real-life presence.

Content productivity: the ratio of *contents* over *internauts*

The world value (and average) is 1: a value higher than 1 means high productivity of connected speakers.

Cyber-globalization index: CGI (L) = (L1 +L2)/L1(L) x S(L) x C(L)

where:

L1+L2/L1 (L) is the ratio of multilingualism of language L (from Ethnologue source)

S(L) is the percentage of world countries which hold speakers of language L (from Ethnologue source)

C(L) is the % of speakers of language L connected to the Internet.(computed by the model)

This is an indicator of the strategic advantages of a language in cyberspace[24].

In addition, regrouping results per language family, the previously shown table *Cyber-geography of languages* is produced by grouping the previous indicators by language families[25], producing interesting global perspective on the situation and trends.

---

[22] Wikipedia, Wiktionnary, WikiBooks, WikiQuote, WikiVoyage,WikiSources, Wikimedia Commons, WikiSpecies, WikiNews, Wikiversity and WikiData.

[23] In the two first versions, as *contents* was an input, the outputs indicators were called *Power*, *Capacity* and *Gradient*, with exactly the same definition as today *Content*, *Virtual Presence* and *Content Productivity*.

[24] In terms of %, English + French hold almost 25% of the weight, followed, somehow faraway, by German, Russian, Spanish and Arabic.

[25] The definition of language families used is the one of Ethnologue.

2.4 ANALYSIS OF BIASES

The following table shows the evolution of biases from V1 to V3 using a subjective rating from 0 (biases so huge that data is meaningless) to 20 (absolutely free of biases), with 10 (notable but bearable biases) in the middle.

Table 2: Bias assessment

BIAS ASSESSMENT Rate over 20	V1 2017	V2 2021	V3 2022
CORE METHOD	**17**	**17**	**17**
METHOD FOR L2	**13**	**19**	**19**
INTERNAUTS	**19**	**16**	**19**
INDEXES	**15**	**18**	**18**
CONTENTS	**5**	**8**	**OUT**
TRAFIC	**13**	**11**	**17**
INTERFACES	**19**	**19**	**19**
USAGES	**12**	**12**	**16**

2.4.1 Core Method

The implicit bias of the core of the model is to consider that all the languages in the same country share the same rate of connectivity to the Internet (the national value provided by ITU). The reality is obviously different as the concept of digital divide also exists within each country.

This working hypothesis provokes a positive bias for speakers of non-European languages living in developed countries (who are probably less connected than the average) and reciprocally a negative bias for European languages speakers in developing countries (who are probably more connected than the average). Being a foundation of the method, this hypothesis cannot be changed and the decisions taken to deal with it are :

- Comparisons between language's performance within a country are not allowed.
- As the risk of important bias grows inversely proportionally to the size of the speaker's population, the study was first limited to languages with more than 5 million L1 speakers and later extended to languages with more than 1 million speakers. Future versions may try to extend this threshold but probably never below 100 000 as the biases could become unavoidable.

2.4.2 Method for L2

For the first time, in 2021, Ethnologue extended its demo-linguistic data per country to L2 speakers. This allowed to remove one of the most important bias of the method (in V1) which resulted in extrapolating data (for example percentage of connected speakers) from L1 to L2, a process which biases positively the results of languages with high presence in developing countries, such as English and French. Indeed, this process assigned to L2 speakers in developing countries Internet connection rates higher than the reality. Starting in V2, with the existence of demo-linguistic data per country for L2 as well as L1, the core method is directly applied to L1+L2 populations and this extrapolation bias disappears but obviously not the core method bias which applied the same for L1, L2 and L1+L2.

It remains that the demo-linguistic source itself has a larger bias for L2 figures than for L1 figures as there is no perfect definition of the level of mastering of a second language required

to be computed as L2. As a matter of fact, the L2 figures' sources vary in huge proportion, especially for English[26].

2.4.3 Internauts

This is, after the demo-linguistic data, the second main element of the model and it needs to be standing in reliable source. As mentioned in 2.2.1 figures from ITU and World Bank are combined to obtain the best and mostly reliable up to date data.

2.4.4 Indexes

With the extension of the sources in V2 reaching close to exhaustivity and a selection from reliable institutions (international organizations and non-Governmental organizations) the selection bias is minimal and the trust in the data is maximal.

2.4.5 Traffic

The available tools providing repartition of traffic per country to a large set of websites (the ones considered as the most visited) are : Alexa.com, SimilarWeb.com, Ahrefs.com and Semrush.com. All are from marketing companies, not totally transparent about their method. For instance, Alexa, the older and most famous, although it has ceased activities in May 2022, performs from a banner that users can download. This banner, associated to a Web browser, reports to Alexa the sites which are visited by the user from this browser. With the collection of all the data sent by all the banners around the world, Alexa builds its outputs, both in term of ranking sites and traffic repartition per country. It is obvious that the geographical repartition of banners could be an indication of probable biases but unfortunately this information is not published.

The process of this indicator has been the most time consuming in order to overcome biases. In version 1, Alexa.com was used, with a selection of 450 websites. It was established, by comparing the traffic figure per country from Alexa with the subscriber's figure per country, collected from various sources, that Alexa was positively biased for English and French and strongly biased against Asian countries and Brazil. In order to fight the unavoidable *selection bias*, the process of the indicator was not realized by simple average but rather by a *reduced mean* with a large 20%, trying to mitigate that way the selection biases.

Trials in version 2 showed that Alexa seemed to have corrected the Asian negative bias but a new bias appeared to affect then European countries. Further trials leaded to the discovering of a bug where the main country in terms of traffic was sometimes not listed and this could be the reason of the observed bias in the results, as it happens specially with European countries. It was then decided to use Alexa only when the sum of percentages offered was higher than 70%, a simple way to eliminate those mistaken cases. Ahrefs and Semrush were tried but not used because of a strong bias in favor of English and for one of them a total of percentages per country often higher than 100%. SimilarWeb provided results relatively close to Alexa.com, after the mentioned correction, and it was decided for Version 3 to use both tools and retain the figure of half the sum of each.

[26] Ethnologue figure for English is 1.348 billion L1+L2 speakers (L1= 370 million, L2 = 978 million) while other sources propose 1.18 billion (https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population) or 1.5 billion (https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/ real source not cited). In 2008, David Crystal expressed the possibility of that figure tending to 2 billion (https://www.cambridge.org/core/journals/english-today/article/two-thousand-million/68BFD87E5C867F7C3C47FD0749C7D417).

After the many tests and experiments conducted, it was observed and concluded that the *selection bias* was definitively a serious problem which needed to be solved in a more drastic way than with the reduced mean.

Finally, version 3 addressed the situation with a new approach which allowed to manage a selection over 1000 websites where the bias has been reduced by all possible means[27].

To reach the objective of unbiased selection it was finally decided to list a selection of the most visited websites in each country, with a number of websites proportionate to the country global traffic. The algorithm was not set to target all the countries for practical reasons but was limited to the 55 countries holding the top places in terms of contents for languages spoken in those countries.

Table 3: List of countries treated for the selection of national sites

Afghanistan	Algeria	Germany	Angola	Saudi Arabia	Argentina
Australia	Bangladeshi	Belgium	Brazil	Bulgaria	Cambodia
China	Hong Kong	Taiwan	Colombia	South Korea	Egypt
United Arab	Emirates	Spain	United States	France	India
Indonesia	Iran	Iraq	Italy	Japan	Kazakhstan
Kuwait	Lithuania	Malaysia	Mexico	Morocco	Mozambique
Nepal	Nigeria	Uzbekistan	Pakistan	Netherlands	Philippines
Poland	Portugal	Romania	UK	Russia	Singapore
Sudan	Sri Lanka	Tanzania	Thailand	Turkey	Ukraine
Vietnam					

The rule was set to select, for each country, at least the first three most visited sites including among them, at least, the first local domain (ccTld[28], such as .fr for France)). This rule was set to avoid that the selection be too heavily concentrated in the global most visited sites (generally .com). Obviously, this did not avoid that the most global websites (such as Google.com or facebook.com) appear in many country's selection and a weighting was performed to respect those figures.

To realize the selection, the four tools were used (Ahrefs, Alexa, Semrush and Similarweb) although on some occasions, due to lack of data in countries with a small population, we had to collect the data from other sources.

A total of 1421 websites were finally selected automatically[29], of which 733 were different websites. The number of occurrences of each website in the 1421 selection was kept for further weighting. For each country, the number of websites corresponding due to the proportion of the world Internet traffic share was also computed and kept for further weighting, this as away to control the selection bias.

This method insured the most unbiased possible selection for the *traffic* measurement and overcome the huge bias against Asian countries which has been carried since the beginning. It

[27] This decision obliged to resign to an interesting but statistically weak outcome of previous versions which consisted to group the websites by theme and draw for some languages some tentative conclusions about its strength or weakness as related to those themes. The issue about determining if those results reflected more the selection biases than some thematic reality of the language presence in the Internet remained at the same time unsolved.

[28] Country Code Top Level Domain.

[29] The selection process was made by computer programming to avoid mistakes or unwanted biases.

definitively enhanced the final results for Chinese, Hindi and Arabic as well as for the rest of languages.

A remaining bias may still exist which penalize the countries (and the associated languages) where the general level of digital literacy is the highest and for which therefore there is significant traffic to sites with scientific or literary content, and in any case excluding social networks and other world-famous sites. This is unfortunately the price to pay to obtain results free of major biases. It is clear that this marginal bias will not favor the languages of developed countries, that is to say most often European languages.

An enhancement possible for version 4 could be to include a new indicator by establishing the proportion by country of sites in the national domain compared to sites in the generic domain; this indicator could be a first step towards measuring the global degree of digital literacy by country and could even be used through a new weighting to compensate for the residual bias in question. In the meantime, the raw results of the model could slightly disadvantage French and English and on the contrary now seem to slightly favor Chinese.

2.4.6 Interfaces

For each language a ranking is established as the number of times the presence of that language exists in the list of applications selected (interface or on-line translation). From that ranking, the weighting operation with the percentage of connected speakers produces the "modulated" percentage expected. Obviously, this indicator is quite "aggressive" as hundreds of languages are totally absent from the list and are therefore attributed a figure of 0%, meaning that there is absolutely no technological support. This harsh measurement is anyway a reflect of the crude reality of that field where too many languages have a level of digital technological support close to none, in spite the growing efforts of the language technology researchers[30].

4.7 Usages

The **subscriber's** element was the reason of a strong pro-occidental biases in versions 1 and 2, due to the absence of non-occidental social networks, and a particular effort has been made in version 3 to complement the 11 initial sources[31] with analogous applications from the rest of the world.

The criteria which has been chosen for complementing was to keep social networks with more than 100 million subscribers. When sources of data per country has been identified (generally repartition of subscribers per country) it is used; when not, the repartition of subscribers per country has been established from the traffic per country, data obtained from SimilarWeb service, and extended to all countries by extrapolation (see 3.4).

The repartition per country, after extrapolation of each element, is weighted in function of the number of subscribers and is finally transformed in percentage per language by weighting with the demo-linguistic matrix.

In version 3, the complementation has considerably reduced the bias against non-occidental countries and indirectly against non-European languages. The complete list of social networks processed is listed in Annex 1.

[30] See the bi-yearly intense conferences and workshops of the LREC researcher's community since 1998 : http://www.lrec-conf.org.

[31] Facebook, Linkedin, Twitter, Instagram, Reddit.

For the **e.Commerce** element, as mentioned previously, the source is unique but quite trustable.

For **video streaming**, the model makes use of only two sources at this stage: the percentage of Netflix subscribers per country. This sub-indicator needs clearly to be extended in the next release of the model with alternative streaming applications beyond YouTube and Netflix with a special effort for non-occidental countries. Meanwhile the element receives a low weighting.

For **open content**, this sub-indicator needs clearly to be extended in the next release of the model with more data related to openness, especially in the field of MOOCs. Meanwhile the element receives a low weighting.

For **infrastructure**, the World Bank figures about fixed lines, mobile and broadband per country are quite reliable and offers a sound basis for the indicator. Summing fixed lines and mobile in a single figure balance the situation between developed countries with large fixed line penetration and developing countries with high mobile penetration.

It remains that, at this stage, the *usage* indicator is the one who have received the less un-bias attention and it needs to be enhanced for next release, although the main objective which was to overcome the occidental bias of working only with the main occidental social networks has been reached for the social networks component and has produced the expected effects on results, revealing the booming presence of Asian countries and languages.

2.4.8 Contents

This indicator is the one, together with *usage*, which has received the higher attention in the work against biases. It is also the one whose biases, inherited from the Wikimedia galaxy, had the major influence on the two first versions results, giving a notable advantage, for indicators independent from speaker's population, to results of the languages with major presence in Wikimedia.

The two main challenges with Wikimedia are, first, that in spite its notable efforts and success to be truly global, it does suffer from an occidental bias and, second, that, some particular languages have invested a lot to participate to the online encyclopedia and show presences hugely disproportionate with the reality of their number of connected speakers[32], while other languages have seen their results in first versions boosted by their heavy presence in Wikimedia services[33]. Furthermore, some languages have artificially boosted their number of articles by translating them from other linguistic version while keeping a rate of updates extremely low.

The focus on unbiasing has been set in those directions. In version 2, a formula was set, and used as indicator, instead of the number of Wikipedia articles, to efficiently remove the mentioned artificial advantage:

$$W(i) = \text{Articles}(i) \times \text{Edits}(i) \times \text{Editors}(i) \times \text{Depth}(i) / \text{L1+L2}(i)^2$$

Where:
i is one of the languages
Articles (i) = the number of Wikipedia articles for language i
Edits (i) = the number of editions of the articles for language i

[32] This is the case of Cebuano, Malagasy and Tagalog.
[33] Like Hebrew, Swedish or Serbo-Croatian.

Editors(i) = the number of editors for the articles for language i
Depth (i) = an indicator of the frequency of updates of articles for language i[34]
L1+L2(i) = the number of speakers first and second language of i.

All elements of the formula are provided in Wikipedia statistics, for more details, see (Pimienta 2021).

For version 3, a profound and systematic effort was dedicated to balance the Wikipedia figures with equivalent figures from other languages. The table exposed in Annex 2 list the online encyclopedias processed with the figure gathered, in terms mainly of number of articles. From that table, the content indicator was built with a fairer representation of languages by cumulating, by languages, the different number of articles. The conclusion of this heavy, necessary, but finally frustrating, effort was that some languages (like Chinese or Turk) have invested massively in online encyclopedias while other appear not interested in that matter. The impact of such violently different figures on the end indicators produced is very high and finally the evidence emerged that online encyclopedias are not honest witnesses of the reality of Web contents and should not be used in the model.

It was a real dilemma to abandon those wonderful statistics of Wikimedia; however, the suppression of *contents* as an input data leaded to the positive renewal of the conception of the approach into a bias light and coherent model.

Instead of calling *power* the main output it was renamed directly **contents** and *capacity* and *gradient,* with the same arithmetic operation turned to become **virtual presence indicator** and **content productivity,** much more natural and understandable concepts. Besides, all the weighting operations which were developed inside the model from version 1 were now reflected coherently in the conceptualization of the approach, as a modulation of content productivity. At the same time, the above-mentioned anomalies on the results, which were driven by the particularities of Wikimedia, disappeared leaving room to more trustable and predictable results[35].

3 MODELISATION

3.1 PRE-PROCESSING

The main part of the data provided by Ethnologue is in the form of an Excel matrix of 11 500 lines in the following format: "ISO639[36]*, Language name, Country name, number of L1 speakers, number of L2 speakers"*, plus a large number of related parameters not used for this method which have been removed.

In order to obtain the format required by the model (a matrix with all the countries considered in columns and all the languages considered in rows), a series of steps was implemented with

[34] See precise definition in https://meta.wikimedia.org/wiki/Wikipedia_article_depth
[35] The best symptom is that Japanese surged in the first place of virtual presence and content productivity which is coherent with the reality of the so pervasive use of the Internet in Japan. Some of the languages which were favored by their high presence in Wikipedia remain in high positions but not at the first places which makes still true the statement that languages of countries (or regions) with highest performance in Information Society parameters benefit from good places in the virtual presence or content productivity indicators.
[36]The 3-character ISO code assigned to each of the 7486 languages identified.

the support of different programs written in the form of VBA macros[37]. One of the most complex steps was to merge all the data from the languages belonging to the same macro-language. This process involved 60 macro-languages comprising 434 different languages[38] (see details in Annex 6).

After completing this step, the process was to reduce the full list of languages to keep only those that are handled by the model, carefully summing all the remaining numbers by country in a single line for the rest of languages.

It is important to understand that the adoption of Ethnologue data entails the acceptance of its rules of presentation, which are based on purely linguistic considerations:

- Grouping of macro-languages[39]
- List of countries and corresponding English denominations.

The list of countries treated by Ethnologue is larger than the one treated by the ITU for the provision of Internet connection rates by country: the ITU, as a United Nations entity, does not separate, for example, Martinique from France. In this case, the ITU rule is the one that prevails and the requirement has been to carefully gather Ethnologue data for the 29 countries not considered by the ITU (for the complete list, see Annex 7) into a single column "Other countries"[40].

3.2 SOURCE MANAGEMENT FOR MICRO-INDICATORS

The whole process of managing sources for micro-indicators is the heaviest and most difficult task of the project, with a high consumption of human resources. Many steps are necessary:

1. For each indicator, check that the sources for 2017 are still available and up-to-date, otherwise search for other comparable sources on the Internet.
2. Select new sources based on their reliability and applicability to the process[41].
3. Collect the selected sources in a format allowing a simplified introduction into the model.
4. Introduce validated sources into the model.
5. Assess source bias.

In Annex 5, the complete list of sources is presented, for each indicator.

To perform step 4, the data must be transformed into Excel format, with the country and language names matching those in the template and in the same sequential order.

In step 3, all sources are collected from a specific URL (see Annex 5 for the complete list of URLs) and most sources are obtained in HTML format. Some sources are in PDF format and a

[37]Virtual Basic Applications, a language used to create executable macros in Excel.
[38]For example, the Arabic macro-language contains 29 languages such as Egyptian Arabic or Moroccan Arabic.
<[39]A significant example is the case of the Serbo-Croatian macro-language whose definition includes, in alphabetical order, Bosnian, Croatian, Montenegrin and Serbian. This grouping does not at all meet geopolitical criteria and could even be considered controversial from this point of view. Moreover, as some sources clearly separate the languages and the countries concerned, this entails a risk of error in the results, even if the entry of the sources has been transformed to take this situation into account (the risk arises when the figures must not be summed but rather averaged as in Wikipedia's depth indicator).
[40] It should be noted that Kosovo does have figures provided by the ITU but is absent from Ethnologue list of countries: for that reason, it does not appear in the results.
[41]It may happen that reliable data is in a format that prohibits automated exploitation.

limited subset (mainly that of the ITU and the World Bank) is in Excel format, the one targeted to transform all sources. The process of converting from PDF to Excel can be relatively simple in most cases, when the tables are well structured, but in some cases, there is an incompatibility and some tricks are needed, such as going through an intermediate .doc format.

The process of transforming from HTML to Excel can often become a real nightmare requiring a lot of imagination, including in some cases the need to go find the data inside the HTML source and try from there to build a table using Excel's conversion function, after cleaning up the HTML code surrounding the data.

In an increasing number of cases, the source offers geographical access to the data (clickable maps) which, except when the number of countries or languages is limited and hand copying is not too cumbersome, makes automated processing impossible or requires the outsourcing of a manual collection work which is tedious and requires high concentration and discipline to avoid errors.

Credit shall be given to the institutions (generally, international organizations or NGOs) that provides the data in a computer-readable format (Wikimedia provides, for example, in its English version, HTML tables that can be transformed directly into Excel format without loss of structure).

Obtaining a copy of the source in Excel or compatible format (usually a table of country names or languages with associated values or percentages) is not the end of the process. With 215 countries and 329 languages to process and, instead of using unambiguous ISO code, the common usage of literal names that can be in different languages and in non-standard spellings, the integration of data into the model cannot be done by hand. Two programs have been written for this process, both of which required recursive tuning[42] to accommodate the different spellings. The program outputs are Excel files that can be used directly to integrate the data into the model. In addition to the appreciable time saving of this computerized method, it guarantees to obtain the data without error.

It should also be noted that the management of macro-languages has made this process even more complex, because the grouping of languages in the corresponding macro-language must be carried out in the source data before processing by the macro. To take a few examples, the frequent occurrences of Egyptian or Moroccan Arabic in the sources have been cumulated into the Arabic macro-language and those of Serbian, Bosnian, Croatian and Montenegrin have been merged into Serbo-Croatian (the number of similar cases being quite high). For the manual processing of unknown spellings reported by the program (incorporation of spellings as synonyms or rejection in the other category), the Ethnologue page descriptive of each language code was used in support[43].

3.3 STRUCTURE OF THE MODEL AND PROCESS

The model is implemented in an Excel file with 17 sheets which are presented below together with the corresponding process.

[42]The recursive process recognizes new spellings and ends when the error check no longer identifies unknown spellings.
[43] https://www.ethnologue.com/language/srp

ITU: a copy of ITU source modified according to pre-processing.

SP: (stand for SPeakers) the matrix of L1+L2 speakers per country.
In lines, the 329 languages, sorted by 3 digits ISO code (ISO369), starting with line 9 with the sum of the rest of languages not processed.
In columns, the 215 processed countries, sorted by 2 digits ISO code, starting in column I with the sum of the rest of countries not processed.

The 8 first lines and columns are reserved for control information:
Control lines: country code 3 characters, country code 2 characters, country name, total country L1+L2 speakers, % persons connected, number of persons connected, world % connected, total or average (number of languages spoken per country), remaining languages.
Control columns: ISO639, language name[44], total L1+L2 speakers, world % of L1+L2 speakers, world % of L1+L2 connected, number of countries with speakers, number of L1 speakers, ratio L1+L2/L1, remaining countries.

This sheet is protected from reading as it contains proprietary information of Ethnologue that cannot be made public.

SP2: (Speakers second data) demo-linguistic secondary data computed from **SP**
For the 329 languages in lines, and the rest of languages: world % L1+L2 speakers, number of L1+L2 speakers, world % of connected L1+L2 speakers, number of L1+L2 connected speakers, world % of connected L1+L2 speakers, world % of connected L1 speakers, world % of L1+L2 internauts.

PL: (Percentage Language) Matrix parallel to **SP** where PL(i,j) = % of language i internauts from the country j connected, computed from **SP and SP2**. It is a redundant information used to simplify the weighting operation performed in sheet **Wut**.

MIl: (Micro-Indicator Language) Holds the list of languages in lines and the value, 0 or 1 of the presence of language in one of the 16 applications used for the *interface* indicator, filled from the sources for languages.

MIc: (Micro-Indicator Country) Holds the list of countries in columns and the value, extracted from the external sources for countries, successively for the *index*, *usage* and *traffic* inputs. For version 3 there is 786 lines.

Note that a pre-processing is required for *usage* in order to integrate the non-occidental social networks; this is done in **MIcU** sheet.

Note that a post-processing is required for *traffic* in order to perform the weighting with the optimal number of websites depending of the % of persons connected per country; this is done in **MIcT** and **MIcT1.**

The control columns are the following:
Col. A: indicates the type of indicator from a lookup of the name from **MATRIX**.
Col; B: indicates the name of the indicator

[44] Followed by « macro » if a macro-language.

Col. C: depending of the type of data computes the average or the total, or a matrix product with the number of connected persons per country of the inputs in each line
Col. D: indicates the type of data, either a world percentage per country or a quantity per country or a percentage within each country
Col. E: indicates if extrapolation is required and which of the two types of extrapolation if so
Col. F: computes the number of countries with source data
Col. H: hold the URL of the source except for *traffic* where it indicates the number of times the website has been cited, in order to allow a further corresponding weighting in **Wut**.

The control lines are the following:

Line 8 indicates for each country the number of websites which have been measured.
Line 9 indicates the ratio for the number of websites which should have been used in order to respect the proportionality of connected persons (the product line 8 by line 9 for each country represents the number exact of websites required for that country in the hypothesis of the actual total of websites (cell C7) . This will be used as a weighting factor to obtain a fair representation of the traffic measurements in **MIcT1** and **MIcT** prior to the weighting by the number of occurrence of websites done in **Wut** (this have been added in V3C to correct an error in V3 where the weighting as made in parallel with the demo-linguistic weighting).

MIcU: (Micro-Indicator Country Usage)The last added sheet included for the new V3 processing of *usage.* Includes a complete copy of the *usage* sources migrated from **MIc**, at the same lines, completed with T-Index and the list of new V3 social networks. For this new list, the traffic measurements per country obtained from SimilarWeb are set, followed by the process of extrapolation (see **EX**). The output is a new and final line called "Social networks weighted" which is obtained by weighting the full list of social networks with the corresponding total of subscribers, balancing finally in a fair manner the occidental social networks with the ones from the rest of the world.

Ma: (Mask absence) A sheet parallel at **MIc** holding 1 when a value is absent for the pair (country, input). Used for extrapolation.

Mp: (Mask presence) A sheet parallel at **MIc** holding 1 when a value exists for the pair (country, input). Used for extrapolation.

EX: (Extrapolation) A sheet parallel at **MIc** where the process of extrapolation is performed. Two different process are used depending of the type of data.

When the data is expressed as a world percentage per country the complement of 100% is split between the countries which have not received data in prorata of their world percentage of connected persons to the Internet. This is typically the case of the *traffic* measurement where the used tools, Alexa and SimilarWeb, do not cover all the countries.

When the data is a rating per country the technics of quartile is used where four quartile values are used depending of where the percentage of connected persons belongs in the interval between : 0%, 15%, 35%, 65%, 85% and 100%.This is typically the case of the *Index* data.

When it appears that either method could not provide meaningful extrapolation the source of data is not included.

In the rare cases where all the countries are informed by the data source obviously no extrapolation is required, as for the NapoleonCat data for percentage of subscribers per social network.

Note that for the V3 process of *usage*, the extrapolation for social networks is not performed in **EX** and has been replicated in **MIcU**.,

Note that the sum of T-Index values for the listed countries is 99.78%, so close to 100% that no extrapolation has been performed.

MIcT1: (Micro-Indicator Traffic1) The sheet is parallel at **MIc** and only is filled for the traffic indicators. Each cell (country, website) contains the product of source traffic from **MIc** added with the extrapolated traffic from **EX**, multiplied by the weighting factor for the country from MIP line 10. The sum of percentage is computed and placed in column G for further normalization to 100% in **MIcT**.

MIcT: (Micro-Indicator Traffic) The sheet is parallel at **MIcT1** and is used for normalization of the figures to 100% for each website by dividing each cell by the total. The results will be used in **Wut** to compute the final traffic repartition per language.

Wut: (Weight usage and traffic) In this sheet the *usage* and *traffic* indicators are processed. The process consists in weighting the values with the percentage of connected speakers per country from **PL**, after applying extrapolation. For the traffic indicator the extrapolation has already been performed in **MIcT** but there is an additional weighting to perform with the figures computed in **MIc** (column H) for the number of occurrences of each website.

Wi: (Weight index) In this sheet, the weighting with demo-linguistic figure in order to obtain figures per language is performed for the *index* indicator, starting in column BA, followed, starting in column 10, with the normalization to 100%.

Pi1: (Process indicator language) In this sheet, the weighting with demo-linguistic figures is performed, in order to obtain figures per language, for the indicators per country *usage* and *traffic*. For *usage*, an additional weighting is performed with the weight attributed to each component of this indicator (see 2.2.4). For traffic, an additional weighting is performed with the number of occurrences of each website in the sampling (see 2.2.5)

RES: (Results) The final results of each indicator per language (usage, traffic, index, interfaces) are computed.

FINAL: This sheet presents the final results with all the associated parameters and offers the results sorted for **contents**, **virtual presence**, **content productivity** and **connected speakers**. It also focuses the 20 first content positions and create the cyber-geography of language result (see table 1). A copy without formula of this sheet is made public as the product of the model (see https:/:funredes.org/lc/Results).

Note that a data base access of these results is scheduled before the end of 2022, with ISO 639-2 codes as the key for access.

MATRIX: The list of all the micro-indicators used in the model for each type (*index, interfaces, usages, traffic*).

3.4 RESULTS OF THE MODEL

The results of the model can be consulted in CC-BY-SA-4.0 in https://funredes.org/lc2022 and can be read in (Pimienta, 2022). Further releases of the model are accessible in https://funredes.org/lc/Results.

3.5 FURTHER CROSS-CHECKING

For the sake of cross-checking the results, the model has been run separately with L1 data only and wit L2 data only (see Annex 9 for the corresponding results which represent a quite positive indirect control of the method).

4 OTHER EXISTING METHODS BIASES

The observation of the presence of languages in the Internet has been quite active in the period 2000-2007 (Pimienta, 2009) but, after that period, as mentioned in introduction, only two options remained available for large public: InternetWorldStats and W3techs.

Both give some highlights of their respective methodology but no peer reviewed scientific paper has addressed their respective biases; their long-time presence without alternative figures have insured them a huge number of citations in diverse papers requiring those figures, too often without the necessary caution that would require the reality of their biases.

4.1 InternetWorldStats biases

The figures of IWS differ slightly from those of the Observatory, mainly because the sources of demo-linguistic data are not the same, and that, especially for L2 figures, the differences between sources could be huge (see note 24). Another difference exists however about the management of L2 figures. The Observatory computes the world language percentages for L1+L2 over the number of L1+L2 speakers, a figure 43% higher than the world population, following Ethnologue source[45], while IWS computes L1+L2 figures over the world population (calling it the zero-sum approach[46]). Unless there is a trick hidden somewhere in the computations, the zero-sum approach seems to provoke an error by overrating the 10 languages mentioned, error hidden in the remaining languages figure, which will become negative at one point of time if the number of languages is extended to the point where the sum of L1+L2 speakers cross the L1 value.

4.2 W3Techs biases

The method used by W3Techs is to apply a language recognition algorithm to the **home page** of 10 million websites which are selected by some Web traffic analysis service (Alexa.com or tranco-list.eu, until the end of 2022) as the most visited.

---

[45] In the 2021 figures, those we are using, Ethnologue counts the world population (total number of L1 speakers) at 7 231 699 136 and the total number of L1+L2 speakers at 10 361 716 756.

[46] Cited from the IWS website *: Indeed, many people are bilingual or multilingual, but here we assign only one language per person in order to have all the language totals add up to the total world population (zero-sum approach).*

The differences between W3Techs figures and Observatory's are huge often in a ratio 1 to 3, sometimes, as for Chinse and Hindi, in a ratio higher than 1 to 10). One of the two sources at least shall be extremely biased. The following table exposes those differences using W3Techs data of 24/8/22 and Observatory data of V3.1 in 8/2022.

Table 4: Comparison of figures W3Techs vs. Observatory

LANGUAGE	W3TECHS		OBSERVATORY	
	Rank	Web %[47]	Rank	Web %
English	1	61.4%	1	19.92%
Russian	2	5.6%	4	3,86%
Spanish	3	3.9%	3	8.09%
Turkish	4	3.2%	12	1.15%
German	5	3,1%	10	2.38%
French	6	3.0%	6	3.43%
Persian	7	2.7%	16	0.89%
Chinese	9	1.7%	2	19.82%
Arabic	13	1.1%	8	3.14%
Hindi	35	0.1%	5	3.67%

The highest differences are to be found in the value for Hindi and Chinese, and obviously the difference of weight of English contents (over 60% versus around 20%) raises concern. In august 2022, the statistics aggregator Statista[48], based on W3Techs figures, states "*English is the Internet's universal language*" while the Observatory, at the same time, states : "*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".* Again, those two statements are not compatible, one at least is wrong.

One could discuss the bias towards English of language recognition algorithms, the bias towards English of selecting the 10 million most visited websites[49]; but those are marginal biases which could not explain such huge differences. The main issue is in the **lack of consideration of multilingualism**, a characteristic of the Web which is ignored by the W3Techs method while the Web is probably still more multilingual than the humanity[50].

In background of that discussion it is important to remind the point stated, and documented in Annex 8, that Internet users prefer to use their mother tongue in the Net as first option and are eager to use their second language(s) in complement.

The problem starts with measuring **home pages** and counting a single language for each one. Many non-English websites may have English abstract or few English words in their home

[47] Note that W3Techs offers figure with only one digit after the point.
[48] https://www.statista.com/chart/26884/languages-on-the-internet/
[49] Following https://news.netcraft.com/archives/category/web-server-survey there are in May 2022 1.16 billion websites of which 270 million are active. The coverage of the most visited is then less than 4% of the total.
[50] It is so if the 270 million active websites offer together more than 400 million different linguistic interfaces, an average in the order of 1.5 by website.

pages and are probably counted as English. Many English websites have many others language' versions which shall be counted also (if, as it is probably the case, the algorithm is set in an English computed environment, the website is counted as English only).

The W3Techs would give quite different figures (and hopefully closer to Observatory's) if the following rules would be added to its algorithm:

- The counting is made on webpages not on websites.
- The algorithm checks the existence of language options in the home page and count each language offered as option.
- The algorithm checks the existence of other language than English in the home page, if this is the case count the website in that language instead of English.
- The algorithm evaluates an approximate number of pages in the website and multiply each language count by this number after dividing by the number of language options.

4 CONCLUSION

For the first time in Internet history, a method is able to offer a variety of meaningful indicators about the presence of 329 languages on the Internet. The model provides results coherent with previous studies made by the Observatory but are in strong contradiction with the results provided by the unique source which have been covering the subject since 2011; in particular it shows that the English contents in the Web are at the same level today that Chinese contents, around 20%, while the media keeps reporting English contents much above 50%.

The method used to obtain those results is completely and transparently exposed and its biases are openly discussed so that the scientific community could analyze.

Those results are simply reflecting a logical step of the evolution of the Web, which have evolved from a first phase English-centered (1992-2000), towards a second step centered in European languages, with English leadership (2000-2010) followed by a more internationalized, with the boom of Asian and Arabic languages and still an important gap leaving behind African languages, with a Web everyday more multilingual (2010-2020). The coming step (2020-2030) will probably see a Web more even in terms of representation of languages, with, hopefully, the digital divide starting to be overcome in Africa, opening the space of local languages of Africa. The rooting of multilingualism in the Web is underway and may be crossing above the one of humanity. Yet, differences in content productivity will prevail, with the maintain of some advantages for some languages with a combination of large L2 population and country coverage (such as English and French).

The surprise should not come for the observatory's figures which are just the reflect of the natural evolution of the world, reflected in its cyber component; they should come from the fact that strongly biased figures have been the rule of the last decade without much reaction of the scientific community.

Hopefully, the full transparency of the method will help more scientific minds to challenge results provided by the marketing world and let this theme where it should belong: the scientific community. Obviously, this includes the challenging of the method exposed here-before and the detection and discussion of possible biases which have not adverted by the authors. Lets the scientific approach prevails over marketing!

REFERENCES

Pimienta D., 2022, “Resource: Indicators on the Presence of Languages in Internet”, *Proc. of SIGUL2022, a workshop of LREC22*, Marseille, 6/2022 http://www.lrec-conf.org/proceedings/lrec2022/workshops/SIGUL/pdf/2022.sigul-1.11.pdf

Pimienta D., 2021, "Internet and linguistic diversity: the cyber-geography of languages with the largest number of speakers" in *LinguaPax Review 2021, Language Technologies and Language Diversity*, pp9. https://www.linguapax.org/wp-content/uploads/2022/02/LinguapaxReview9-2021-low.pdf

Pimienta D., 2021, “New and enhanced version of an alternative approach to measure the presence of languages in the Internet”, *Observatory of Linguistic and Cultural Diversity on the Internet*, 8/2021 https://funredes.org/lc2021/ALI%20V2-EN.pdf

Pimienta D., 2019, "Indicators of Languages in the Internet", in *Proc. of International Conference Language Technologies for All (LT4All)*, 4-6 December 2019, UNESCO, Paris; pp315, https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.79.pdf

Pimienta D., 2017, “An alternative approach to produce indicators of languages in the Internet”, *Proc.* of *Global Expert Meeting Multilingualism in Cyberspace for Inclusive Sustainable Development*, Khanty-Mansiysk, Russian Federation, June 4-9, 2017 http://funredes.org/lc2017/Alernative%20Languages%20Internet.docx

Pimienta D., 2016, "Medición de la presencia de la lengua española en la Internet: métodos y resultados", en *Revista Española de Documentación Científica 39(3)*, julio-septiembre 2016, e141. ISSN-L:0210-0614. - doi: http://dx.doi.org/10.3989/redc.2016.3.1328

Pimienta D., 2014, "Le français dans l'Internet", *Rapport 2014 "La langue française dans le monde"*, pp501, OIF, Nathan, http://www.francophonie.org/Rapports-Publications.html

Pimienta D., Prado D., Blanco Á, 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- http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=187016 (français, English)

ANNEX 1: USAGE INDICATOR SOURCES

Table 5: Social networks selected and total subscribers

SOCIAL NETWORK	TOTAL SUBSCRIBERS (Million)
Whatsapp	2000
Wechat	1225
Tiktok	732
Douyin	600
Telegram	600
QQ	595
Snapchat	528
Weibo	521
Qzone	517
Kuaishou	481
Quora	300
Skype	300
Tieba	300
Viber	260
IMO	200
LINE	169
picsart	150
Likee	150
Discord	140
Twitch	140
Stack Exchange	100
VK	650
Odnoklassniki	200
Douban	200
MOJ	160
JOSH	115
ShareChat	160
FACEBOOK %users per country (NapoleonCat 2021)	1455
INSTAGRAM %users per country (NapoleonCat 2021)	1200
MESSENGER %users per country (NapoleonCat 2021)	1300
LINKEDIN %users per country (NapoleonCat 2021)	155
FACEBOOK World% from IWS 2021	1455
Linkedin %user by country (ApolloTech 2021)	155
Twitter %users per country (Statista 2021)	396
Pinterest audience % (Statista 2021)	460
REDDIT % users per country (Statista 2021)	430

Table 6: Sources for social network figures

SOCIAL NETWORK FIGURES	SOURCE
FACEBOOK %users per country (NapoleonCat 2021)	https://napoleoncat.com/stats/
INSTAGRAM %users per country (NapoleonCat 2021)	https://napoleoncat.com/stats/
MESSENGER %users per country (NapoleonCat 2021)	https://napoleoncat.com/stats/
LINKEDIN %users per country (NapoleonCat 2021)	https://napoleoncat.com/stats/
Linkedin %user by country (ApolloTech 2021)	https://www.apollotechnical.com/linkedin-users-by-country/
Twitter %users per country (Statista 2021)	https://www.statista.com/statistics/242606/ number-of-active-twitter-users-in-selected-countries/
FACEBOOK World% from IWS 2021	https://www.internetworldstats.com/stats1.htm + stats2.htm+…stats6.htm
Facebook audience % (Statista 2021)	https://www.statista.com/statistics/268136/ top-15-countries-based-on-number-of-facebook-users/
YouTube % of connected within country (Statista 2021)	https://www.statista.com/statistics/1219589/youtube-penetration-worldwide-by-country/
Netflix % subscribers per country (CompariTech 2020)	https://www.comparitech.com/tv-streaming/netflix-subscribers/
Pinterest audience % (Statista 2021)	https://www.statista.com/statistics/328106/pinterest-penetration-markets/
REDDIT % users per country (Statista 2021)	https://backlinko.com/reddit-users
Cumul. 2012/21 % OpenOffice downloads per country	http://www.openoffice.org/stats/countries.html
# Secure Internet servers	https://data.worldbank.org/indicator/IT.NET.SECR
% Fixed broadband subscr. within country (WB 2021)	https://data.worldbank.org/indicator/IT.NET.BBND.P2
% Fixed Tel.+mobile subscr. within country (WB 2021)	https://data.worldbank.org/indicator/IT.MLT.MAIN.P2 + https://data.worldbank.org/indicator/IT.CEL.SETS.P2

ANNEX 2: ONLINE ENCYCLOPEDIAS ANALYZED FOR CONTENT INDICATOR

Table 7: Online encyclopedias

LANGUAGE	ENCYCLOPEDIA	NUMBER OF ARTICLES (Millions)	OTHER INFORMATION
Various	Encyclopedia of Life (EOL)	0.75 (2010) 1.9 today	Supported languages: Arabic, Brazilian Portuguese, English, Finnish, French, Macedonian, Piedmontese, Traditional Chinese and Turkish Interface languages: the same plus German, Spanish, Dutch, Turkish and Ukrainian.
Various	thefreedictionary.com/ Free with ads or paid	no statistics	English, Spanish, German, French, Italian, Chinese, Portuguese, Dutch, Norwegian, Greek, Arabic, Polish, Turkish, Russian, Hebrew It is not clear if they are parallel version or language specific.
Various	en.metapedia.org/ neonazi version of wikipedia	Marginal (5000 articles in English)	Czech, Danish, German, Spanish, English, Hungarian, Dutch, Portuguese, Romanian, Slovenian, Swedish, Estonian, Croatian, Icelandic, Norwegian, Macedonian
Chinese	Baidu Baike	24.5	194 million edits 7.5 million publishers
Chinese	Baike (Hudong)	18	5.8 million publishers (2013)
Chinese	SogouBaike	????[51]	
Arab	Marefa	0.136636	2.4 million pages
Arab	Mawdoo3	0.15	45 (2018)
Bengali and English	bengaldia	0.0057	1450 publishers
Croatian	enciklopedija.hr	0.067	Print version data
Croatian	proleksis.lzmk.hr	0.062	
Danish	Den Store Danske	0.161	1100 publishers 1 million users
Dutch	winklerprins.com	0.0115	by subscription
English	britannica.com		limited free access
English	Everipedia Articles copied from wikipedia	?	7000 active publishers (2019) 3M users (2017) free access but also blockchain market
English	Citizenship	0.017	Statistics stopped in 2014 close to stopping
English	Conservapedia	0.0518	800 million pageviews 1.5 million edits
English	Scholarpedia	0.0018	Marginal digits
English	Encyclopedia.com	0.3	Formal encyclopedia aggregator
English	Colombia Encyclopedia		Aggregated by Encyclopedia.com
English	digitaluniverse.net		offline
French	Larousse	0.317	
German	retro bib	0.3	
Hebrew English	Hamichlol	0.28	Censored version of Wikipedia for a hyper-religious audience
Korean	Doopedia	0.588	

[51]Sogou Baike is considered at least as important as Baidu Baké and the same value of number of articles has been assumed.

Malay Sundanese Javanese	Superpedia	0.02	
Italian	Treccani	0.9	
Malayalam	Sarvavijnanakosam	0.007	
Marathi	Viswakosh	0.016	
Norwegian Bokmal and Nynorsk	Store norske leksikon	0.2 (2019)	3M users/month read 500K articles
Polish	encyclopedia.interia.pl	0.12 (2006)	
Polish	encyclopedia.pwn.pl	0.08	
Russian	Great Russian Encyclopedia	0.012 (2016)	
Russian	Krugosvet	0.012	
Spanish	https://www.ecured.cu/ Cuban	0.237	73,000 active 537 publishers
Spanish	Encyclonet	0.185	
Spanish	enciclopedia.us.es/	0.053	https://wikiapiary.com/wiki/
Swedish	ne.se/	0.26 (2005)	
Tamil	not online		
Turkish	Ekşi Sozluk	8M admissions in 2009[52]	400,000 users 110,000 publishers 4M new admissions/year in 2013[53] Open for publication, each entry is kept after moderation.
Vietnamese	seems to have disappeared		Go to archive.org - https://bachkhoatoanthu.vass.gov.vn

[52]https://www.researchgate.net/publication/242100750_Web_Based_Authorship_in_the_Context_of_User_Generated_Content_An_Analysis_of_a_Turkish_Web_Site_Eksi_Sozluk

[53]https://www.researchgate.net/publication/271521393_SOCIAL_MEDIA_IN_A_DICTIONARY_FORMAT_ONLINE_COMMUNITY_OF_eksisozlukcom/figures?lo=1

ANNEX 3: INTERFACE INDICATOR SOURCES

Table 8: Sources for interface indicator

Translation languages of Bing Translator	https://www.bing.com/translator/
Amazon Kindle direct Publishing supported languages	https://kdp.amazon.com/en_US/help/topic/G200673300
Languages supported by Cortana	https://en.wikipedia.org/wiki/Cortana
Word Reference languages supported	https://www.wordreference.com
WordLingo Translation languages	http://www.worldlingo.com/en/languages/
Facebook supported languages	https://www.facebook.com/language.php
Facebook In-Stream Ads languages supported	https://www.facebook.com/business/help/267128784014981
Free Translator languages supported	http://www.free-translator.com
Google Play Console supported languages	https://support.google.com/googleplay/android-developer/table/4419860?hl=en
Google Cloud supported languages	https://cloud.google.com/translate/docs/languages?hl=en
Google Translate supported languages	https://en.wikipedia.org/wiki/Google_Translate
Google Scholar supported languages for search	https://scholar.google.com/scholar_settings?sciifh=1&hl=en&as_sdt=0,5#1
Language supported by Paralink Translator	http://paralink.com
Online Translator languages supported	https://www.online-translator.com/traduction
Reverso translator languages supported	https://www.reverso.net/text_translation.aspx?lang=EN
Free Translation supported languages	https://www.freetranslations.org
Skype Supported languages	https://support.skype.com/en/faq/FA34781/what-languages-are-supported-in-skype
Systran translate supported languages	https://support.systran.net/systranlinks/faq/

ANNEX 4: INDEX INDICATOR SOURCES

Table 9: Sources for index indicator

E-Government Index	https://publicadministration.un.org/egovkb/Data-Center
E-Participation Index	https://publicadministration.un.org/egovkb/Data-Center
Online Service Index	https://publicadministration.un.org/egovkb/Data-Center
Human Capital Index	https://publicadministration.un.org/egovkb/Data-Center
Telecommunication Infrastructure Index	https://publicadministration.un.org/egovkb/Data-Center
Cisco Global Digital Readiness Index 2019	https://www.cisco.com/c/dam/en_us/about/csr/reports/global-digital-readiness-index.pdf
Government AI Readiness Index 2020	https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf
Internet Freedom Scores	https://freedomhouse.org/countries/freedom-net/scores
Global Connectivity Index	https://www.huawei.com/minisite/gci/en/country-rankings.html
Global Cybersecurity Index 2018	https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf
UNCTAD B2C E-commerce index, 2020	https://unctad.org/system/files/official-document/tn_unctad_ict4d17_en.pdf
The Global Open Data Index	https://index.okfn.org/place/
World Digital Competitiveness Ranking 2020	https://www.imd.org/globalassets/wcc/docs/release-2020/digital/digital_2020.pdf
Readiness For Frontier Technologies Index	https://unctad.org/system/files/official-document/tir2020_en.pdf
Global Innovation Index	https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2020.pdf
Access to Basic Knowledge	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Access to Information and Communications	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Access to Advanced Education	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Access to electricity (% of pop.)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Access to quality education (0=unequal; 4=equal)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Access to online governance (0=low; 1=high)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Media censorship (0=frequent; 4=rare)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Freedom of expression (0=no freedom; 1=full freedom)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Quality weighted universities (points)	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Citable documents	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Women with advanced education	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Years of tertiary schooling	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx

ANNEX 5: TRAFFIC INDICATOR WEBSITES SELECTION

Table 10: Selection of websites for traffic indicator

WEBSITE	NUMBER OF TIMES
10086.cn	1
10jqka.com.cn	1
122.gov.cn	1
12306.cn	1
12371.cn	1
12377.cn	1
12388.gov.cn	1
1688.com	1
17ok.com	1
189.cn	1
1c-bitrix.ru	1
22.cn	1
24.kg	1
24h.com.vn	1
2gis.ru	1
300.cn	1
360.cn	1
4.cn	1
6.cn	1
66law.cn	1
999.md	1
abc.com.py	1
abc-communication.dz	1
abril.com.br	1
accuweather.com	2
activemind.de	1
actuniger.com	1
ad.iq	1
admin.ch	1
adminbuy.cn	1
Adobe.com	1
afrikmag.com	1
agenziaentrate.gov.it	1
ah.gov.cn	1
ahoraeg.com	1
ahram.org.eg	1
aktuality.sk	1
alakhbar.info	1
aliexpress.com	1
alipay.com	1
allegro.pl	1
allevents.in	1
almasryalyoum.com	1
alraziuni.edu.ye	1
alwakeelnews.com	1
alwatanvoice.com	1
amazon.ae	1
amazon.ca	1
amazon.cn	1
amazon.co.jp	1
amazon.co.uk	1
amazon.com	20
amazon.com.br	1
amazon.de	1
amazon.eg	1
amazon.es	1
amazon.fr	1
amazon.in	1
amazon.it	1
ameblo.jp	1
amritmahotsav.nic.in	1
andersnoren.se	1
anpc.gov.ro	1
anyxxx.com	1
ap.gov.in	1
aparat.com	2
apple.com	5
arabiaweather.com	1
argentina.gob.ar	1
aruba.it	1
autohome.com.cn	1
avaz.ba	1
babytree.com	1
baharain.bh	1
baidu.com	4
band.us	1
bangladesh.gov.bd	1
bankmandiri.co.id	1
bayern.de	1
bb.com.br	1
bbc.co.uk	1

bbc.com	1
bbc.in	1
bcel.com.la	1
beian.gov.cn	1
beijing.gov.cn	1
belgium.be	1
belizebank.com	1
belonnanotservice.ga	2
bet365.com	1
bgeneral.com	1
bih.nic.in	1
Bing.com	3
biobiochile.cl	1
bitrix24.ru	1
bjx.com.cn	1
blogger.com	3
bnonline.fi.cr	1
boc.cn	1
Bongacams.com	3
borneobulletin.com.bn	1
bri.co.id	1
britannica.com	2
bshare.cn	1
bt.bt	1
bt.cn	1
bukalapak.com	1
bund.de	1
businessday.ng	1
businessinsider.in	1
businesstoday.in	1
businessworld.in	1
cac.gov.cn	1
cafebazaar.ir	1
caixa.gov.br	1
cambridge.org	1
Canva.com	5
cao.ir	1
careers.sl	1
cas.cn	1
cbec.gov.in	1
cbic.gov.in	1
cbos.gov.sd	1
cbse.gov.in	1
cbse.nic.in	1
ccdi.gov.cn	1
ccgp.gov.cn	1
ccm.gov.cn	1
ce.cn	1
centrafrique-presse.over-blog.com	1
chase.com	1
chaturbate.com	2
china.cn	1
china.com.cn	1
chinadaily.com.cn	1
chinanews.com.cn	1
chinatax.gov.cn	1
chsi.com.cn	1
cib.com.cn	1
cmbc.com.cn	1
cmseasy.cn	1
cnil.fr	1
cninfo.com.cn	1
cnipa.gov.cn	1
cnnindonesia.com	1
cnpq.br	1
cnr.cn	1
cntv.cn	1
coinmarketcap.com	2
comores-infos.net	1
conac.cn	1
consultant.ru	1
coremail.cn	1
correios.com.br	1
corriere.it	1
coupa.ng	1
court.gov.cn	1
covid19.go.id	1
cowin.gov.in	1
cpdp.bg	1
cq.gov.cn	1
creditchina.gov.cn	1
cri.cn	1
cricbuzz.com	1
cro.ma	1
csdn.net	1
csrc.gov.cn	1
customs.gov.cn	1
cvc.nic.in	1
cyberpolice.cn	1
dahe.cn	1

dailypost.ng	1
dakaractu.com	1
daraz.pk	1
data.gov.in	1
dataprotection.gov.cy	1
daum.net	1
defimedia.info	1
detik.com	1
dg-datenschutz.de	1
dictionary.com	1
digikala.com	1
digitalindia.gov.in	1
dinesh-ghimire.com.np	1
discord.com	1
ditaduraeconsenso.blogspot.com	1
dlszywz.cn	1
dns4.cn	1
docdro.id	1
dpboss.net	1
dr.dk	1
draugiem.lv	1
duckduckgo.com	1
dwz.cn	1
e.gov.kw	1
ebay.com	1
ebay.de	1
ebs.org.cn	1
eci.gov.in	1
education.gov.in	1
elcomercio.com	1
eldeber.com.bo	1
elnuevodia.com	1
elpais.com	1
elsalvador.com	1
eluniverso.com	1
emansion.gov.lr	1
emploi.cg	1
ems.com.cn	1
enamad.ir	1
enimerotiko.gr	1
eol.cn	1
e-recht24.de	1
ernet.in	1
espn.com	1
espncricinfo.com	1

estadao.com.br	1
eta.gov.lk	1
ethiojobs.net	1
etnet.com.hk	1
facebook.com	80
facebook.com.br	1
fandom.com	3
fazenda.gov.br	1
fijivillage.com	1
file-upload.com	1
findlaw.cn	1
firefox.com.cn	1
fiverr.com	1
flipkart.com	1
flydubai.com	1
fmprc.gov.cn	1
focus.cn	1
follow.it	1
Force.com	1
free.fr	1
freebitco.in	1
freeindianporn2.com	1
freepik.com	1
fs.fed.us	1
ftc.go.kr	1
fujian.gov.cn	1
gansu.gov.cn	1
garanteprivacy.it	1
gd.gov.cn	1
geni.us	1
gesetze-im-internet.de	1
ghanaweb.com	1
gismeteo.ru	1
globo.com	1
gmw.cn	1
gogo.mn	1
gome.com.cn	1
goo.ne.jp	1
google,com	1
google.ad	1
google.ae	1
google.at	1
google.az	1
google.be	1
google.bf	1

google.bg	1
google.ca	2
google.cd	1
google.cg	1
google.ch	1
google.ci	1
google.cl	1
google.cn	1
google.co.id	1
google.co.il	1
google.co.in	1
google.co.jp	1
google.co.ke	1
google.co.kr	1
google.co.ma	1
google.co.mz	1
google.co.nz	1
google.co.th	1
google.co.tz	1
google.co.ug	1
google.co.uk	1
google.co.uz	1
google.co.ve	1
google.co.za	1
google.co.zm	1
google.co.zw	1
google.com	146
google.com.af	1
google.com.ar	1
google.com.bd	1
google.com.bn	1
google.com.bo	1
google.com.br	1
google.com.bz	1
google.com.co	1
google.com.cu	1
google.com.do	1
google.com.eg	1
google.com.hk	2
google.com.jm	1
google.com.kw	1
google.com.lb	1
google.com.ly	1
google.com.mm	1
google.com.mt	1

google.com.mx	1
google.com.na	1
google.com.ng	1
google.com.ni	1
google.com.np	1
google.com.om	1
google.com.pa	1
google.com.pe	1
google.com.pg	1
google.com.ph	1
google.com.pk	1
google.com.pr	1
google.com.py	1
google.com.qa	1
google.com.sa	1
google.com.sb	1
google.com.sg	1
google.com.sl	1
google.com.sv	1
google.com.tj	1
google.com.tr	1
google.com.tw	1
google.com.ua	1
google.com.uy	1
google.com.vn	1
google.de	1
google.dj	1
google.dk	1
google.dz	1
google.ee	1
google.es	2
google.fr	3
google.ge	1
google.gr	1
google.gy	1
google.hn	1
google.ht	1
google.ie	1
google.iq	1
google.is	1
google.it	1
google.jo	1
google.kg	1
google.kz	1
google.la	1

google.lk	1
google.lt	1
google.lu	1
google.lv	1
google.md	1
google.me	1
google.mg	1
google.mk	1
google.ml	1
google.mn	1
google.mw	2
google.nl	1
google.no	1
google.pl	1
google.ps	1
google.pt	1
google.ro	1
google.rs	1
Google.ru	3
google.rw	1
google.se	1
google.si	1
google.sk	1
google.sn	1
google.so	1
google.sr	1
google.st	1
google.td	1
google.tg	1
google.tl	1
google.tm	1
google.tn	1
google.tt	1
gosuslugi.ru	1
gov.bw	1
gov.ls	1
govtrack.us	1
grid.id	1
grupobancolombia.com	1
gst.gov.in	1
gsxt.gov.cn	1
guardian.co.tt	1
guardian.ng	1
gujarat.gov.in	1
gxzf.gov.cn	1

gz.gov.cn	1
haberler.com	1
hainan.gov.cn	1
haosou.com	1
hatena.ne.jp	1
hd315.gov.cn	1
hdfcbank.com	1
healthline.com	1
heartland.us	1
henan.gov.cn	1
herald.co.zw	1
hi.is	1
hindustantimes.com	1
homedepot.com	1
hoster.kz	1
hotlog.ru	1
hotpepper.jp	1
hotstar.com	2
huanqiu.com	1
hubei.gov.cn	1
hunan.gov.cn	1
hurriyet.com.tr	1
ibps.in	1
ibw.cn	1
icbc.com.cn	1
icicibank.com	1
icio.us	1
ico.org.uk	1
idnes.cz	1
iitb.ac.in	1
iitkgp.ac.in	1
ijavhd.com	1
imageshack.us	1
imdb.com	2
imjo.in	1
in.gr	1
incometax.gov.in	1
incometaxindia.gov.in	1
incometaxindiaefiling.gov.in	1
index.hr	1
index.hu	1
india.com	1
india.gov.in	1
indiamart.com	1
indianrailways.gov.in	1

indianvisaonline.gov.in	1
indiapost.gov.in	1
indiatimes.com	1
indiatoday.in	1
inflibnet.ac.in	1
instagram.com	47
instructure.com	1
intoday.in	1
iol.co.za	1
ionos.de	1
iplt20.com	1
irctc.co.in	1
irembo.gov.rw	1
irna.ir	1
is.fi	1
isna.ir	1
itau.com.br	1
jamaica-gleaner.com	1
japanpost.jp	1
jc001.cn	1
Jd.com	1
jiangsu.gov.cn	1
jiangxi.gov.cn	1
jiji.ng	1
jl.gov.cn	1
jne.co.id	1
jotform.us	1
jrj.com.cn	1
jumia.ci	1
jumia.com.ng	1
juraforum.de	1
justindianporn.me	1
kancloud.cn	1
kar.nic.in	1
karnataka.gov.in	1
kaskus.co.id	1
kemdikbud.go.id	1
kemenag.go.id	1
kemkes.go.id	1
kenh14.vn	1
kerala.gov.in	1
khaberni.com	1
knet.cn	1
knetreg.cn	1
kominfo.go.id	1

kompas.com	1
kriesi.at	1
kuaishang.cn	1
kuenselonline.com	1
kumparan.com	1
kupujemprodajem.com	1
lanouvelletribune.info	1
laodong.vn	1
laprensa.com.ni	1
laprensa.hn	1
lawtime.cn	1
lazada.co.id	1
leader.ir	1
lefigaro.fr	1
legifrance.gouv.fr	1
legit.ng	1
lemonde.fr	1
lex.uz	1
licindia.in	1
line.me	2
linkd.in	1
linkedin.com	13
liputan6.com	1
list.am	1
listindiario.com	1
live.com	19
liveinternet.ru	1
livroreclamacoes.pt	1
lnkd.in	1
ltn.com.tw	1
ltt.ly	1
m.in	1
macaodaily.com	1
mahaonline.gov.in	1
maharashtra.gov.in	1
mail.ru	2
mana.pf	1
mastercard.us	1
mayoclinic.org	2
medcol.mw	1
mediacongo.net	1
mercadolibre.cl	1
mercadolibre.com.co	1
mercadolibre.com.ve	1
mercadolivre.com.br	1

merdeka.com	1
merriam-webster.com	1
meskerem.net	1
meteo.nc	1
metruyenchu.com	1
mhlw.go.jp	1
microsoft.com	25
microsoftonline.com	4
milliyet.com.tr	1
mk.by	1
mof.gov.tl	1
moh.go.tz	1
moip.gov.mm	1
mol.gov.om	1
monetizze.com.br	1
msn.com	3
myshopify.com	5
namibian.com.na	1
namnak.com	1
naver.com	1
ncdc.gov.ng	1
nessma.tv	1
netafrique.net	1
netflix.com	13
nethouse.ru	1
nettruyengo.com	1
news24.com	1
niagahoster.co.id	1
notion.so	1
novinky.cz	1
nsw.gov.au	1
nzherald.co.nz	1
odnoklassniki.ru	1
office.com	8
ok.ru	3
okezone.com	1
onlinehome.us	1
orange.fr	1
orient.tm	1
otr.tg	1
ouest-france.fr	1
oxu.az	1
ozon.ru	1
pagcor.ph	1
pagesjaunes.fr	1
paypal.com	1
paystack.com	1
pikiran-rakyat.com	1
pinterest.com	11
pinterest.de	1
pinterest.es	1
pinterest.fr	1
pinterest.it	1
pixnet.net	1
planalto.gov.br	1
pornhub.com	4
portaldoconhecimento.gov.cv	1
post.ir	1
postcourier.com.pg	1
postimees.ee	1
premierbet.co.ao	1
premierleague.com	1
prensa-latina.cu	1
prensalibre.com	1
presidence.gov.bi	1
president.ir	1
president.tj	1
prom.st	1
prom.ua	1
public.lu	1
pulse.ng	1
punchng.com	1
qq.com	2
r01.ru	1
rae.es	1
rakuten.co.jp	1
rambler.ru	1
reddit.com	9
reg.ru	1
repubblica.it	1
republika.co.id	1
ria.ru	1
rijksoverheid.nl	1
rt.com	1
rte.ie	1
rtvslo.si	1
s.id	1
sabay.com.kh	1
sacoronavirus.co.za	1
sahibinden.com	1

sakura.ne.jp	1
salesforce.com	1
salla.sa	1
sana.sy	1
sante.gov.dz	1
sante.gov.gn	1
sapo.pt	1
sapp.ir	1
saude.gov.br	1
scielo.br	1
sekolahku.web.id	1
seneweb.com	1
serveriai.lt	1
service-public.fr	1
setn.com	1
seznam.cz	1
shopee.co.id	1
shopee.co.th	1
shopee.tw	1
shopee.vn	1
shop-pro.jp	1
singaporepools.com.sg	1
smarturl.it	1
sohu.com	2
solomonstarnews.com	1
soy502.com	1
spiegel.de	1
stackoverflow.com	1
standardmedia.co.ke	1
state.co.us	1
state.fl.us	1
state.il.us	1
state.ma.us	1
state.md.us	1
state.mn.us	1
state.nj.us	1
state.nm.us	1
state.nv.us	1
state.ny.us	1
state.oh.us	1
state.or.us	1
state.pa.us	1
state.tx.us	1
suara.com	1
sucursalelectronica.com	1
suribet.sr	1
sympla.com.br	1
syri.net	1
t.me	2
taobao.com	2
theguardian.com	1
thethao247.vn	1
tiktok.com	10
time.mk	1
times.co.sz	1
timesofmalta.com	1
timeweb.ru	1
tmall.com	1
tokopedia.com	1
t-online.de	1
tradingview.com	2
trendyol.com	1
tribunnews.com	1
tripadvisor.com.br	1
tripadvisor.fr	1
tripadvisor.it	1
turkiye.gov.tr	1
twitch.tv	5
twitter.com	32
ucoz.ru	1
uem.mz	1
ultimahora.com	1
uol.com.br	1
ura.go.ug	1
usp.br	1
vanguardngr.com	1
vg.no	1
vk.com	7
vkontakte.ru	1
vnexpress.net	1
walmart.com	1
wbs-law.de	1
weather.com	1
webmd.com	1
whatsapp.com	22
wikipedia.org	29
wiktionary.org	2
wildberries.ru	1
wizard.id	1
www.gob.mx	1

www.gob.pe	1
www.gov.br	1
www.gov.pl	1
www.gov.uk	1
xhamster.com	1
xnxx.com	3
xosodaiphat.com	1
xvideos.com	6
yahoo.co.jp	1
yahoo.com	25
yandex.ru	5

yasour.org	1
yelp.com	1
ynet.co.il	1
youm7.com	1
youtu.be	1
youtube.com	103
youtube.org	1
zalo.me	1
zambiaimmigration.gov.zm	1
zhizhuchi.cm	1
zoom.us	15

ANNEX 6: MACROLANGUAGES

As defined by Ethnologue.

Table 11: List of macro-languages

ISO CODE	*MACRO LANGUAGES*	*NUMBER OF LANGUAGES MERGED*
ara	**Arabic**	***29***
aym	**Aymara**	***2***
aze	**Azerbaijani**	***3***
bal	**Baluchi**	***3***
bik	**Bikol**	***8***
bnc	**Bontok**	***5***
bua	**Buryat**	***3***
chm	**Mari**	***2***
cre	**Cree**	***6***
del	**Delaware**	***2***
den	**Slave**	***2***
din	**Dinka**	***5***
doi	**Dogri**	***2***
est	**Estonian**	***2***
fas	**Persian**	***2***
ful	**Fulfulde**	***9***
gba	**Gbaya**	***6***
gon	**Gondi**	***3***
grb	**Grebo**	***5***
grn	**Guarani**	***5***
hai	**Haida**	***2***
hbs	**Serbo-croatian**	***4***
hmn	**Hmong**	***25***
iku	**Inuktitut**	***2***
ipk	**Inupiatun**	***2***
jrb	**Judeo-arabic**	***5***
kau	**Kanuri**	***3***
kln	**Kalenjin**	***9***
kok	**Konkani**	***2***
kom	**Komis**	***2***
kon	**Congo**	***3***
kpe	**Kpell**	**2**
kur	**Kurdish**	***3***
lah	**Lahnda**	***7***
lav	**Latvian**	***2***
luy	**Luiya**	***14***
man	**Mandingo**	***6***
mlg	**Malagasy**	***11***
mon	**Mongolian**	***3***
msa	**Malay**	***36***
mwr	**Marwari**	***6***
nep	**Nepalese**	***2***
oji	**Ojibway**	***7***
ori	**Oriya**	***2***
orm	**Galla**	***4***
pus	**Pashto**	***3***
que	**Quechua**	***42***
raj	**Rajasthan**	**6**
rom	**Romani**	***6***
sqi	**Albanian**	***4***
srd	**Sardinian**	***4***
swa	**Swahili**	***2***
syr	**Syriac**	***2***
tmh	**Tamashek**	***4***
uzb	**Uzbek**	***2***
yid	**Yiddish**	***2***
zap	**Zapotec**	***57***
zha	**Zhuang**	***16***
zho	**Chinese**	***15***
zza	**Dimli**	***2***

ANNEX 7: LIST OF COUNTRIES OR TERRITORIES WITH NO ITU DATA

Table 12: List of countries with no ITU data

ISO code	NAME OF THE COUNTRY	POPULATION
AX	**Åland Island**	**27,652**
AS	**American Samoa**	**55,990**
IO	**British Indian Ocean Territory**	**4,000**
QB	**Caribbean Netherlands**	**18,740**
CX	**Christmas Island**	**1,170**
CC	**Cocos (Keeling) Islands**	**630**
CK	**Cook Islands**	**15,000**
CW	**Curacao**	**140,000**
GF	**French Guiana**	**366,590**
GP	**Guadeloupe**	**454,800**
GU	**Guam**	**139,550**
IM	**Isle of man**	**88,085**
QM	**Martinique**	**377 100**
NC	**Norfolk Island**	**1,500**
KP	***North Korea***	***25,579,000***
PM	**Northern Mariana Islands**	**53,280**
PW	**Palau**	**17,550**
PN	**Pitcairn**	**36**
RE	**Réunion**	**751,580**
BL	**Saint Barthélemy**	**7,850**
FM	**Saint-Martin**	**28,500**
PM	**Saint Pierre and Miquelon**	**6,340**
SX	**Saint-Martin**	**33,470**
CT	**Turks and Caicos Islands**	**30 170**
GO	***Vatican State***	***330***
HE	***Western Sahara***	***544 150***
	TOTAL	**28,689,463**

There are two possible reasons why the country or territory is excluded from ITU data:

1) It is a territory whose data is included in a given country
2) There is no source or estimate of the percentage of people connected to the Internet (in italics in the table).

ANNEX 8 : SOURCES ABOUT LANGUAGE BEHAVIOR OF INTERNAUTS

https://motsdici.be/wp-content/uploads/2019/04/Article-cant-read-wont-buy.pdf
2006 Common Sense Advisory Report

https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556
2011 Union European Survey Report "*Digital Agenda: more than half EU Internet surfers use foreign language when online*"
Citation : *"While 90% of Internet surfers in the EU prefer to access websites in their own language, 55% at least occasionally use a language other than their own when online according to a pan-EU Eurobarometer".*

https://hbr.org/2012/08/speak-to-global-customers-in-t
2012 Harvard Business Review*: "Speak to Global Customers in Their Own Language"*
Citation: *"72.1% of consumers spend most or all of their time on websites in their own language"*

https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf
2017 KPMG/Google study "*Indian languages – Defining India's Internet*"
Citation : "*Indian languages Internet users are expected to account for nearly 75% of Indias's Internet user base in 2021*."

https://insights.csa-research.com/reportaction/305013126/Marketing
https://csa-research.com/Blogs-Events/CSA-in-the-Media/Press-Releases/Consumers-Prefer-their-Own-Language
2020 CSA Research Report *"Can't Read, Won't Buy – B2C Analyzing Consumer Language Preferences and Behaviors in 29 Countries"*
Citation: *Survey of 8,709 Consumers in 29 Countries Finds that 76% Prefer Purchasing Products with Information in their Own Language*

https://octopustranslations.com/e-commerce-and-the-impact-of-language-on-consumer-behavior/
2021 Octopus Translation report *E-commerce and the Impact of Language on Consumer Behavior*
Citation: *55% of consumers around the world make their purchases online only in their mother tongue*

https://www.businesswire.com/news/home/20211026005375/en/Unbabel%E2%80%99s-2021-Global-Multilingual-CX-Survey-Reveals-68-of-Consumers-Prefer-to-Speak-with-Brands-in-Their-Native-Language
2021 BusinessWire Report "*Unbabel's 2021 Global Multilingual CX Survey Reveals 68% of Consumers Prefer to Speak with Brands in Their Native Language"*

https://www.prweb.com/releases/2014/04/prweb11725995.htm
2022. PRWeb Market Research *"Survey of 3,000 Online Shoppers across 10 Countries Finds that 60% Rarely or Never Buy from English-only Websites"*

ANNEX 9: SEPARATE MODEL RUN FOR L1 AND L2

As a method for cross-checking the validity of the model, which is based on L1+L2 demo-linguistic data, two additional runs were made, one with L1 figures only and another one with L2 figures only.

Table 13: Model run with L1 only

		Internet users	Population	Speakers	Contents	Presence	Productivity
L1		L1	L1	Connected	L1	Virtual	Contents
1	Chinese	22.34%	18.33%	71.18%	**25.55%**	**1.39**	**1.14**
2	English	7.82%	5.12%	89.24%	**12.96%**	**2.53**	**1.66**
3	Spanish	8.14%	6.52%	72.95%	**8.76%**	**1.34**	**1.08**
4	Arab	5.33%	4.80%	64.91%	**4.15%**	**0.86**	**0.78**
5	Portuguese	3.91%	3.21%	70.99%	**3.91%**	**1.22**	**1.00**
6	Japanese	2.77%	1.75%	92.63%	**3.47%**	**1.99**	**1.25**
7	Russian	3.00%	2.13%	82.36%	**3.22%**	**1.51**	**1.07**
8	Hindi	3.35%	4.73%	41.34%	**2.93%**	**0.62**	**0.88**
9	French	1.59%	1.10%	84.59%	**2.08%**	**1.89**	**1.31**
10	German	1.62%	1.06%	89.51%	**1.96%**	**1.85**	**1.21**

If we only consider first language speakers, French would be in position 9 and quite logically Chinese will show a big advantage over English, despite its very large virtual presence and content productivity. The virtual presence and content productivity for French are very high, despite this ninth place.

Table 14 : Model run with L2 only

		Internet users	Population	Speakers	Contents	Presence	Productivity
L2		L2	L2	Connected	L2	Virtual	Contents
1	English	32.53%	31.25%	55.64%	**37.91%**	**1.21**	**1.17**
2	Chinese	8.68%	6.38%	72.65%	**10.68%**	**1.67**	**1.23**
3	French	6.47%	5.99%	57.81%	**6.90%**	**1.15**	**1.07**
4	Hindi	6.32%	8.25%	40.93%	**5.96%**	**0.72**	**0.94**
5	Spanish	3.37%	2.28%	78.82%	**5.47%**	**2.39**	**1.62**
6	Russian	4.82%	3.33%	77.32%	**5.12%**	**1.54**	**1.06**
7	Malay	5.37%	5.21%	55.08%	**4.52%**	**0.87**	**0.84**
8	German	3.10%	1.87%	88.72%	**3.61%**	**1.93**	**1.17**
9	Thai	1.86%	1.28%	77.84%	**1.55%**	**1.21**	**0.83**
10	Urdu	1.81%	5.15%	18.86%	**1.15%**	**0.22**	**0.63**
11	Portuguese	0.68%	0.81%	44.81%	**0.89%**	**1.10**	**1.32**

If we only consider second language speakers, English logically takes a big lead in first place and French takes third place above Spanish.

As a reminder, here are the results for L1+L2.

Table 15: Model results for L1+L2

L1		Internet users	Population	Speakers	Contents	Presence	Productivity
L2		L1+L2	L1+L2	Connected	L1+L2	Virtual	Contents
1	Chinese	18.46%	14.72%	71.38%	**21.60%**	**1.47**	**1.17**
2	English	14.83%	13.01%	64.86%	**19.60%**	**1.51**	**1.32**
3	Spanish	6.79%	5.24%	73.72%	**7.85%**	**1.50**	**1.16**
4	Hindi	4.19%	5.80%	41.16%	**3.76%**	**0.65**	**0.90**
5	Russian	3.51%	2.49%	80.32%	**3.76%**	**1.51**	**1.07**
6	French	2.98%	2.58%	65.80%	**3.33%**	**1.29**	**1.12**
7	Portuguese	2.99%	2.49%	68.43%	**3.13%**	**1.26**	**1.05**
8	Arab	3.97%	3.53%	63.99%	**3.09%**	**0.87**	**0.78**
9	Japanese	1.99%	1.22%	92.63%	**2.66%**	**2.18**	**1.34**
10	German	2.04%	1.30%	89.17%	**2.37%**	**1.82**	**1.16**

A consistency check between the 3 results is made, the third having to flow logically from the first two.

Table 16: Control of L1 and L2 results

	WORLD	WORLD	WORLD	ENGLISH	ENGLISH	ENGLISH	
	POPULATION	**CONNECTED**	**% CONN.**	**POP.**	**CONN.**	**% CONN.**	Control
L1	**7,231,699,136**	**4,223,428,027**	**58.40%**	5.12%	7.82%	89.24%	89.24%
L2	**3,130,017,620**	**1,673,121,762**	**53.45%**	31.25%	32.53%	55.64%	55.64%
L1+L2	**10,361,716,756**	**5,896,549,789**	**56.91%**	13.01%	14.83%	64.86%	64.86%
Control			56.91%	13.01%	14.83%	64.86%	

In green the checks are carried out: it is a question of calculating the same values directly and thus verifying that the two models L1 and L2 have functioned correctly: the proof is made.

The second series of controls is more complex and one should not expect perfect matches (because modeling is not a linear process with respect to demo-linguistic data).

Table 17: Checking L1 and L2 results (continued)

	English	Chinese	Spanish	French	Hindi	Portuguese	Russian	German
Content L1	12.96%	25.55%	8.76%	2.08%	2.93%	3.91%	3.22%	1.96%
L2 content	37.91%	10.68%	5.47%	6.90%	5.96%	0.89%	5.12%	3.61%
Contents L1+L2	19.60%	21.60%	7.85%	3.33%	3.76%	3.13%	3.76%	2.37%
Control	20.04%	21.33%	7.83%	3.45%	3.79%	3.05%	3.76%	2.43%

The first three lines show the results of the three respective models. The control line in green is calculated by weighting the respective L1 and L2 percentages with respect to the respective connected populations. So, for English, 20.04% is obtained by the following formula: ((12.96 x 4,233,428,027) + (37.91 x 1,673,121,762)) / 5,896,549,789

It is both remarkable and very reassuring, regarding the validity of the model, that the results obtained by the two methods (the L1+L2 model or the prorating of the results of the L1 and L2 models in relation to the respective connected populations) are so close.