Enhanced and second version of an alternative approach to produce indicators of languages in the Internet

Daniel Pimienta
Observatory of linguistic and cultural diversity in the Internet
http://funredes.org/lc

August 2021

Credits: The work leading to this enhanced version has been made possible thanks to the support of the *Cultural and Educational Department of the Brazilian Ministry of Foreign Affairs* within the frame of the International Institute of the Portuguese Language and the coordination of the UNESCO Chair on language policies for multilingualism. Credits are given also to Daniel Prado, who first conceived the idea of collecting diverse sources for measuring languages in the Internet as well as to transform figures per country into figures per language.

Thanks: To Professor Gilvan Müller de Oliveira for the support on linguistic matters and coordination with the funders, to Alvaro Blanco for the support in writing tricky programs which change radically the management of so many sources and so many orthographs of languages and countries and to David Pimienta who wrote the program required to transform the Ethnologue format into the required format for that study and process macro-languages.

Warning : The following study is essentially a statistical work based on a large variety of input sources. Adopting a major source in such type of works also implies logically adopting the rules sustaining the data of that source. The author is therefore not responsible for the list of considered countries and territories established by ITU, a United Nation agency, for the statistics of percentage of persons connected to the Internet, nor for the list of languages with more than five million L1 speakers according to Ethnologue and for the regrouping into macro-languages adopted by Ethnologue in concordance with standard ISO 693.3.

ABSTRACT

In a context of scarcity of reliable data about the space of languages in the Internet, the 2017 alternative approach to compute indicators of behavior in the Internet, for the 140 languages with more than 5 million speakers, has been enhanced and actualized. The enhancements of this approach based on the collection of a series of micro-indicators that measure languages or countries in various Internet spaces or applications are exposed. The use of the last Ethnologue Global Data Set allows not only to dispose of the most reliable and up to date demo-linguistic data but also give the ground to overcome one of the major bias of the method related to the process of the L2 speakers. The five indicators of languages in the Internet which has been defined and exposed in 2017 (*Internet users, traffic, use, contents, societal indexes and interfaces*), and 4 macro-indicators which are deduced from them (*power, capacity, gradient and content productivity*) are reproduced with all inputs updated in 2021. The results are showing the trends with English decreasing close to 25% and Chinese getting stronger while Spanish is comforted in third position. French shares now the third place with Hindi, with a reduced advance over a group of languages in very close positions: Portuguese, Russian, Arabic and German. As in 2017 edition, all possible biases derived from the method, assumptions or sources are discussed and finally an estimate is proposed that consider those biases. It is forecasted for the end of 2021 a new set of enhancements with the high possibility to extend the results for the 332 languages with more than 1 million L1 speakers, a limit that this method shall not cross to avoid stronger biases.

Keywords: Languages, Internet, linguistic diversity, indicators, bias

Contents

LIST OF TABLES AND FIGURES

TABLES

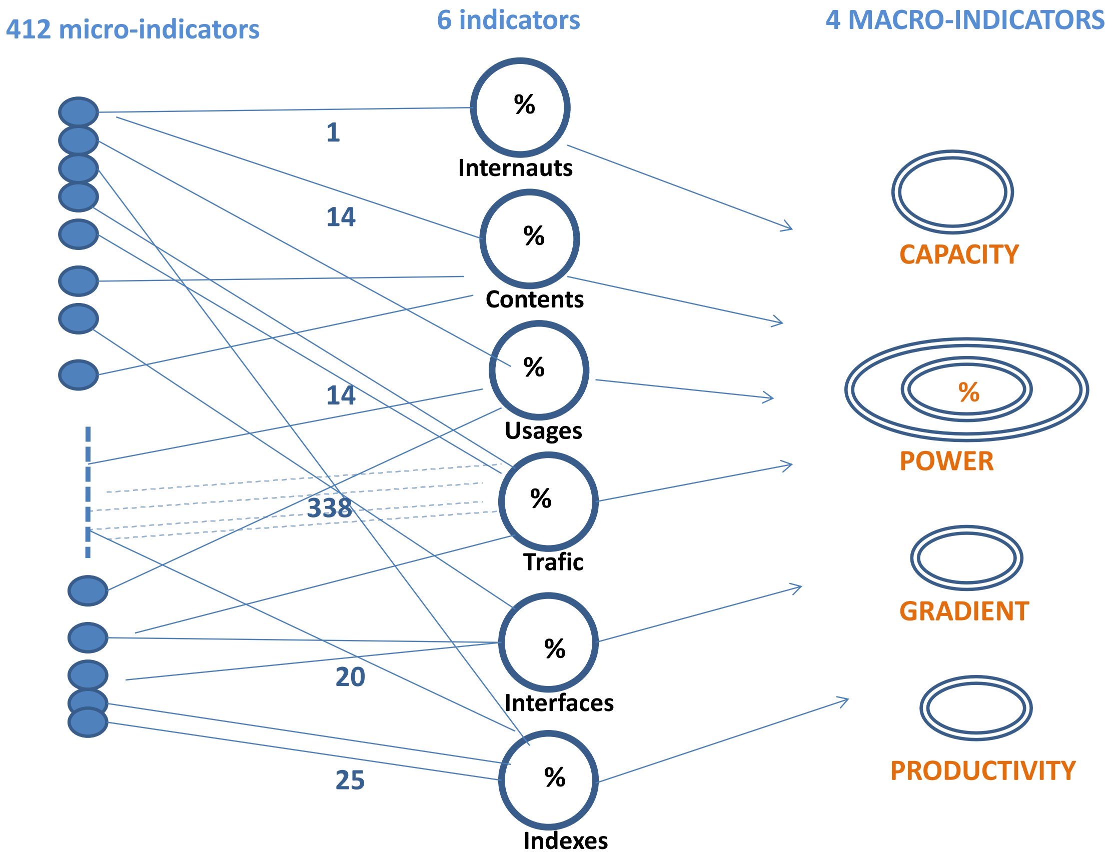
FIGURES

BACKGROUND

The first edition of this method to produce indicators of language presence in the Internet has been realized in 2017 and documented under the title "*An alternative approach to produce indicators of languages in the Internet" ([1])* accessible in the website of the Observatory in 4 linguistic versions (English, French, Portuguese and Spanish)[1]. The reader is invited to consult it previous to the reading of this paper which is written as a complement of the first version. The first version presented both the method and the results; this paper presents the differences in the method and the new results.

As a reminder, the method addresses the 138 languages with quantity of L1[2] speakers higher than 5 million[3] and produce indicators for each of them, under the following scheme (which numbers are updated for the second version).

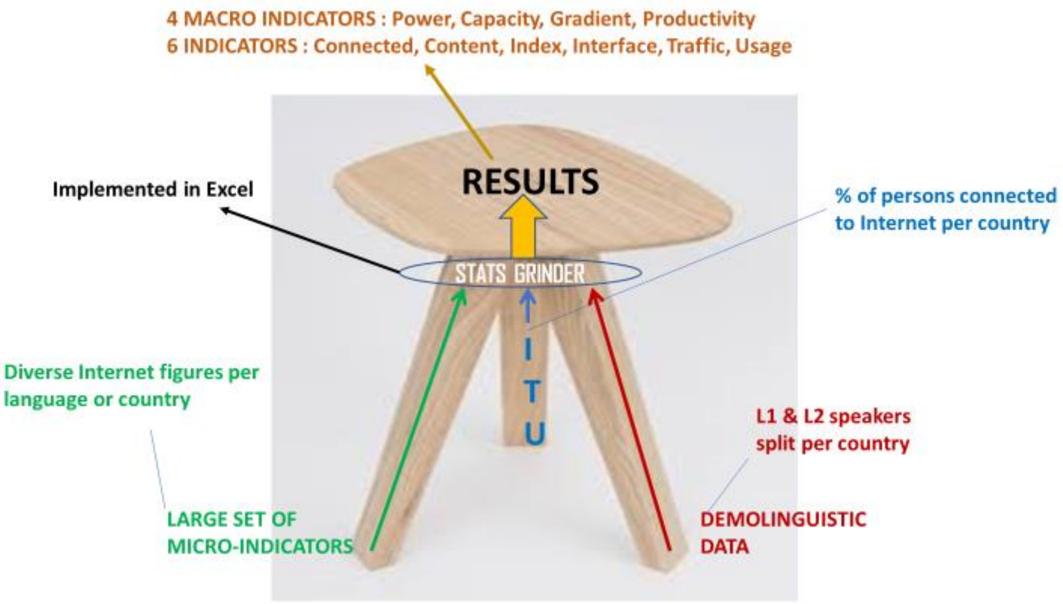
Figure 1: From micro-indicators to macro-indicators



The method relies in 3 type of inputs and 10 outputs as represented in the following figure.

[1] http://funredes.org/lc2017
[2] The convention used is to call L1 the mother tongue (or first language) and L2 the second languages, providing a sufficient level of control to be accepted in that category.
[3] As a matter of fact, the total is 128 : in order to be able to make comparisons with the 2017 study, 10 languages with less than 5 million speakers have been left because they appeared in the 2017 study, in order to be able to make controls and comparisons. Those languages are : Awhadi, Belarusian, Bikol, Bugis, Dugri, Armenian, Kimbundu, Luyia, West Flemissh and Southern Thai.

Figure 2: The input/output process of the model


4 MACRO INDICATORS : Power, Capacity, Gradient, Productivity
6 INDICATORS : Connected, Content, Index, Interface, Traffic, Usage
RESULTS
Implemented in Excel
% of persons connected to Internet per country
STATS GRINDER
I
T
U
Diverse Internet figures per language or country
L1 & L2 speakers split per country
LARGE SET OF MICRO-INDICATORS
DEMOLINGUISTIC DATA

The process of the model stands on weighting mechanisms able to **transform figures per country into figure per language**, **extrapolation technics** for completing sources with limited figures per country and **weighting mechanisms** with the figure of world repartition of Internet connected persons per country to produce world percentages of the different sources.

Table 1 : The 2 types of weightings used.

	Demo-linguistic	Internauts per language
TYPE	% per Country ---> % per Language	% Criterion ---> % worldwide
INPUT	Data by country	Given in % by specific criteria
OUTPUT	Data by language	Data in worldwideL1+L2 %
DATA WEIGHTING	L1+L2 Speakers per country matrix	% of persons connected to the Internet per country
SCOPE	All sources by country	Index and interfaces indicators.
IMPLIED ASSUMPTION	Independence of languages in the country	Modulation rate connection to the Internet according to the criterion

The model is implemented in Excel within a spreadsheet of 7 Megabytes with 17 correlated worksheets organized around the 215 countries considered, the 138 languages processed and the 412 micro-indicators collected. The model so implemented allows to verify in fraction of second the impact of any hypothesis (including prospective analysis).

1. INTRODUCTION

This second version of the referenced method to create indicators of the presence of languages on the Internet brings a set of **tangible enhancements** which improve considerably the reliability of the method and reduce the biases.

The major improvements derive for the **adoption of the Ethnologue Global Dataset 24**[4] of March 2021 which not only update the demo-linguistics data (the quantity of speakers of each language in each country) but also provide the most trustable data overall on the subject, even if perfect exactitude on that matter is unattainable, and additionally, in this last version, provide for the first source of L2 speakers of each language, split per country.

2. DIFFERENCES FROM FIRST VERSION

Many differences on the method or sources occurs from version 1 in the spirit of enhancing the quality of the method and the products.

2.1 Adoption of Ethnologue as demo-linguistic source

The main part of the Ethnologue source is in the form of an Excel matrix of 11500 lines with the following format: ISO639[5], Language Name, Country Name, number of L1 speakers, number of L2 speakers, plus a large set of associated parameters not used for this method.

In order to get the format required by the model (a matrix with all considered countries on column and all considered languages on lines) a set of cautious steps has been implemented, with the support of different computer programs written as macros for Excel. One of the most complex steps has been to fusion all figures for the languages belonging to each macro-language into a single one. This process has been involving 60 macro-languages regrouping 434 different languages[6] (see in Annex 2 the list of macro-languages).

After completing this step, the process consisted in reducing the large list of languages into the list of languages being processed by the model[7], summing carefully all the remaining figures per country into a single line “REST”.

It is important to understand that the adoption of the Ethnologue data implies the conformity with the imbedded rules which are based in pure linguistic considerations:

- Macro-language regrouping[8]
- List of countries and corresponding English naming.

The list of countries in the Ethnologue source is larger than the list processed by ITU[9] for the providing of the Internet connection rate per country (ITU as a UN entity does not separate, for instance Martinique from France). In that case, the ITU rule is the obliged one and the requirement has been to

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

[5] The ISO code with 3 characters assigned to each of the 7486 languages identified.

[6] For instance, Arabic macro language holds 29 languages such as Egyptian or Moroccan Arabic.

[7] At that stage 138 languages with the number of L1 speakers higher than 5 million.

[8] A significative example is the case of Serbo-Croatian macro language which definition regroups, in alphabetic order, Bosnian, Croatian, Montenegrin and Serbian. This obliged grouping does not answer at all to geo-political criteria and could even be considered as polemical from this standpoint. Additionally, as some sources separate clearly the involved languages and countries this produce some risk of error in the results even though the sources input has been transformed to pay attention to that situation (the risk occurs when the figures are not to be added but rather averaged like in the Depth indicator of Wikipedia).

[9] The International Telecommunications Unit (http://itu.int), the organ of United Nations which provide telecom stats including the percentage of persons connected to the Internet per country.

sum up all the Ethnologue figures for the 29 countries which appears in Ethnologue but not in ITU (for complete list see Annex 3) into a single column « Remaining countries ».

2.2 Management of L2 and multilingualism

The inclusion of the last Ethnologue data on the model allowed, as a by-product, to eliminate the major bias of the method which was linked to the process of the second language (L2) in the model. For the first time there is a trustable source which completes the number of L1 speakers per country with **the number of L2 speakers per country.** In the 2017 version, the L2 figures for persons connected were computed from the total of L2 speakers worldwide, applying the Internet connectivity rate computed by the model for L1 speakers. An important bias resulted from the fact that for some major languages (as for example French and English) a high proportion of L2 speakers belongs to developing countries where the average Internet connection rate is much lower than what is computed in average for L1 speakers. This bias inflated the results for English and French (and some other languages) and obliged to a "manual" bias correction.

Another positive consequence of the use of Ethnologue data is the ability to get an "official figure" for **multilingualism**. The world ratio (L1+L2)/L1 was established in 2017 edition by projecting data available for the processed countries: it resulted to be around 1.25. Now the figure is provided indirectly by Ethnologue data and its value is 1.43.

Following Ethnologue figures:

- ✓ The total worldwide (L1) population is given as: 7 231 699 136
- ✓ The total worldwide L1+L2 speakers is given as: 10 361 716 756
- ✓ The "multilingualism ratio" is then 10 361 716 756/7 231 699 136 = 1.4328
  (in other terms **43% of the population speaks more than one language**).

This figure of 43% is clearly much better than the 25% used in the first version and this is not an anecdotical element of the model but one of the key elements. As shown in the first study, the most common and critical bias of the figures offered on languages is the fact that they are not considering correctly the L2 speakers (issue which expresses fully in the Internet where most internauts do use their L2 languages and many websites are multilingual[10]). Not paying due attention to multilingualism conduces to tremendous errors, often hidden in "the rest of languages", as world percentages are computed over a total of 7 billion (the world population) where it should be over a population of 10 billion (the L1+L2 speakers).

In that second version, the **principle of measuring everything in terms of L1+L2 population** (instead of the world population) has been fully adopted to insure accuracy to the results. For that reason (and also because of other improvements) comparison between 2017 and 2021 results are to be made with caution. As a matter of fact, all the macro-indicators, *power* but also *capacity* and *gradient*, are now following this rule of being computed over the L1+L2 population instead of the L1 population (and will then appear lower than in 2017 version).

[10] As a matter of fact, the 5 indicators processed by the study are by nature multilingual: internauts visit websites and generated traffic in the different languages they manage, often websites are multilingual, interfaces are multilingual, translation services cover different languages…

2.3 Source for persons connected to the Internet

Until 2017, ITU used to provide each year an update of its figures[11] on the *percentage of individuals who use the Internet per country*, including its own estimates whereas there is no official source in a given country. This input, which is indeed the most important element of the method, was considered one of the most reliable sources. Unfortunately, after 2017, ITU has decided to stop providing its own estimates, which leaves many countries (almost all developing countries[12]) with old figures of 2017 in 2021.

This posed a serious issue to the method and after many iterations drove the decision to violate, in that case, a strong principle which is basic in this type of statistical tasks: never change the data of the sources, take it as it is.

The World Bank provides its own figures[13] for the same concept, which are clearly retaken from ITU, but, in many cases, overcomes the ITU limitation and does offer new data where ITU has left 2017 data. This is a progress; however, many countries still remain out of the update from 2017 and this would impact negatively the languages spoken in those countries and prevent to perceive possible progress.

Finally, it was decided to use the World Bank data when they are different from ITU's and, for the many remaining cases lacking actualization, do, for each concerned country, an Internet search for reliable data and provide estimates based, when there is no evidence of arguments against, in the approximate linear progression from previous data.

One case remained an issue: **India** has now a 2021 official figure of 20.1% while the 2017 ITU estimate was 32%... and many sources on the Internet claim a boost of the Internet in India in the last years with figures around 50%[14]! After failing to obtain answer from the official source and from Indian colleagues consulted, it was decided, due to the paramount importance of India in the study context[15], to exceptionally violate a still stronger principle: not to change official sources. The working hypothesis made is that the figure provided by the *Indian Ministry of Statistics and Program Implementation* concerns only the fixed type of connections and leave outside the mobile connections to the Internet. Based on that hypothesis, the conservative figure of 40% was set. Note that the sensitivity of this figure on the results is not marginal. Hereafter the different model results for Hindi and Bengali depending of the figure selected.

Table 2: Sensitivity of India figures for percentage of persons connected to the Internet

India % Connected persons	20.08%	30%	40%	50%
Hindi Power (ranking)	2.42% (10)	2.91% (8)	3.38% (5)	3.81% (4)
Bengali Power (ranking)	0.75% (17)	0.82% (15)	0.88% (15)	0.95% (14)

[11] https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2021/PercentIndividualsUsingInternet.xlsx

[12] Only 80 countries have provided official figures in 2019.

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

[14] See for instance in https://www.statista.com/statistics/255146/number-of-internet-users-in-india/ or https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

[15] With major languages such as Hindi and Bengali and also 34 languages which are part of the list of considered languages.

2.4 Management of sources for micro-indicators

The whole process of sources management for micro-indicators is the most heavy, cumbersome and challenging task of the project, with high consumption of human resources. There are many steps involved:

1. For each category of indicator, search the Internet for sources
2. Select sources based on reliability and applicability to the process
3. Collect sources in a format able to allow automatic integration to the model
4. Integrate sources to the model and associate a theme
5. Evaluate biases of the sources

In annex 1, the full list of sources, for each type of indicator, is presented.

In order to do step 4, the data needs to get transformed into an Excel format with the appropriate names of Countries and languages, in the same order than the one used in the model.

As for step 3, all the sources are collected from a specific URL (see Annex 1 for the complete list of URLs). Most of the sources are obtained in HTML format, some others in PDF format and a limited subset (mainly ITU and World Bank's) in Excel format, which is the target to transform all the sources. The process of transformation from PDF format into Excel could be relatively straightforward in most cases, however in some cases there is incompatibility and some tricks are required, such as passing first by an intermediary DOC format.

The process of transforming from HTML format to Excel format can often turn into a real nightmare, requiring a lot of imagination and tricks, including in several case trying to retrieve the data inside the HTML source and attempting, from there, to construct a table using the convert function of Excel.

In a growing number of cases, the source offers a geographic access to the data (clickable maps) which, except when the number of countries or languages is limited and copying by hand is not too heavy, makes it impossible to process or requires subcontracting a person for a hand collection job which is tedious but require high concentration and discipline to avoid errors. The collection of traffic data involving hundreds of micro-indicators was subcontracted that way.

Credits must be given to the institutions (in general, international organizations or NGOs) which provide the data in a computer exploitable format (Wikimedia for example provides, in its English version, HTML tables which are always transformed directly in Excel format, without trouble).

The transformation of the source into an Excel file (in general, a table of country names and numerical percentages or values) is not the end of the game. With 214 countries or hundreds of languages to be processed and rare utilization of ISO codes, but instead literal names which can be in different languages and non-standardized orthographs, the setting into a model, which bear its own meaningful order for the countries and languages, is not feasible by hand. Two programs have been written for that process, which in both cases needs some recursing tuning[16] in order to integrate the various orthographs (which has been conserved in a file used by the programs). The final output of those programs is an Excel file directly usable to copy entirely, or line by line, the sources into the appropriate spreadsheet of the model. Besides the huge gain of time to that method it also warrants to get the data from the sources without errors.

Note also that the decision to match Ethnologue formats and to treat all the languages part of a macro language as a unity has made this process still more complex, as macro regrouping needs to be processed into the very sources, prior to process. To take some examples, frequent occurrences of Egyptian or Moroccan Arabic in sources has been cumulated to Arabic and Serbian, Bosnian, Croatian and Montenegrin data has been merged into Serbo-Croatian (the number of similar cases being quite high). For the manual process of the list of unidentified languages identified by the program, extensive use of the Ethnologue page https://www.ethnologue.com/language/srp has been made.

2.4.1 INDEX

The deadline came too early when production of the 2017 version was made and this indicator came short with a single source providing only 5 micro-indicators. This time, the required attention was given and an almost **exhaustive collection** has been realized for this indicator. A large variety of parameters characterizing the progress of countries in the Information society have been included, with 25 micro-indicators now, from electricity stability to artificial intelligence, crossing to Governance and many other parameters (see Annex 1 for the full range).

2.4.2 CONTENT

As explained before, the sources for languages figures on the Internet are extremely scarce and this makes this indicator rely heavily on Wikimedia outstanding statistics. The fact, discovered in that second version, is that the presence of languages in Wikimedia is not proportionate to their presence in the real world, as shown in the section of the document analyzing biases, nor could be considered a faithful reflect of their very presence in the Web. Some way to balance the results of Wikimedia stats has been introduced and the painful diagnostic is that *Content* is the weaker indicator of this method and at the same time a very sensible and sensitive one (changes in this indicator can provoke important impact in the resulting macro-indicators). While one of the main goals of the project is to know the content repartition per language in the Internet, this objective remains hard to get with a frustrating difficulty to weight correctly the *contents*[17] and *power*, a holistic macro-indicator, remains yet the best approximation of the presence of languages in the Web.

[16] The recursive process ends when the process of the sources produces no more unknown orthographs.

[17] As shown in the first edition, the commendable effort of W3Techs to offer updated figures for contents is biased at many different levels (the strongest but not unique being the lack of consideration of multilingualism and the fact that most multilingual websites including English are probably computed as English only). This source projects values for English contents in the Web which are extremely exaggerated (above 50% whereas the reality is probably today below 25%). The lack of sources fuels the myth in the media that more than half of websites are

To try to control better the excessive influence of Wikimedia figures on this indicator two decisions were made. The first one concern exclusively Wikipedia: instead of having one indicator for each of the figures provided (*number of articles, active editors, edits and depth*[18]) a formula has been set up to define a single micro-indicator:

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

This formula expresses more accurately the Wikipedia overall activity per language, not giving so much importance to languages where bots[19], instead of humans, are used to create articles by translating from another language version and hardly updating the articles further[20].

The following table shows how the formula manages to reflect better the reality. The last column (presence) is the ratio between the number of articles and the L1+L2 population (number of articles per speaker) is a clear demonstration of why the presence of languages in Wikipedia is not a good indicator of the overall presence of languages in the Internet… Note that the depth value for Vietnamese was not informed and a value of 1 was set to avoid a null formula[21].

Table 3: Wikipedia factors and the formula

Language	Articles	Edits	Active Users	Depth	FORMULA	PRESENCE
English	6332139	1027716498	125399	1073	481775	0,47
Cebuano	5853095	32075254	186	2	275	36,71
Swedish	3050759	49330695	2148	12	22759	23,37
German	2593827	212207089	18119	93	50897	1,92
French	2342875	183969129	18054	242	26424	0,88
Dutch	2060512	59302602	3933	17	13742	8,45
Russian	1736736	115035192	10425	137	4286	0,67
Italian	1703284	121418801	8085	172	62435	2,51
Spanish	1698331	136390848	15694	210	2590	0,31
Polish	1480982	63723938	4235	32	7742	3,64
Japanese	1277204	84188217	15173	85	8683	1,01

in English. This was the case between 2007 and 2009 (see [3]), but since the exponential growth of Chinese, Hindi, Arabic, Turkish, Bengali, Vietnamese, Urdu, Persian and Marathi, to name new languages in the first 20 ranks and together weighting close to 28% of contents, has radically changed the situation and English represents today only a quarter of the content. Between 2000 and 2007, the persistent myth was that English occupied 80% of the Web and this disinformation finally disappeared after 2009 with the publication by UNESCO of reports (see [3] and [4]) which established a presence of English around 50%. How come English would have kept stable at 50% during 14 years while the Internet was changing demography and the number of connected English speakers (L1+L2) has decreased from 32% of the total of connected persons in 2007 (source : https://web.archive.org/web/20120511104604/http://dtil.unilat.org/LI/2007/es/resultados_es.htm) to only 13% today?

[18] Quoted from Wikimedia: *Depth*, which is defined as [Edits/Articles] × [Non-Articles/Articles] × [1 − Stub-ratio]), is a rough indicator of a Wikipedia's quality, showing how frequently its articles are updated. It does not refer to *academic* quality.

[19] A bot is a computer program behaving like a human from the point of view of the application interface.

[20] Without this formula Cebuano, with huge number of articles but very low depth, appeared with the highest *capacity* score.

[21] The low value of *depth* is a reflect of the fact that 67% of articles are been made by bots, not by humans (source: https://www.wikiwand.com/en/Vietnamese_Wikipedia).

Vietnamese	1266628	65110373	2476	1	35	1,65
Chinese	1208732	66159632	8940	202	62	0,08
Arabic	1123561	54279052	5189	227	536	0,31
Ukrainian	1100281	32831286	2773	53	4823	3,32
Portuguese	1067241	61371751	9508	176	1651	0,41

In the chapter discussing biases, a deep analysis is made of the Wikimedia statistics.

The second decision made to balance the Wikimedia influence on the *content* indicator is a system of weighting, implemented in regard to each *content* micro-indicator, which gives more importance to the T-Index of Translated[22] than to the whole Wikimedia collection of indicators. Playing with different configurations of weighting factors showed the high sensitiveness of the value of this indicator, basically due to the very low number of sources and the fact that some languages have disproportionate presence in some Wikimedia items.

The configuration of weighting finally implemented is the following:

Table 4: Weighting of content indicators

ITEM	WEIGHT
Amazon US - number of books 2017[23]	0,5
Wikipedia formula	1
Number of WikiBooks per language	0,5
WikiQuote articles per language	0,1
Number of WikiSource articles per language	0,1
Number of articles Wikiversity per language	0,1
Number of articles Wiktionnary per language	0,1
Number of articles WikiNews per language	0,1
Number of articles WikiVoyages per language	0,1
T-Index for e-commerce Projection 2021	3

2.4.3 TRAFFIC

This step has also been very dense with a lot of trial and errors. In 2017, it was established that the Alexa Traffic data were extremely biased against Asian countries (especially India and China), and Brazil, and somehow biased also in favor of French and English. Four years after, the Alexa data collection showed strange patterns (the output would not show traffic in the country of creation of some sites[24]) and the feeling was that European countries traffic was underestimated, while, in the other hand, India appears quite high in all sites, not so much China.

[22] This index, accessible at https://translated.com/les-langues-qui-comptent, is an attempt to measure the potential of languages in electronic commerce, from the number of internauts per language, multiplied by the estimated online expenses. It uses World Bank and ITU figures and proposes a 2021 projection which is the figure selected for the model. It is, besides Wikimedia data, one of the extremely few serious sources available for languages in the Internet.

[23] The lack of equivalent accessible data for 2021 and the situation with Wikimedia drove the decision to keep this micro-indicator in spite not being actualized.

[24] As examples, theses.fr showed zero traffic in France, the same with spip.net, a CMS mainly used in France.

A study comparing the traffic data with the subscription data for 5 main social networks first confirmed the intuitive findings. In summary, Brazil traffic seems largely underestimated compared to the level of subscription, as well France, Germany, Italy, Spain and United Kingdom; on the other hand, India, Japan, Korea appear largely overestimated (see the Chapter discussing the biases for more details).

In front of those un-trustable results, it was decided to look for alternative measurement tool. SimilarWeb.com looks as a possible alternative and the test was intended prior to buying subscription. Unfortunately, it was impossible to reach the country data in the website, and, in spite of many intents thru different channels, including the interactive chat of the company, no answer was ever obtained.

Facing this blocking situation, another provider, Semrush.com, was tested and country figures were collected for the same set websites. Semrush, at difference of Alexa, provides, for each measured site, the results for all countries, which was an attractive prospect, leaving out the need for extrapolation. However, it happens that in some cases the total goes short of 100% (which is not a problem) and some other times it goes over 100% (which is a problem). Finally, the figures were normalized to exact 100% using a pro-rata rule before introduction to the model.

After running the model, transforming country data to language data, the results were not convincing: Chinese value was quite too low, the same for Hindi and Arabic and for the "remaining languages".

The extreme differences between Alexa and Semrush results, after running the model for the same set of websites, are an alarm signal about the reliability of such tools and a worry for future plans to extend the number of websites studied and allow theme differentiation results for some languages.

2.4.4 INTERFACES

The list of languages supported in important application's interfaces, or as a possible target for translation services in the Web, does not pose any particular problem. The list of applications selected can be consulted in Annex 1. In order to reduce the importance of the Wikimedia figures on the model the decision was made to remove from this indicator the Wikimedia sources.

2.4.5 USAGES

No particular difficulties either for this indicator, except to find free of charge figures for the main social networks (mainly number of subscribers per country). Finally, the coverage managed to include the following applications: Facebook, Instagram, Linkedin, Messenger, Pinterest, Reddit and Twitter. Additionally, some sources not related to social networks were included (as for example the number of downloads of OpenOffice per country), see the full list in Annex 1.

2.5 Summary of Indicators

The following table summarizes the description of each of the indicators and explain how it is built from micro-indicators.

Table 5: Description of indicators

INDICATOR	DEFINITION	TECHNICAL	RELIABILITY/BIAS
A: INTERNAUTS	Mono indicator derived from ITU and World Bank figures of world % of people connected per country extrapolated where recent figures are lacking.	weighting country -> language without extrapolation	High reliability Very marginal bias although increasing because of lack of update for many countries.
B: USAGES	Includes 14 micro indicators with 2021 data: - Fixed + mobile % per country - Broadband % per country - Cumulative OpenOffice download - Facebook, Instagram, LinkedIn, Messenger, Netflix, Pinterest Twitter, YouTube, % subscribers per country	weighting country-> language extrapolated proportionally Mean of micro indicators	Strong reliability. Low bias.
C: TRAFFIC	Alexa measured traffic per country to a selection of 338 websites.	weighting country-> language extrapolated proportionally Truncated mean to 20%	Relatively good reliability But strong European negative bias of Alexa confirmed by comparisons of traffic and number of subscribers per country.
D: INDEXES	Includes 25 indexes from various sources measuring parameters such as: - E.government - Universal Access - E.participation - General infrastructure (See Annex 1 for complete list)	weighting country-> language extrapolated by quartile method then transformed into world percentage weighting data with ITU Mean of micro indicators	Good reliability and marginal bias (subjective data quantified by a competent body).
E: CONTENTS	Includes 13 micro indicators with associated weighting. T-Index of Translated a measure of the potential for e-commerce of a list of languages (2021) - Number of books at Amazon (2017) - 11 language micro indicators from Wikimedia: articles, users or editors; all Wikipedia indicators are synthetized with a formula.	Direct use of figures per language weighted to balance Wikimedia importance. Merge of Wikipedia 4 indicators with a formula. Truncated mean to 20% of micro indicator	Very strong for Wikimedia and Amazon. But quite biased due to very low presence of some major Asian languages. The number of micro-indicators would need to be extended to give more strength to the mean.
F: INTERFACE (and translation languages)	Includes 23 binary micro-indicators	Presence % on all 23 micro indicators. Word % by weighting with ITU figures.	Perfect.

3. RESULTS

The following tables show the results, sorted by the various macro-indicators, for each of the indicators and macro-indicators, except productivity[25].

The following table shows all the summary results for the 15 most "*powerful*" languages in the Internet. All percentages are made on the basis of L1+L2 population.

Table 6 : Indicators for the top 15 languages in terms of power

	W.Conn.	W.Pop	TRAFIC	L.Conn.	USAGE	CONT.	INTERF.	INDEX	POWER	Capac.	Grad.
English	15,30%	13,01%	37,44%	64,33%	27,92%	38,61%	21,73%	17,87%	**26,48%**	**2,04**	**1,73**
Chinese	17,65%	14,72%	7,79%	65,59%	5,47%	8,18%	25,07%	19,38%	**13,92%**	**0,95**	**0,79**
Spanish	7,00%	5,24%	10,72%	73,08%	11,74%	5,42%	9,94%	7,59%	**8,73%**	**1,67**	**1,25**
French	3,00%	2,58%	2,64%	63,67%	3,75%	5,40%	4,26%	3,21%	**3,71%**	**1,44**	**1,24**
Hindi	4,26%	5,80%	4,81%	40,18%	3,16%	0,28%	4,03%	3,71%	**3,38%**	**0,58**	**0,79**
Portuguese	3,05%	2,49%	1,42%	67,16%	5,53%	3,30%	3,85%	2,92%	**3,35%**	**1,35**	**1,10**
Russian	3,51%	2,49%	1,81%	77,20%	2,28%	3,38%	3,88%	3,78%	**3,11%**	**1,25**	**0,88**
Arabic	3,89%	3,53%	2,30%	60,14%	3,02%	2,05%	4,29%	3,01%	**3,09%**	**0,88**	**0,80**
German	2,09%	1,30%	1,32%	87,65%	1,95%	5,84%	2,97%	2,98%	**2,86%**	**2,19**	**1,37**
Japanese	2,07%	1,22%	1,98%	92,62%	1,76%	3,55%	2,77%	3,01%	**2,52%**	**2,07**	**1,22**
Malay	2,20%	2,36%	0,89%	51,00%	2,79%	0,79%	1,91%	1,99%	**1,76%**	**0,75**	**0,80**
Italian	0,91%	0,66%	0,51%	75,65%	0,97%	3,39%	1,22%	1,20%	**1,37%**	**2,09**	**1,51**
Turkish	1,21%	0,85%	1,03%	77,98%	1,59%	0,94%	1,43%	1,22%	**1,24%**	**1,46**	**1,02**
Korean	0,93%	0,79%	0,93%	64,73%	0,99%	0,85%	1,10%	0,95%	**0,96%**	**1,22**	**1,03**
Bengali	1,14%	2,58%	1,22%	24,15%	1,13%	0,26%	0,72%	0,84%	**0,88%**	**0,34**	**0,78**
REST	**31,79%**	**40,39%**	**23,19%**		**25,95%**	**17,77%**	**10,81%**	**26,34%**	**22,64%**		
TOTAL	**100%**	**100%**	**100%**		**100%**	**100%**	**100%**	**100%**	**100%**		

W.Conn. : percentage of speakers of that language connected to the Internet related to total speakers connected to the Internet
W. Pop. : percentage of speakers of that language related to the total world L1+L2 population
L. Conn. : percentage of L1+L2 speakers of that language who are connected to the Internet
REST : represents the results for the full set of all languages of the world except the 15 languages listed in the table.

It must remain clear that the ranking in terms of power favors the languages that have the largest number of speakers. The capacity and gradient macro-indicators offer results independently of the number of speakers.

Reminder:
Power[26] has been defined as the mean of the 5 indicators.
Capacity[27] is the value of power divided by the % of L1+L2 speakers

[25] This indicator will be revisited in the chapter *Correction of biases*. The *power* indicator, which integrates all the elements would probably be, at this stage, a better approximation to the distribution of contents per language data which remains very difficult to get in a trustable manner as of today.

[26] The term **power** has been used instead of *weight* to avoid confusion with the heavy transversal use of weighting in the method. It represents the absolute presence of a language in the Internet, integrating all factors.

[27] The **capacity** is the relative presence of a language in the Internet, independently of its number of speakers; it indicates the dynamism of a language in the Internet.

Gradient[28] is the value of power divided by the % of connected L1+L2 speakers.

The following table is sorted by connected languages, the most connected first.

Table 7 : Languages sorted by percentage of people connected

INTERNAUT SORT	Internauts	Capacity	Gradient
Danish	97,82%	2,19	1,22
Swedish	93,49%	2,61	1,53
Japanese	92,62%	2,07	1,22
Dutch	92,02%	2,26	1,34
German, Swiss	91,56%	1,21	0,72
West Flemish	90,43%	1,12	0,68
Finnish	89,67%	3,42	2,09
Bavarian	87,68%	0,97	0,61
German	87,65%	2,19	1,37
Hebrew	85,46%	5,24	3,35
Slovak	82,47%	1,30	0,86
Belarusian	82,27%	1,00	0,66
Czech	81,37%	1,70	1,14
Polish	81,17%	1,88	1,26
Hungarian	79,92%	1,79	1,22
Tatar	78,05%	0,87	0,61
Turkish	77,98%	1,46	1,02
Serbo-Croatian	77,78%	3,14	2,21
Greek	77,71%	1,75	1,23
Russian	77,20%	1,25	0,88
Kazakh	76,98%	0,90	0,64
Romanian	75,66%	1,18	0,86
Italian	75,65%	2,09	1,51
Albanian	75,48%	1,12	0,81
Azerbaijani	74,76%	0,94	0,69
Napoletano-Calabrese	74,39%	0,84	0,62
Spanish	73,08%	1,67	1,25
Kurdish Macro	73,02%	0,89	0,67
Bulgarian	70,34%	1,18	0,92
Armenian	69,86%	1,41	1,11
Vietnamese	69,04%	1,07	0,85
Guaraní	68,83%	0,64	0,51
Portuguese	67,16%	1,35	1,10

The following table is sorted by capacity.

[28] The **gradient** indicates the dynamism of the connected speakers; the term *gradient*, expressing a derivate and therefore a trend or a drive, has been chosen because a high gradient is a promise of increasing capacity.

Table 8 : Languages sorted by capacity

CAPACITY SORT	Internauts	Capacity	Gradient
Hebrew	85,46%	5,24	3,35
Finnish	89,67%	3,42	2,09
Serbo-Croatian	77,78%	3,14	2,21
Swedish	93,49%	2,61	1,53
Dutch	92,02%	2,26	1,34
German	87,65%	2,19	1,37
Danish	97,82%	2,19	1,22
Italian	75,65%	2,09	1,51
Japanese	92,62%	2,07	1,22
English	64,33%	2,04	1,73
Polish	81,17%	1,88	1,26
Hungarian	79,92%	1,79	1,22
Greek	77,71%	1,75	1,23
Czech	81,37%	1,70	1,14
Spanish	73,08%	1,67	1,25
Turkish	77,98%	1,46	1,02
French	63,67%	1,44	1,24
Armenian	69,86%	1,41	1,11
Portuguese	67,16%	1,35	1,10
Slovak	82,47%	1,30	0,86
Russian	77,20%	1,25	0,88

And finally, the last table, sorted by gradient, highlights the dynamism of people connected. The presence of Malagasy so high[29] is a consequence of the dynamism of its speakers in some Wikimedia indicators.

Table 9 : Languages sorted by gradient

GRADIENT	%	Capacity	Gradient
SORT	Internauts		
Hebrew	85,46%	5,24	3,35
Serbo-Croatian	77,78%	3,14	2,21
Malagasy	9,79%	0,40	2,21
Finnish	89,67%	3,42	2,09
English	64,33%	2,04	1,73
Swedish	93,49%	2,61	1,53
Italian	75,65%	2,09	1,51
German	87,65%	2,19	1,37
Dutch	92,02%	2,26	1,34
Polish	81,17%	1,88	1,26
Spanish	73,08%	1,67	1,25
French	63,67%	1,44	1,24
Greek	77,71%	1,75	1,23

[29] Such a ranking for Malagasy, a language with less than 10% of speakers connected, and a very low *capacity*, can legitimately provoke surprise: this is the result of a “mathematical accident” due a hugely disproportionate presence in one of the *content* micro-indicators and is indeed a symptom of the weakness of this indicator which is discussed hereafter.

Danish	97,82%	2,19	1,22
Hungarian	79,92%	1,79	1,22
Japanese	92,62%	2,07	1,22
Czech	81,37%	1,70	1,14
Armenian	69,86%	1,41	1,11
Portuguese	67,16%	1,35	1,10

Beyond the quite logical fact that the national languages of countries acknowledged for their proactive policies for the information society appear in the top positions, it is remarkable that several languages rate above English in spite its strategic advantage in the Internet to be the preferred language of choice for multilingual content and the belief of many it is the Internet lingua franca.

Those results have to be taken paying attention to the biases mentioned in the document, especially the difficulties with the *content* indicator whose changes may impact considerably those macro-indicators[30].

4. RESULTS ANALYSIS

Although comparisons with 2017 results is to be made with caution due to the importance and nature of the changes (specially the decision to express percentages in relation with the total world L1+L2 population), some phenomena can be highlighted.

The expected growth of Hindi which compete now with French for the 4th place and the apparition of Turkish in the list of top languages. As expected also, the differences between the group of followers of French are too close to consider the results are beyond the confidence interval; Portuguese, Russian, Arabic and German. However, the demographics may in the close future separate the respective positions at the speed of digital divide reduction.

As for the macro-indicators independent of the number of speakers, the apparition of Serbo-Croatian has to be taken with caution due to the process of the indicators resulting to the decision to adopt the Ethnologue classification as macro-language. And clearly, the indicator *content* and its actual high dependency on Wikimedia statistics, in spite the effort made to counterbalance it, clearly favors languages whose speakers have invested in Wikimedia presence. See the table below those languages., first sorted by the ratio 1000 x Number or articles/L1+L2 speakers and then sorted by the result of the formula set up (factor)

Table 10: Wikipedia presence of top languages

Language	Articles	Edits	Active Users	Depth	FACTOR	%FACTOR/L1+L2	%FACTOR/CONN	ART/L1+L2
Swedish	3050759	49330695	2148	12	22759	1,74	1,86	233,68
Finnish	512026	19813368	1752	40	21354	3,70	4,13	88,74
Dutch	2060512	59302602	3933	17	13742	0,56	0,61	84,51
Serbo-Croatian	1514114	78699318	1959	92	53779	2,69	3,46	75,77
Belarusian	281379	6093511	384	61	2620	0,67	0,81	71,87
Danish	267641	10777444	767	64	4486	0,80	0,82	47,64

[30] Prior to the introduction of the Wikipedia formula and the Wikimedia weighting, Cebuano, the second language in terms of number of Wikipedia articles, close to English, therefore with a content presence two order of magnitude higher than its speaker's presence, appeared first in the *gradient* table…

Hungarian	489514	23958462	1561	59	6871	0,55	0,69	39,04
Polish	1480982	63723938	4235	32	7742	0,19	0,23	36,44
Czech	484445	20095461	2242	46	5593	0,42	0,51	36,16
Ukrainian	1100281	32831286	2773	53	4823	0,15	0,23	33,16
Bulgarian	273163	11023721	789	27	942	0,11	0,16	33,10
Hebrew	298053	31660591	3335	258	92147	9,82	11,49	31,75
Italian	1703284	121418801	8085	172	62435	0,92	1,22	25,10
German	2593827	212207089	18119	93	50897	0,38	0,43	19,21
Japanese	1277204	84188217	15173	85	8683	0,07	0,07	10,11
Persian	816984	32472834	5416	172	3534	0,04	0,07	9,77
French	2342875	183969129	18054	242	26424	0,10	0,16	8,78
English	6332139	1027716498	125399	1073	481775	0,36	0,56	4,70

The following table shows clearly why some languages, such as Hebrew, Finnish and Serbo-Croatian, have gotten an advantage in the final results sorted by gradient.

Table 11: Wikipedia presence sorted by formula figures

Language	FACTOR	%FACTOR/L1+L2	%FACTOR/CONN
Hebrew	92147	9,82	11,49
Finnish	21354	3,70	4,13
Serbo-Croatian	53779	2,69	3,46
Swedish	22759	1,74	1,86
Italian	62435	0,92	1,22
Danish	4486	0,80	0,82
Belarusian	2620	0,67	0,81
Hungarian	6871	0,55	0,69
Dutch	13742	0,56	0,61
English	481775	0,36	0,56
Czech	5593	0,42	0,51
German	50897	0,38	0,43
Polish	7742	0,19	0,23
Ukrainian	4823	0,15	0,23
Bulgarian	942	0,11	0,16
French	26424	0,10	0,16
Japanese	8683	0,07	0,07
Persian	3534	0,04	0,07

Those considerations naturally lead to the discussion on biases.

5. BIASES ANALYSIS

There are three main categories of biases susceptible to affect the results:

- Biases proper of the method
- Biases from source’s selection
- Biases from sources

5.1 Biases proper of the method

One of the main biases proper of the method, which result of giving the same figure of percentage of L1 speakers connected to the Internet for L2 speakers, has been eliminated with the switch to Ethnologue data, gaining the repartition of L2 speakers per country. This strong bias affected particularly the languages with an important L2 population in countries with low connectivity rate (such as French and English). This is a paramount progress for the trust of the figure produced by the established model.

The second main bias proper of the method is to consider that, within a given country, all language speakers hold the same connectivity percentage (in other terms the national percentage of persons connected to the Internet is applied to all speakers, independently of their mother tongue). This bias forbids to distinguish speakers of different languages within a country with the method (for example, Catalan speakers in Spain are given the same connectivity percentage than Spanish speakers and no differentiate advantage can be analyzed, the same with Martinique creole in France, the same with the many languages of India). It is understandable intuitively that this assumption is not verified in many cases (the national digital divide could be linked to linguistic considerations) and that the impact of this bias is as strong as the language population is low. Marginal effect is expected if the model is limited to speaker's population higher than 5 million (although in the case of India it is not so obvious). The next launch of the model, forecasted to conclude before the end of 2021, will try to push the limit to languages with more than 1 million speakers.

Other marginal biases of the model may result of the adoption of structures implied by main sources. For instance, the split into countries has been derived from ITU classification and do not distinguish some territories.

5.2 Biases from sources' selection

There is obviously a "*selection bias* ", which is not proper of the methodology but belongs to the application of the method, where the decision on what source selection is made implicitly favor criteria proper of one's cultural background and ignore unconsciously data from countries too remote from one's experience. This may apply to each of the indicator and impact specially *traffic* where the selection of websites is hardly even between countries and can be influent even if the number of websites is counted in hundreds. The use of the *truncated mean at 20%* has been implemented to reduce such biases, after verifying that 20% was a large span capable to eliminate the large majority of results centered in websites with high language locality.

5.3 Biases from sources

The biases resulting from sources are discussed in the table below, rating each indicator with a value from 0 (totally unreliable) to 20 (bias-free).

Table 12: Bias rating by indicator

INDICATOR	RATING	COMMENTS
INTERNAUTS	19→16	This indicator derives from a unique micro-indicator. The main source is ITU. In 2017, this was the best rated sources with a 19/20 but in this release the rating drops to 16 because ITU has stopped to provide its own estimation when the country does not produce official data. ITU figures has been completed by World Bank's whenever possible and a linear projection of previous year's data has been set for the other cases. This indicator is key in the method as it serves as weighting of the results in several situation, however the factor analysis showed that the impact of small variation is moderate. As an example, if the connection rate for Brazil will be set at 80% instead of the actual value of 74% Portuguese *power* value would increase from 3.26% to 3.39%.
INDEX	15→18	This indicator derives from a mix of 25 micro-indicators rating different parameters of countries characterizing Information Society. The sources are either international organizations, large NGOs or universities. Bias-free rating does not exist but if biases exist they are certainly marginal. The selection bias is now extremely low as we are closer to exhaustivity in the set of micro-indicators.
CONTENT	5→8	There are only 13 micro-indicators to build this indicator and 11 of them derive from Wikimedia. Repartition of web content by language is a hidden continent of the Internet and existing sources are, first, extremely scarce and, second, highly biased. Unfortunately, the actual stage of the model does not escape to that situation. As it relies strongly in Wikimedia excellent statistics it carries the biases of Wikimedia where the presence of Asian languages is way below their proportion in the Internet. Obviously, the selection bias in that case, which is hugely dependent on Wikimedia stats, is extremely important. A weighting system has been put in place to reduce that dependence as much as possible (which in any case is certainly not enough, this is why the rating has been upgraded from a very low 5 to an insufficient 8). The bias proper to *content* indicator is not only important but quite **sensitive** (meaning that small variations may produce strong impacts in results) as we could experiment playing with the weighting method and the Wikipedia formula we designed (see below). Some ideas to try to remediate that issue will be implemented in the next measurement campaign. Meanwhile biases are overcome "by hand" using some technics (see Bias correction).
TRAFFIC	11	This indicator derives from the measurement of traffic by country using Alexa.com on a selection of 338 websites. In 2017 the bias analysis showed that this source was strongly biased disfavoring Asian countries and Brazil. In 2021, it appears that the bias against Asian country has been corrected

		(may be too much in case of India!) and new biases are detected disfavoring now European countries. The selection bias is obvious in that case and the next release will increase seriously the number of websites measured. The possibility to fusion in even proportion the results of Semrush and Alexa needs to be explored in order to contain the existing biases.
INTERFACES	19	Those are objective data (presence or not of a language in the interface or as a target for translation). The selection bias may exist and we may need to extend the list but its impact is marginal. Intuitively it is perceived an increase, compared to 2017, of the number of languages supported in interfaces or translation; however, this remains a "radical indicator" which leaves out the great majority of world languages and concentrate in a very subset.
USAGE	12	This indicator relies mainly on data of subscription to social networks by country. While the data collected can be considered as reliable, the method implies a bias disfavoring non-occidental country having alternate applications to Facebook, Twitter, Linkedin, etc. The next measurement campaign will try to identify the alternate applications subscriber populations to balance the results and try to reduce the bias. Meanwhile bias correction has to be made by hand. The selection bias does not really exist as the selection is dictated by the narrowness of the existing options. Next release will benefit of a small budget for not toll-free data base which will allow extending somehow the number of micro-indicators.

If the confidence weighting shown in that table is applied to the results in the building of the *power* macro-indicator (weighted average instead of simple average), in order to acknowledge the relative trust of the different indicators into the model, some changes are obtained to the results which are to be compared with the previous one (on the right of the table) and help understand the effect of the biases.

Table 13 : Macro Indicators for the top 15 languages after weighting indicators

	POWER	Capac.	Grad.	POWER	Capac.	Grad.	Effect
English	24,23%	1,86	1,58	26,48%	2,04	1,73	---
Chinese	15,77%	1,07	0,89	13,92%	0,95	0,79	+++
Spanish	8,80%	1,68	1,26	8,73%	1,67	1,25	+
Hindi	3,63%	0,63	0,85	3,38%	0,58	0,79	+++
French	3,62%	1,40	1,21	3,71%	1,44	1,24	-
Portuguese	3,37%	1,36	1,10	3,35%	1,35	1,10	+
Arabic	3,28%	0,93	0,85	3,09%	0,88	0,80	++
Russian	3,24%	1,30	0,92	3,11%	1,25	0,88	++
German	2,72%	2,08	1,30	2,86%	2,19	1,37	--
Japanese	2,51%	2,06	1,22	2,52%	2,07	1,22	
Malay	1,87%	0,79	0,85	1,76%	0,75	0,80	++
Turkish	1,27%	1,49	1,05	1,24%	1,46	1,02	+

Italian	1,23%	1,88	1,36	1,37%	2,09	1,51	--
Korean	0,97%	1,24	1,04	0,96%	1,22	1,03	
Bengali	0,91%	0,35	0,79	0,88%	0,34	0,78	+

5.3.1 Wikimedia biases

Wikipedia statistics are impeccable; however, it shall be understood that, in spite of being one of the most global Internet applications, it shows figures for some Asian languages which are much below their relative presences in the Internet. The following table compares the ratios between number of Wikipedia articles and number of Internet users; huge variance with abnormally low values for most Asian languages appear.

Table 14: Sorted by number of Wikipedia articles

Language	Articles	% TOTAL ART.	Weighted %	Art./L1+L2
English	6332139	12,92%	0,28%	7
Cebuano	5853095	11,94%	22,16%	851
Swedish	3050759	6,22%	14,11%	250
German	2593827	5,29%	1,16%	22
Arabic	2433772	4,97%	0,40%	11
French	2342875	4,78%	0,53%	14
Dutch	2060512	4,20%	5,10%	92
Chinese	1752600	3,58%	0,07%	2
Russian	1736736	3,54%	0,41%	9
Italian	1703284	3,47%	1,51%	33
Spanish	1698331	3,46%	0,19%	4
Serbo-Croatian	1514114	3,09%	4,57%	97
Polish	1480982	3,02%	2,20%	45
Japanese	1277204	2,61%	0,61%	11
Vietnamese	1266628	2,58%	1,00%	24
Ukrainian	1100281	2,24%	2,00%	52
Portuguese	1067241	2,18%	0,25%	6
Malay	936876	1,91%	0,23%	8
Persian	816984	1,67%	0,59%	15
Korean	543656	1,11%	0,40%	10
Finnish	512026	1,04%	5,36%	99
Hungarian	489514	1,00%	2,36%	49
Czech	484445	0,99%	2,18%	44
Romanian	421153	0,86%	1,06%	23
Armenian	420677	0,86%	6,60%	156
Azerbaijani	420677	0,86%	1,06%	24
Turkish	410954	0,84%	0,28%	6
Tatar	299494	0,61%	3,42%	73
Hebrew	298053	0,61%	1,92%	37
Belarusian	281379	0,57%	4,34%	87
Bulgarian	273163	0,56%	2,00%	47
Danish	267641	0,55%	2,88%	49
Slovak	237210	0,48%	1,98%	40
Kazakh	228493	0,47%	1,05%	23

Greek	195481	0,40%	0,89%	19
Urdu	164062	0,33%	0,04%	3
Hindi	148545	0,30%	0,01%	1
Uzbek	140894	0,29%	0,25%	9
Tamil	138490	0,28%	0,10%	4
Thai	137351	0,28%	0,14%	3
Bengali	109438	0,22%	0,02%	2

To be noticed, the presence of Cebuano from Philippines in second position, the relative presence of Chinese and languages from India. It is useful to check a weighted percentage in function of the number of L1+L2 speakers: English does not appear disproportionate and some languages appear to have a strong presence compared to their L1+L2 population, by order of importance: Cebuano, Swedish, Armenian, Finnish, Dutch, Serbo-Croatian Macro, Belarusian, and Tatar, for the first ones.

Wikimedia is probably at the same time the cyberplace with the major linguistic diversity and the only one which systematically provides reliable and clear linguistic statistics on all its activities. Adding the central importance of its function in the Web and its focus on openness, no doubt it is an uncontainable indicator when *contents* are discussed. Unfortunately, serious analysis shows that in no way it could reflect a close indication of what we are looking for: the repartition of *contents* by language. The importance of languages in Wikimedia is not always related to their real importance in cyberspace and some languages have invest heavily this cyberplace, independently of their overall presence in the Web. This is clearly visible across the various Wikimedia indicators we have collected hereafter, showing the first positions.

As explained before, the number of articles is not an excellent indicator because, for some languages, bots have been implemented which have created articles from translation which later are not maintained. In order to control that, one has to pay attention to the number of active editors, the number of edits during a given year and the depth, an indicator created to reflect the degree of actualization of articles. A formula has been elaborated to integrate those factors and presented previously. The results sorted by this formula and presented in percentage are the following:

Table 15: Wikipedia articles sorted by formula

English	53,96%
Hebrew	10,32%
Italian	6,99%
Serbo-Croatian	6,02%
German	5,70%
French	2,96%
Swedish	2,55%
Finnish	2,39%
Dutch	1,54%
Japanese	0,97%
Polish	0,87%

Armenian	0,84%
Hungarian	0,77%
Czech	0,63%
Ukrainian	0,54%
Danish	0,50%
Russian	0,48%
Persian	0,40%
Belarusian	0,29%
Spanish	0,29%
Portuguese	0,18%
Arabic	0,16%
Romanian	0,13%
Bulgarian	0,11%
Korean	0,10%
Turkish	0,10%
Greek	0,07%
Slovak	0,04%
Cebuano	0,03%
Azerbaijani	0,02%
Malay	0,02%
Thai	0,01%
Chinese	0,01%
Malayalam	0,00%
Kazakh	0,00%
Afrikaans	0,00%
Tatar	0,00%
Bengali	0,00%
Mongolian	0,00%
Tagalog	0,00%

This is clearly a fairer representation of the reality with Wikipedia, paying balanced attention to the number of editors, edits and depths, then weighted in function of the number of speakers L1+L2. To be noted that Cebuano is penalized now for its policy of using bots but another language from Philippines is getting its way to the top: Tagalog! The predominance of English on Wikimedia appears also more clearly with this approach.

There is more in Wikimedia than Wikipedia and stats exist also for each of the other indicators: WikiBooks, WikiQuote, WikiSource, Wikiversity, Wiktionnary, WikiNews and WikiVoyages for which the number of articles per language is accessible. For those elements of Wikimedia, the sources are presented in absolute, without weighting by function of the number of speakers, showing only the top ones.

Table 16: Number of Wikibooks

English	3851195	35,72%
German	961696	8,92%
French	657991	6,10%
Portuguese	473196	4,39%
Italian	411671	3,82%
Polish	403336	3,74%
Hungarian	401256	3,72%
Spanish	396546	3,68%
Dutch	349987	3,25%
Vietnamese	256386	2,38%
Russian	205469	1,91%
Japanese	178783	1,66%
Arabic	174452	1,62%
Hebrew	164355	1,52%
Chinese	141302	1,31%
Finnish	131314	1,22%
Persian	112964	1,05%
Malay	89019	0,83%
Hindi	73969	0,69%

Table 17: Number of Quotes

English	33897	14,28%
Italian	30799	12,98%
Polish	28960	12,20%
Russian	13148	5,54%
Czech	9263	3,90%
Persian	8495	3,58%
German	7879	3,32%
Portuguese	7443	3,14%
Spanish	7116	3,00%
Serbo-Croatian	7022	2,96%
French	5923	2,50%
Ukrainian	5798	2,44%
Slovak	4547	1,92%
Turkish	4503	1,90%
Bulgarian	4389	1,85%
Hebrew	4202	1,77%

Table 18: Number of Wikisources

French	2609546	25,3%
English	2204231	21,3%
Chinese	778716	7,5%
Bengali	722295	7,0%
Polish	669381	6,5%
Russian	642705	6,2%

German	431714	4,2%
Italian	415032	4,0%
Tamil	411502	4,0%
Hebrew	214947	2,1%
Swedish	84882	0,8%
Arabic	80708	0,8%
Multilingual Wikisource	78809	0,8%
Armenian	75487	0,7%
Portuguese	73139	0,7%

Table 19: Number of Wikiversity

German	49011	36,9%
English	38612	29,0%
French	17553	13,2%
Russian	5883	4,4%
Czech	5195	3,9%
Portuguese	4692	3,5%
Italian	4472	3,4%
Spanish	2662	2,0%
Finnish	1914	1,4%
Slovene	1252	0,9%
Swedish	858	0,6%
Greek	644	0,5%
Japanese	207	0,2%

Table 20: Number of Wiktionnary entries

English	5923218	19,2%
Malagasy	5466228	17,7%
French	3392407	11,0%
Chinese	1239843	4,0%
Serbo-Croatian	1177979	3,8%
Russian	1002462	3,2%
Spanish	885649	2,9%
German	737337	2,4%
Dutch	686499	2,2%
Swedish	674872	2,2%
Polish	649612	2,1%
Kurdish	635201	2,1%
Lithuanian	616313	2,0%
Greek	462897	1,5%
Italian	434058	1,4%
Korean	398737	1,3%
Finnish	374056	1,2%

It is important to try to understand what happened with Malagasy and wonder if its abnormal third ranking in the *gradient* macro-indicator invalids the method. This language ranks second in this micro-indicator and shows a hugely disproportionate 17% of entries compare to its population (18 million speakers) and still much more to its very low number of connect speakers (1.8 million). Even though the weight of this micro-indicator has been set to 0.1 (the same as all the Wikimedia's except Wikipedia formula and Wikibooks) the disproportion is so giant it does affect a weighted average with only 9 elements and in cascade the *power* and *gradient* macro-indicators. This situation does not delegitimate the definition of g*radient* but it is indeed a symptom of the weakness of the *content* indicator.

Table 21: Number of Wikinews

English	21687	14,9%
French	20761	14,3%
Russian	17649	12,1%
Polish	14357	9,9%
Spanish	11312	7,8%
Chinese	8559	5,9%
Arabic	7578	5,2%
Serbo-Croatian	5650	3,9%
Czech	5608	3,9%
Catalan	4056	2,8%
Tamil	3363	2,3%
Swedish	3317	2,3%
Greek	3084	2,1%
Ukrainian	1738	1,2%
Romanian	1697	1,2%
Persian	1645	1,1%
Bulgarian	1562	1,1%
Portuguese	1474	1,0%
German	1386	1,0%

Table 22: Number of articles in Wikivoyages

English	28852	28,1%
German	16545	16,1%
Persian	8674	8,5%
Italian	7619	7,4%
French	7407	7,2%
Polish	6946	6,8%
Russian	5438	5,3%
Dutch	3671	3,6%
Portuguese	3624	3,5%
Chinese	2972	2,9%
Spanish	2524	2,5%
Hebrew	2072	2,0%
Vietnamese	1624	1,6%
Swedish	1522	1,5%

Greek	1408	1,4%
Romanian	917	0,9%
Ukrainian	779	0,8%

The diversity of results depending on each subject prevents to make a systematic conclusion from the analysis of those figures, however some general statements could be made:

- English trusts, generally, but not always, the first place, although the proportion of English is less predominant than in Wikipedia, and remains in the range 14% - 36%, averaging 23.5% (versus 29.4% in Wikipedia indicators)[31].
- French and German score high in most of Wikimedia indicators.
- Chinese, Hindi, Bengali and Persian make their way in some of the indicators
- Some unexpected languages appear in top positions for some of the indicators: Malagasy and Tamil (besides Cebuano).

In conclusion, Wikimedia remains, from far, the more linguistic diverse place of the Internet with some unexpected languages managing to score high but it hardly reflects the real diversity of contents in the Web. English is largely predominant but not as much it used to be. In any case, the method needs as a next priority to enhance the quality of the *content* indicator.

5.3.2 Alexa Biases

The following table shows the different test and comparisons realized between Alexa and Semrush and, for Alexa, between the two years of use (2017 and 2021). For Alexa 2017, the previous traffic figures has been inserted in the 2021 model to make a fair comparison. The comparison is not made from the input (per country) but from the model outputs (per language); in other terms the comparison is made with the product of the model inserting each of the respective traffic figures.

Table 23: Comparisons of different traffic measurements

	SEMRUSH 2021	ALEXA 2021	2021 (S-A)/S	ALEXA 2017	A21-A17/A21
English	52,50%	35,83%	32%	45,40%	-27%
Chinese	1,88%	7,67%	-308%	4,94%	36%
Spanish	14,45%	10,14%	30%	7,53%	26%
French	4,48%	2,56%	43%	6,35%	-148%
Russian	1,88%	1,83%	3%	1,68%	8%
German	2,61%	1,33%	49%	2,94%	-122%
Portuguese	2,18%	1,46%	33%	1,63%	-12%
Arabic	1,02%	2,51%	-145%	2,54%	-1%
Hindi	1,26%	5,37%	-327%	1,60%	70%
Japanese	0,65%	1,94%	-198%	1,90%	2%
Malay	0,68%	0,98%	-44%	1,23%	-27%
Italian	0,89%	0,53%	41%	0,91%	-72%
Turkish	0,60%	1,03%	-74%		
Polish	0,47%	0,31%	34%	0,63%	-100%

[31] Those percentages refer to the number of items for English divided by the total number.

Korean	0,50%	0,90%	-78%	0,72%	20%
REST	13,95%	25,34%	-82%	18,99%	25%
TOTAL	100,00%	100,00%	0%	100,00%	0%

The comparisons highlight (in red in the table) numerous anomalies.

1) Clearly Semrush and Alexa do not reflect the same repartition of traffic per country for the same set of websites, not even close in too many cases. In the other hand, Semrush seems to ignore Asian and Arabic countries.
2) Comparing Alexa results from 2017 to 2021, one would expect not too extreme changes. This is not the case for French, German, Italian and Polish which figures drop in a suspicious manner, a confirmation of the feeling obtained during the measurement about European traffic being underestimated.

Finally, those comparisons tend to confirm the first impressions in using Alexa2021 and will be used at the time of biases correction:

- English, Spanish, Hindi may be overestimated
- French, German, Italian and Polish results looks quite underestimated
- Portuguese and Malay looks underestimated

For the next edition some attention needs to be given to this indicator to try to overcome the situation, maybe a fusion of the existing services data could be an alternative to compensate the biases?

5.4 BIAS CORRECTION

At this stage, there is no intention to apply bias correction to all the languages of the study and the focus in only on the 15 first languages in terms of *power*.

There is a method which has been used in 2017 to produce an estimation of the percentage of contents based on the coherence of the productivity factor (ratio content over connected population) for each language considered and, very important, for the rest of languages. This method applied in 2021 leads to the following rough estimation:

Table 24: Bias correction first method

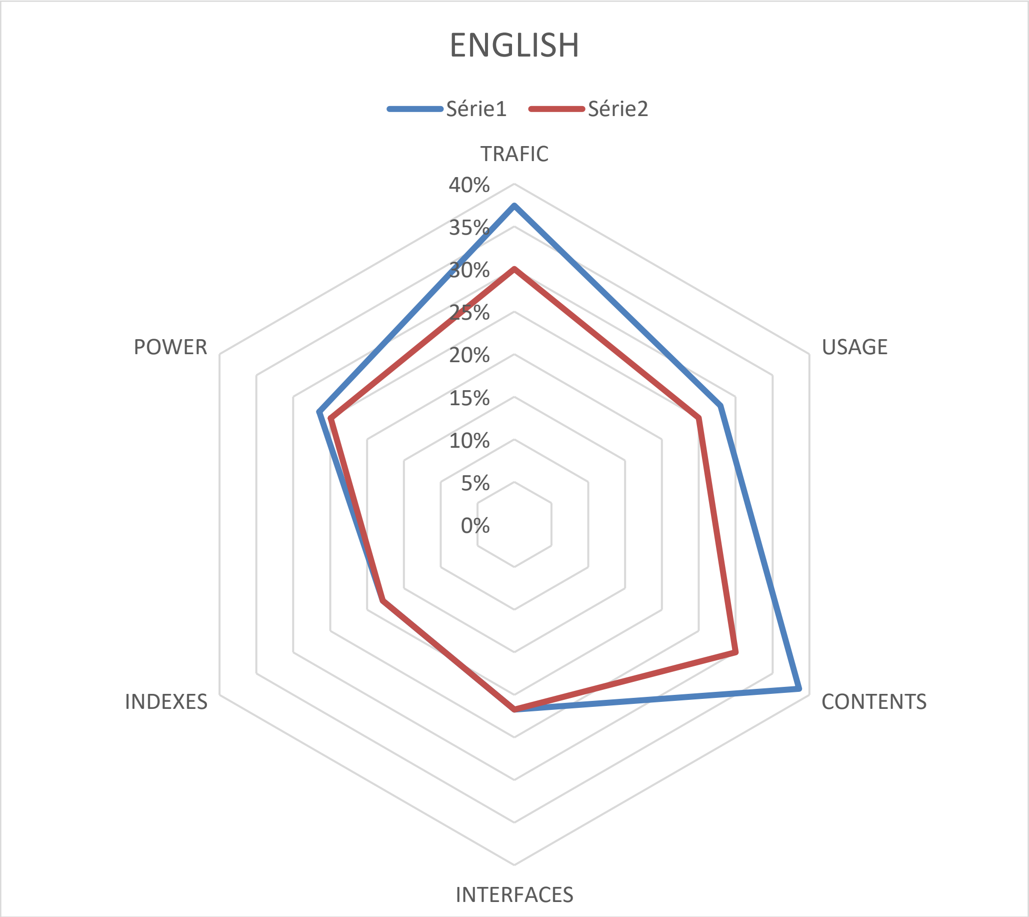
LANG.	CONTENTS	PRODUCTIVITY
English	**25,00%**	1,92
Chinese	**15,00%**	1,02
Spanish	**7,00%**	1,34
French	**4,00%**	1,55
Hindi	**4,00%**	0,69
Portuguese	**3,50%**	1,41
Russian	**3,50%**	1,41
Arabic	**2,50%**	0,71

German	**2,50%**	1,92
Japanese	**2,50%**	2,05
Malay	**1,80%**	0,76
Italian	**1,40%**	2,14
Turkish	**1,20%**	1,41
Korean	**1,20%**	1,53
Bengali	**1,20%**	0,46
Vietnamese	**0,70%**	0,94
RESTE	**23,00%**	**0,58**

This time a new approach to bias correction has been added, working specifically and directly on the respective biases of each indicator, as commented here-before. The scheme of the result on the language is examined, indicator by indicator, at the light of what we know about biases, and a new possible figure is consigned. From that a new "power" figure is computed with round values.

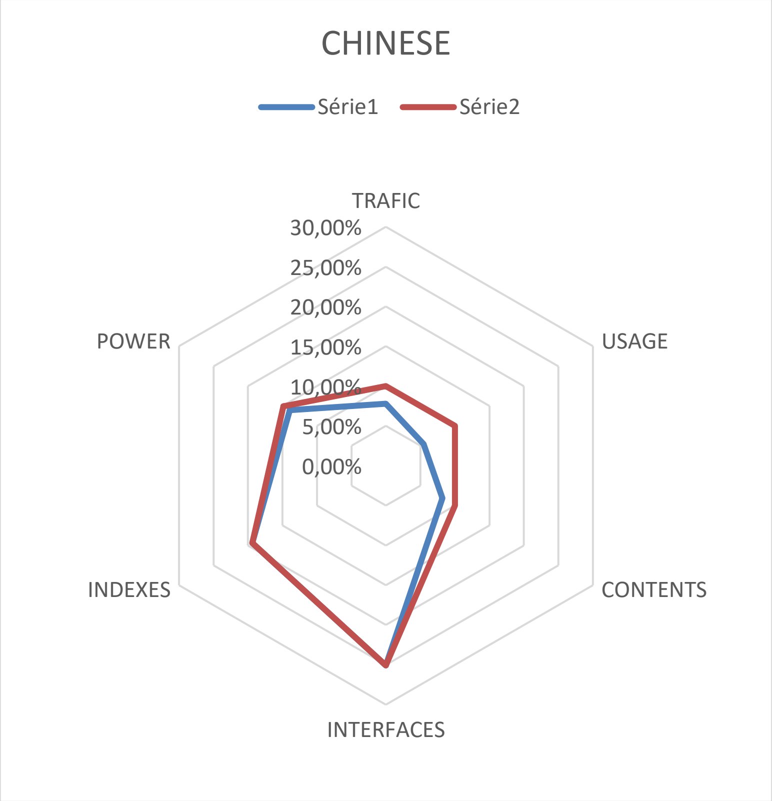
Table 25: Bias correction 2nd method

English	**TRAFIC**	**USAGE**	**CONTENTS**	**INTERFACES**	**INDEXES**	**POWER**
MODEL	0,3744	0,2792	0,3861	0,2173	0,1787	**0,2648**
BIAS CORRECTION	30%	25%	30%	22%	18%	25%

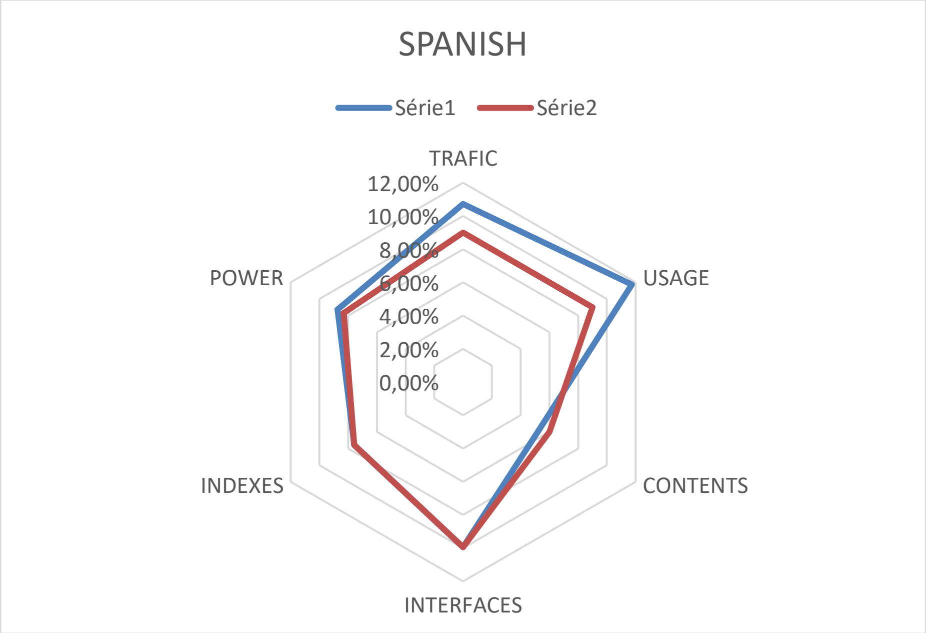


Chinese	**TRAFIC**	**USAGE**	**CONTENTS**	**INTERFACES**	**INDEXES**	**POWER**

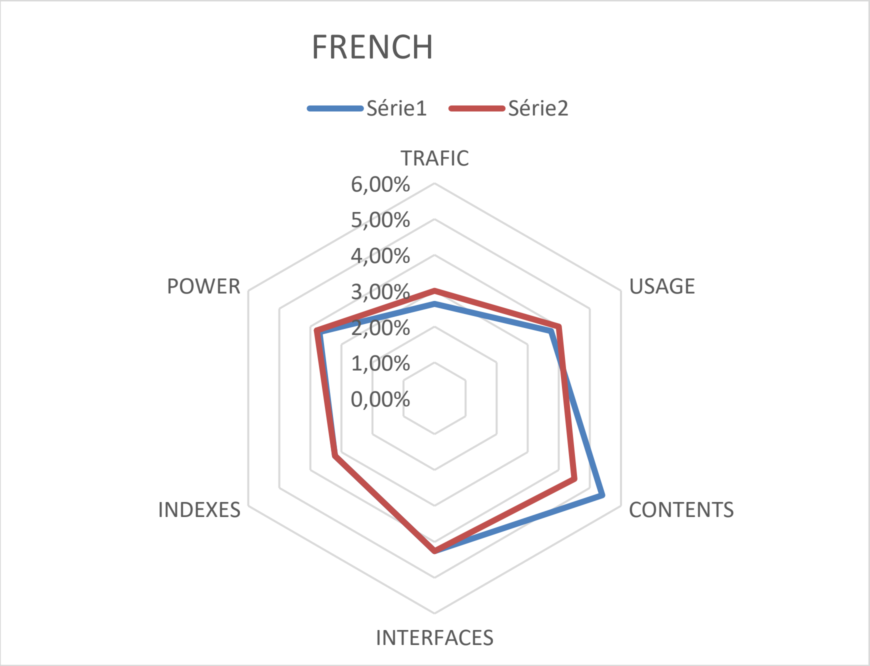
MODEL	7,79%	5,47%	8,18%	25,07%	19,38%	**13,92%**
BIAS CORRECTION	10%	10%	10%	25%	19%	15%



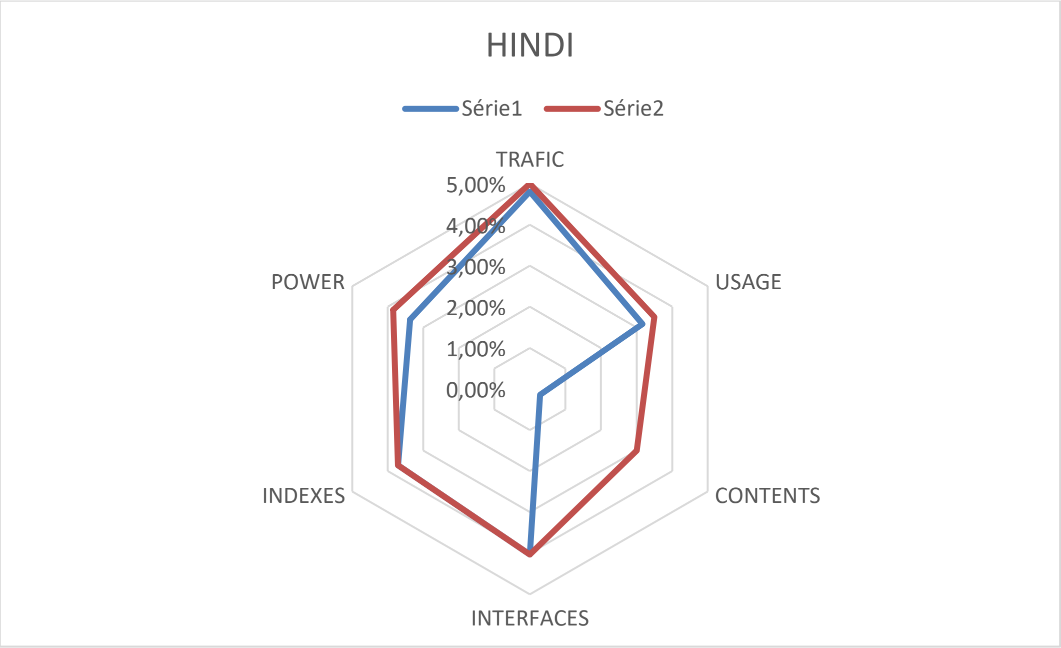
Spanish	**TRAFIC**	**USAGE**	**CONTENTS**	**INTERFACES**	**INDEXES**	**POWER**
MODEL	10,72%	11,74%	5,42%	9,94%	7,59%	**8,73%**
BIAS CORRECTION	9%	9%	6%	10%	8%	8%



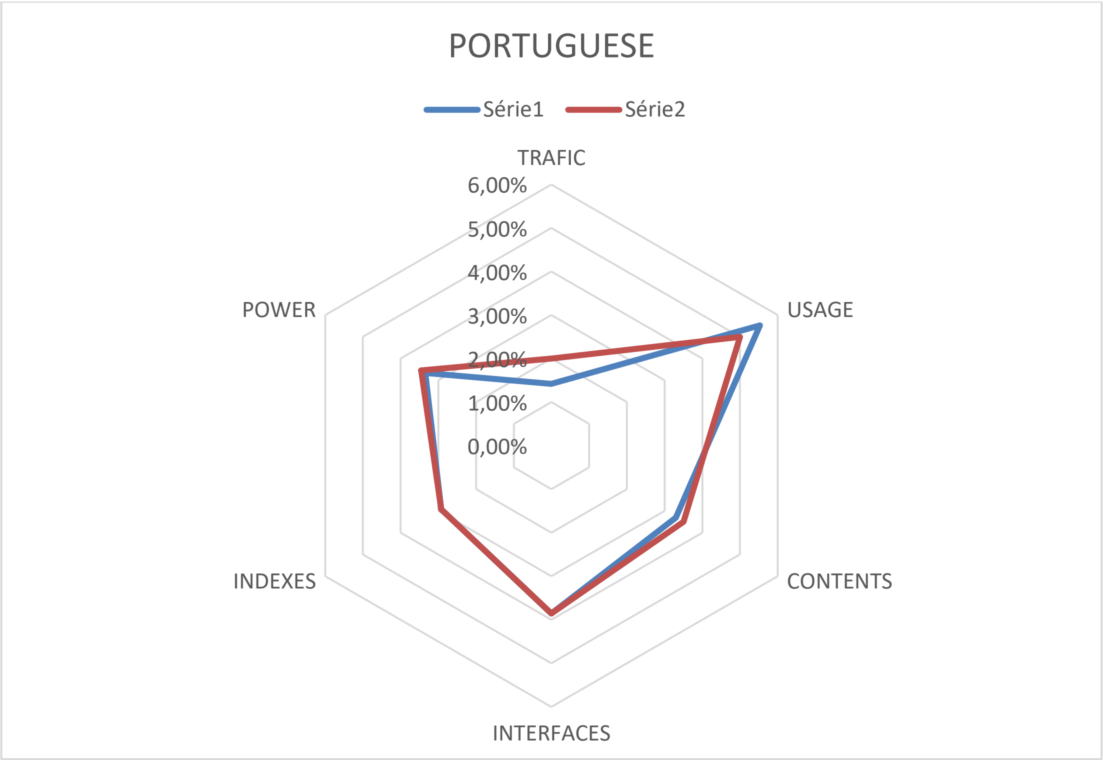
French	TRAFIC	USAGE	CONTENTS	INTERFACES	INDEXES	POWER
MODEL	2,64%	3,75%	5,40%	4,26%	3,21%	3,71%
BIAS CORRECTION	3,0%	4,0%	4,5%	4,3%	3,2%	3,8%



Hindi	TRAFIC	USAGE	CONTENTS	INTERFACES	INDEXES	POWER
MODEL	4,81%	3,16%	0,28%	4,03%	3,71%	3,38%
BIAS CORRECTION	5,0%	3,5%	3,0%	4,0%	3,7%	3,8%



Portuguese	TRAFIC	USAGE	CONTENTS	INTERFACES	INDEXES	POWER
MODEL	1,42%	5,53%	3,30%	3,85%	2,92%	3,35%
BIAS CORRECTION	2,0%	5,5%	3%	3,9%	2,9%	3,5%



The result of this bias correction exercise is presented here after and compared with the results from the first method of correction:

Table 26: Bias correction results

	SECOND METHOD		FIRST
	POWER	CONTENT	METHOD
English	25%	30,0%	25%
Chinese	15%	10%	15%
Spanish	8%	6%	7%
French	3.8%	4.5%	4%
Hindi	3.8%	3.0%	4%
Portuguese	3.5%	2.8%	3.5%

Interestingly, the results from the two different methods are quite close.

6. CONCLUSIONS AND PERSPECTIVES

This second version of the method to produce indicators of the presence of languages on the Internet show some interesting enhancements, especially in demo-linguistic data more reliable and in the process of L2. It also makes a coherent move on the process of establishing world percentage related to the total number of L1+L2 speakers and presents now an *index* indicator more complete. The method has upgraded the analysis of the biases produced by using systematically Wikimedia statistics and present two complementary ways to compensate those biases.

The method encounters however new challenges with the behavior of *traffic* measurement tools, with the *content* indicator still too dependent on Wikimedia figures, and clearly no reflecting correctly the reality, and with the fact that ITU does not provide any more estimates for the percentage of persons connected to the Internet per country (with a particular issue about the exact percentage for India).

It is forecasted a new version before the end of 2021 which will try to address those challenges and try to enlarge the number of languages treated, pushing the boundary to languages with more than 1 million L1 speakers. The objective of the future release will also be to extend the number of websites measured in terms of traffic so to be able to provide more accurate and trustable differentiate results for some given languages by themes.

As for the results, the trend of relative reduction of the dominance of English continues with now an estimated presence around 25% (versus 30% in 2017), the growth of Chinese and the appearance of Hindi as a probable fourth language of the Internet, together with French today, and probably above French in the coming years.

REFERENCES

[1] D. Pimienta, “An alternative approach to produce indicators of languages in the Internet”, 2017
http://funredes.org/lc2017/Alernative%20Languages%20Internet.docx

[2] - MAAYA, “NET.LANG: Towards a multilingual cyberspace”, C&F Editions, 2012 –
http://net-lang.net/lang_en

[3] - D. Pimienta, D. Prado, Á. Blanco, “Twelve years of measuring linguistic diversity in the Internet: balance and perspectives”, UNESCO, 2009 –
http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

[4] - J. Paolillo, D. Pimienta, D. Prado, et al., “Measuring linguistic diversity on the Internet”, UNESCO,/2005-
http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-list/measuring-linguistic-diversity-on-the-internet/

[5] – D. Pimienta "Indicators of Languages in the Internet", International Conference Language Technologies for All (LT4All), 4 - 6 December 2019, UNESCO, Paris
http://funredes.org/lc2019/Indicators%20Language%20Internet.pdf

ANNEX 1. LIST OF MICRO INDICATORS AND SOURCES

MICRO-INDICATOR	TYPE	THEME	URL OF SOURCE
Amazon US - number of books 2017	CONTENT	Book	Retaken from 2017
Value of Wikipedia depth	CONTENT	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Number of active Wikipedia users	CONTENT	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Number of Wikipedia edits	CONTENT	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
Number of Wiki Books per language	CONTENT	Book	https://meta.wikimedia.org/wiki/Wikibooks/Table
Number of Wikipedia article by language	CONTENT	Ency	https://meta.wikimedia.org/wiki/List_of_Wikipedias
WikiQuote articles per language	CONTENT	Book	https://stats.wikimedia.org/wikiquote/FR/Sitemap.htm
Number of WikiSource articles per language	CONTENT	Book	https://stats.wikimedia.org/wiktibooks/EN/Sitemap.htm
Number of articles Wikiversity per language	CONTENT	S/T	https://stats.wikimedia.org/wikiversity/EN/Sitemap.htm
Number of articles Wiktionnary per language	CONTENT	Dict	https://stats.wikimedia.org/wiktionary/EN/Sitemap.htm
Number of articles WikiNews per language	CONTENT	News	https://stats.wikimedia.org/wikinews/EN/Sitemap.htm
Number of articles WikiVoyages per language	CONTENT	Tur	https://stats.wikimedia.org/wikivoyage/EN/Sitemap.htm
T-Index for e-commerce Projection 2021	CONTENT	e.com	https://translated.com/les-langues-qui-comptent
E-Government Index	INDEX	S/T	https://publicadministration.un.org/egovkb/Data-Center
E-Participation Index	INDEX	S/T	https://publicadministration.un.org/egovkb/Data-Center
Online Service Index	INDEX	Infra	https://publicadministration.un.org/egovkb/Data-Center
Human Capital Index	INDEX	ICT	https://publicadministration.un.org/egovkb/Data-Center
Telecommunication Infrastructure Index	INDEX	Gov	https://publicadministration.un.org/egovkb/Data-Center
Cisco Global Digital Readiness Index 2019	INDEX	S/T	https://www.cisco.com/c/dam/en_us/about/csr/reports/global-digital-readiness-index.pdf
Government AI Readiness Index 2020	INDEX	ICT	https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf
Internet Freedom Scores	INDEX	Book	https://freedomhouse.org/countries/freedom-net/scores
Global Connectivity Index	INDEX	Gov	https://www.huawei.com/minisite/gci/en/country-rankings.html
Global Cybersecurity Index 2018	INDEX	Gov	https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf
UNCTAD B2C E-commerce index, 2020	INDEX	Gov	https://unctad.org/system/files/official-document/tn_unctad_ict4d17_en.pdf
The Global Open Data Index	INDEX	Infra	https://index.okfn.org/place/
World Digital Competitiveness Ranking 2020	INDEX	Secu	https://www.imd.org/globalassets/wcc/docs/release-2020/digital/digital_2020.pdf
Readiness For Frontier Technologies Index	INDEX	Econ	https://unctad.org/system/files/official-document/tir2020_en.pdf
Global Innovation Index	INDEX	AI	https://wipo.int/edocs/pubdocs/en/wipo_pub_gii_2020.pdf
Access to Basic Knowledge	INDEX	Econ	https://legacy.socialprogress.org/assets/downloads/2011-2020-Social-Progress-Index.xlsx
Access to Information and Communications	INDEX	Gov	“ ”
Access to Advanced Education	INDEX	Gov	“ ”
Access to electricity (% of pop.)	INDEX	Infra	“ ”
Access to quality education (0=unequal; 4=equal)	INDEX	S/T	“ ”
Access to online governance (0=low; 1=high)	INDEX	Econ	“ ”
Media censorship (0=frequent; 4=rare)	INDEX	Infra	“ ”
Freedom of expression (0=no freedom; 1=full freedom)	INDEX	Gov	“ ”
Quality weighted universities (points)	INDEX	e.com	“ ”
Citable documents	INDEX	Gov	“ ”
Women with advanced education	INDEX	Econ	“ ”

Years of tertiary schooling	INDEX	S/T	“ ”
Translation languages of Bing Translator	INTERFACE	Tra	https://www.bing.com/translator/
Amazon Kindle direct Publishing supported languages	INTERFACE	Inter	https://kdp.amazon.com/en_US/help/topic/G200673300
Languages supported by Cortana	INTERFACE	Tra	https://en.wikipedia.org/wiki/Cortana
Word Reference languages supported	INTERFACE	Inter	https://www.wordreference.com
WordLingo Translation languages	INTERFACE	Inter	http://www.worldlingo.com/en/languages/
Facebook supported languages	INTERFACE	Tra	https://www.facebook.com/language.php
Facebook In-Stream Ads languages supported	INTERFACE	Tra	https://www.facebook.com/business/help/267128784014981
Free Translator languages supported	INTERFACE	Tra	http://www.free-translator.com
Google Play Console supported languages	INTERFACE	Tra	https://support.google.com/googleplay/android-developer/table/4419860?hl=en
Google Cloud supported languages	INTERFACE	Inter	https://cloud.google.com/translate/docs/languages?hl=en
Google Translate supported languages	INTERFACE	Inter	https://en.wikipedia.org/wiki/Google_Translate
Google Scholar supported languages for search	INTERFACE	Inter	https://scholar.google.com/scholar_settings?sciifh=1&hl=en&as_sdt=0,5#1
Language supported by Paralink Translator	INTERFACE	Inter	http://paralink.com
Online Translator languages supported	INTERFACE	Tra	https://www.online-translator.com/traduction
Reverso translator languages supported	INTERFACE	Tra	https://www.reverso.net/text_translation.aspx?lang=EN
Free Translation supported languages	INTERFACE	Tra	https://www.freetranslations.org
Skype Supported languages		Tra	https://support.skype.com/en/faq/FA34781/what-languages-are-supported-in-skype
Systran translate supported languages		Tra	https://support.systran.net/systranlinks/faq/
163.com	TRAFFIC	GAM	https://www.alexa.com/siteinfo
17ok.com	TRAFFIC	?	https://www.alexa.com/siteinfo
1and1.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
360.cn	TRAFFIC	Secu	https://www.alexa.com/siteinfo
4shared.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
500px.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
6.cn	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
A2hosting.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Abilogic.com	TRAFFIC	DIR	https://www.alexa.com/siteinfo
About.me	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Academia.edu	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Adam4Adam.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Adictinggames.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
adobe.com	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Adultfriendfinder.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Aim.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Alexa.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Aliexpress.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Alipay.com	TRAFFIC	Econ	https://www.alexa.com/siteinfo
Alivedirectory.com	TRAFFIC	DIR	https://www.alexa.com/siteinfo
Amazon.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Amazonaws.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
Anastasiadate.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Android	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Angel.co	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Anobii.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Answers.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Aparat.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Apple	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Apple music	TRAFFIC	SN-Mu	https://www.alexa.com/siteinfo
Apple.com/Safari	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Archive.org	TRAFFIC	Book	https://www.alexa.com/siteinfo
Archives-ouvertes.fr	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Armorgames.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Arvixe.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
Arxiv.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Ashleymadison.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Ask.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Ask.fm	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Atom.io	TRAFFIC	App	https://www.alexa.com/siteinfo

Avvo.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Babytree.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Badoo.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Baidu.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Bandcamp.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Bartleby.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Base-search.net	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Bet365.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Beyond.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Bilibili.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Bing.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Bit.ly	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Bitbucket.org	TRAFFIC	App	https://www.alexa.com/siteinfo
Bitcoin.com	TRAFFIC	Econ	https://www.alexa.com/siteinfo
Bitshare.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Bl.uk	TRAFFIC	Book	https://www.alexa.com/siteinfo
Blackle.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Blog.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Blogadda.com/	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Blogcatalog.com/	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Blogger.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Blogspot.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Bluehost.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
Blurtit.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Bnf.fr	TRAFFIC	Book	https://www.alexa.com/siteinfo
Bongacams.com	TRAFFIC	Porn	https://www.alexa.com/siteinfo
booking.com	TRAFFIC	Tur	https://www.alexa.com/siteinfo
Books.google.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Box.com	TRAFFIC	App	https://www.alexa.com/siteinfo
Brackets.io	TRAFFIC	App	https://www.alexa.com/siteinfo
Business.com	TRAFFIC	DIR	https://www.alexa.com/siteinfo
Busuu.com	TRAFFIC	EDU	https://www.alexa.com/siteinfo
C9.io	TRAFFIC	Cloud	https://www.alexa.com/siteinfo
Cafemom.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Cairn.info	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Canva.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Care2.com	TRAFFIC	Advo	https://www.alexa.com/siteinfo
Caringbridge.org	TRAFFIC	Health	https://www.alexa.com/siteinfo
Chacha.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Chaturbate.com	TRAFFIC	Porn	https://www.alexa.com/siteinfo
Chrome.com	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Classmates.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Codeanywhere.com	TRAFFIC	Cloud	https://www.alexa.com/siteinfo
Codepen.io	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Commonsensemedia.org	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Contentful.com	TRAFFIC	APP	https://www.alexa.com/siteinfo
Couchsurfing.com	TRAFFIC	Tur	https://www.alexa.com/siteinfo
Coursera	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Creativecommons.org	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Crunchyroll.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Csdn.net	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Cyworld.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Dailymotion.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Dart-europe.eu	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Daum.net	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Deezer.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Del.icio.us	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Depositfiles.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Deviantart.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Discordapp.com	TRAFFIC	App	https://www.alexa.com/siteinfo
disneyplus.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Dmoz.org	TRAFFIC	DIR	https://www.alexa.com/siteinfo
Doaj.org	TRAFFIC	DIR	https://www.alexa.com/siteinfo
Douban.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
doubleclick.net	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Draugiem.lv	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Dreamhost.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
Dreamwidth.org	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Dropbox.com	TRAFFIC	App	https://www.alexa.com/siteinfo
Drupal.org	TRAFFIC	CMS	https://www.alexa.com/siteinfo
Duckduckgo.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo

DXY.cn	TRAFFIC	Health	https://www.alexa.com/siteinfo
ebay.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Eclipse.org	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Edx.org	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Egnyte.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Eharmony.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Etoro.com	TRAFFIC	Econ	https://www.alexa.com/siteinfo
Etsy.com	TRAFFIC	Econ	https://www.alexa.com/siteinfo
Europeana.eu	TRAFFIC	Book	https://www.alexa.com/siteinfo
Exalead.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Excite.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Experienceproject.com	TRAFFIC	Dead	https://www.alexa.com/siteinfo
Fandom.com	TRAFFIC	VC	https://www.alexa.com/siteinfo
Fetlife.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Filefactory.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Fileserve.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Filmaffinity.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Filmow.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Flickr.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Flipboard.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Flixster.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
FNAC.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Force.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Fotki.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Fotolog.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Foursquare.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Fun-mooc.fr	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Funnyordie.com	TRAFFIC	Hum	https://www.alexa.com/siteinfo
Futurelearn.com	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
G2a.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Gaiaonline.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Gameblog.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Gamefaqs.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Geni.com	TRAFFIC	Gen	https://www.alexa.com/siteinfo
Gfycat.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Ghost.org	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Gigablast.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Gigasize.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Girlsaskguys.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Github.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Gmx.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Gmx.net	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Godaddy.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
GOG.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Goodreads.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Google.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Gotinder.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Gravatar.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Grindr.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Gutenberg.org	TRAFFIC	Book	https://www.alexa.com/siteinfo
Haosou.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Hathitrust.org	TRAFFIC	Book	https://www.alexa.com/siteinfo
Hi5.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Hightail.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Hostgator.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Hotmail.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Huanqiu.com	TRAFFIC	News	https://www.alexa.com/siteinfo
Hubpages.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Hulu.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Hushmail.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Ibiblio.org	TRAFFIC	Book	https://www.alexa.com/siteinfo
Icloud.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Icq.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
imdb.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Imgur.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Indiblogger.in	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Influenster.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Inmotionhosting.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
Instagram.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Iqiyi.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Isbn.org	TRAFFIC	Book	https://www.alexa.com/siteinfo

Italki.com	TRAFFIC	EDU	https://www.alexa.com/siteinfo
Itch.io	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Jasminedirectory.com	TRAFFIC	DIR	https://www.alexa.com/siteinfo
jd.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Jekyllrb.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Jetbrains.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
joinclubhouse.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Joomla.com	TRAFFIC	CMS	https://www.alexa.com/siteinfo
Journalseek.net	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Jstor.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Jurn.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Justanswer.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Kaixin001.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Kakao.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Kompas.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Kongregate.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Last.fm	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Library.harvard.edu	TRAFFIC	Book	https://www.alexa.com/siteinfo
Librarything.com	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Line.me	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Linkedin.com	TRAFFIC	SN-pr	https://www.alexa.com/siteinfo
Linux.org	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Liquidweb.com	TRAFFIC	Host	https://www.alexa.com/siteinfo
Live.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Livejasmin.com	TRAFFIC	Porn	https://www.alexa.com/siteinfo
Livejournal.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Liveleak.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Logoslibrary.eu	TRAFFIC	Book	https://www.alexa.com/siteinfo
Lycos.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Mail.aol.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Mail.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Mail.google.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Mail.ru	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Mail.yandex.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Mamba.ru	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Match.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Mediafire.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Medium.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Meetic.fr	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Meetup.com	TRAFFIC	SN-pr	https://www.alexa.com/siteinfo
Mega.io	TRAFFIC	Cloud	https://www.alexa.com/siteinfo
Mendeley.com	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Messenger.yahoo.com/	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Metacafe.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Metafilter.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Microsoft.com	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Metropoles.com	TRAFFIC	News	https://www.alexa.com/siteinfo
Microsoftonline.com	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Miniclip.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Mixi.jp	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Mocospace.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Moodle.org	TRAFFIC	CMS	https://www.alexa.com/siteinfo
Mouthshut.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Mozilla.org	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Msn.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Mubi.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Myheritage.com	TRAFFIC	Gen	https://www.alexa.com/siteinfo
Mylife.com	TRAFFIC	Dead	https://www.alexa.com/siteinfo
Myshopify.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Myspace.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Napster.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Naver.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Netcraft.com	TRAFFIC	Secu	https://www.alexa.com/siteinfo
Netflix.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Newgrounds.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Nicovideo.jp	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Ning.com	TRAFFIC	SN-pr	https://www.alexa.com/siteinfo
Notepad-plus-plus.org	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Novoed.com	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Oatd.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Odnoklassniki.ru	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo

Office.com	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Ok.ru	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Okcupid.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Okezone.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Oovoo.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Openclassrooms.com	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Opengrey.eu	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Openlibrary.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Openoffice.org	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Openthesis.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Opera.com	TRAFFIC	ICT	https://www.alexa.com/siteinfo
Origin.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Outlook.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Panda.tv	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Paypal.com	TRAFFIC	Econ	https://www.alexa.com/siteinfo
Pen.io	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Periscope.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Periscope.tv	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Photobucket.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Pikiran-rakyat.com	TRAFFIC	News	https://www.alexa.com/siteinfo
Pinterest.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Playstation.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Playstore.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Plurk.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Pornhub.com	TRAFFIC	Porn	https://www.alexa.com/siteinfo
Primevideo.com	TRAFFIC	Film	https://www.alexa.com/siteinfo
Protonmail.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Qq.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Question.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Quora.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Qwant.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Rapidshare.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Ravelry.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Reddit.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Rediff.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Rediffmail.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Renren.com	TRAFFIC	SN-Fr	https://www.alexa.com/siteinfo
Researchgate.net	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Reverbnation.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Roblox.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Rumble.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Rutube.ru	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Salesforce.com	TRAFFIC	App	https://www.alexa.com/siteinfo
Sapo.pt	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Savefrom.net	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Scielo.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Scienceopen.com	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Search.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Secondlife.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Semanticscholar.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Sharecare.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Similarweb.com	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Sina.com.cn	TRAFFIC	Port	https://www.alexa.com/siteinfo
Sitebuilder.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Skype.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Skyrock.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Slack.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Slideshare.net	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Smugmug.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Snapchat.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
so.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Socolar.com	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Sogou.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
sohu.com	TRAFFIC	Port	https://www.alexa.com/siteinfo
Somuch.com	TRAFFIC	DIR	https://www.alexa.com/siteinfo
Sony.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Soso.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Soundcloud.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo
Spaces.ru	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Spip.net	TRAFFIC	CMS	https://www.alexa.com/siteinfo
Spotify.com	TRAFFIC	SN-mu	https://www.alexa.com/siteinfo

Squarespace.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Stackexchange.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Stackoverflow.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Startpage.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Steam.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Steampowered.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Straightdope.com	TRAFFIC	Q/A	https://www.alexa.com/siteinfo
Stumbleupon.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Sublimetext.com	TRAFFIC	App	https://www.alexa.com/siteinfo
Svbtle.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Tagged.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Taobao.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Taringa.net	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Teamspeak.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Teamviewer.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Technorati.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Telegram - interface	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Telegram.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Telegram.org	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Theblogchatter.com/	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Theses.fr	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Tianya.cn	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Tiktok.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Tinyurl.com	TRAFFIC	Tool	https://www.alexa.com/siteinfo
Tmall.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Trombi.com	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Tudou.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Tuenti.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Tumblr.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Twitch.tv	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Twoo.com	TRAFFIC	SN-Da	https://www.alexa.com/siteinfo
Typepad.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Udacity.com	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Udemy.com	TRAFFIC	MOOC	https://www.alexa.com/siteinfo
Uploaded.net	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Uploading.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Veoh.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Viadeo.com	TRAFFIC	SN-pr	https://www.alexa.com/siteinfo
Viber.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Vimeo.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Vk.com	TRAFFIC	SN-Mu	https://www.alexa.com/siteinfo
Wattpad.com	TRAFFIC	SN-fr	https://www.alexa.com/siteinfo
Wayn.com	TRAFFIC	Tur	https://www.alexa.com/siteinfo
Wdl.org	TRAFFIC	Book	https://www.alexa.com/siteinfo
Webcrawler.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Webometrics.info	TRAFFIC	Mktg	https://www.alexa.com/siteinfo
Wechat.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Weebly.com	TRAFFIC	e.com	https://www.alexa.com/siteinfo
Weheartit.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Weibo.com	TRAFFIC	Blog	https://www.alexa.com/siteinfo
Wetransfer.com	TRAFFIC	FiSh	https://www.alexa.com/siteinfo
Whatsapp.com	TRAFFIC	MSG	https://www.alexa.com/siteinfo
Wistia.com	TRAFFIC	SN-Im	https://www.alexa.com/siteinfo
Wix.com	TRAFFIC	App	https://www.alexa.com/siteinfo
Wolframalpha.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Wordpress.com	TRAFFIC	CMS	https://www.alexa.com/siteinfo
Worldcat.com	TRAFFIC	Book	https://www.alexa.com/siteinfo
Worldwidescience.org	TRAFFIC	S/T	https://www.alexa.com/siteinfo
Xbox.com	TRAFFIC	Gam	https://www.alexa.com/siteinfo
Xhamster.com	TRAFFIC	Porn	https://www.alexa.com/siteinfo
Xing.com	TRAFFIC	SN-pr	https://www.alexa.com/siteinfo
Xinhuanet.com	TRAFFIC	News	https://www.alexa.com/siteinfo
Xvideos.com	TRAFFIC	Porn	https://www.alexa.com/siteinfo
yahoo.com	TRAFFIC	Mail	https://www.alexa.com/siteinfo
Yammer.com	TRAFFIC	SN-pr	https://www.alexa.com/siteinfo
Yandex.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Yelp.com	TRAFFIC	SEng	https://www.alexa.com/siteinfo
Youku.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
YouTube	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Yy.com	TRAFFIC	Vid	https://www.alexa.com/siteinfo
Zhanqi.tv	TRAFFIC	Vid	https://www.alexa.com/siteinfo

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

TYPOLOGY	QTY	THEME
?	1	
Advo	1	Advocacy
App	10	Applications
Blog	20	
Book	18	
Cloud	3	
CMS	5	Content Management System
DIR	7	Directory
e.com	9	E-Commerce
Econ	5	Economy
EDU	2	Courses
FiSh	11	File Sharing
Film	8	Movies on demand
Gam	20	Games
Gen	2	Genealogy
Health	2	Health
Host	7	Web Hosting
Hum	1	Humor
ICT	13	
Mail	17	
Mktg	10	Marketing
MOOC	8	
MSG	23	Messaging
News	4	
Porn	6	

Port	8	Portal
Q/A	13	Question/Answer
S/T	22	Science & Technology (research)
Secu	2	Security
SEng	26	Search Engine
SN-Da	20	Dating Social Networks
SN-Fr	28	Friendship Social Networks
SN-Im	24	Images Social Networks
SN-Mu	10	Music Social Networks
SN-pr	6	Professional Social Networks
Tool	14	
Tur	3	Tourism
VC	1	Virtual Community
Vid	13	Video

ANNEX 2: MACROLANGUAGES

ISO CODE	***MACRO LANGUAGES***	***NUMBER OF LANGUAGES FUSIONED***
ara	***Arabic***	***29***
aym	***Aymara***	***2***
aze	***Azerbaijani***	***3***
bal	***Balochi***	***3***
bik	***Bikol***	***8***
bnc	***Bontok***	***5***
bua	***Buriat***	***3***
chm	***Mari***	***2***
cre	***Cree***	***6***
del	***Delaware***	***2***
den	***Slavey***	***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	***Guaraní***	***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	***Komi***	***2***
kon	***Kongo***	***3***
kpe	**Kpelle**	**2**
kur	***Kurdish***	***3***
lah	***Lahnda***	***7***
lav	***Latvian***	***2***
luy	***Luyia***	***14***
man	***Mandingo***	***6***
mlg	***Malagasy***	***11***

mon	***Mongolian***	***3***
msa	***Malay***	***36***
mwr	***Marwari***	***6***
nep	***Nepali***	***2***
oji	***Ojibwa***	***7***
ori	***Oriya***	***2***
orm	***Oromo***	***4***
pus	***Pashto***	***3***
que	***Quechua***	***42***
raj	**Rajasthani**	**6**
rom	***Romani***	***6***
sqi	***Albanian***	***4***
srd	***Sardinian***	***4***
swa	***Swahili***	***2***
syr	***Syriac***	***2***
tmh	***Tamasheq***	***4***
uzb	***Uzbek***	***2***
yid	***Yiddish***	***2***
zap	***Zapotec***	***57***
zha	***Zhuang***	***16***
zho	***Chinese***	***15***
zza	***Dimli***	***2***

ANNEX 3: LIST OF COUNTRIES OR TERRITORIES WHERE ITU DOES NOT OFFER DATA

Country CODE	COUNTRY NAME	POPULATION
AX	Aland Islands	27 652
AS	American Samoa	55 990
IO	British Indian Ocean Territory	4 000
BQ	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
MQ	Martinique	377 100
NF	Norfolk Island	1 500
KP	*North Korea*	*25 579 000*
MP	Northern Mariana Islands	53 280
PW	Palau	17 550
PN	Pitcairn	36
RE	Réunion	751 580
BL	Saint Barthélemy	7 850
MF	Saint Martin	28 500
PM	Saint Pierre and Miquelon	6 340
SX	Sint Maarten	33 470
TC	Turks and Caicos Islands	30 170
VA	*Vatican State*	*330*
EH	*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 which data are included in a given country
2) There is no source nor estimates for the percentage of connected people to the Internet (in italic in the table).

ANNEX 4: RESULTS FOR ALL PROCESSED LANGUAGES

Rank			W.Connect.	W.Pop.	TRAFIC	L.Connec.	USAGE	CONT.	INTER.	INDEX	POWER	CAP.	GRAD.
	ISO	**TOTAL OR AVG---->**	**100%**	**100%**	**100%**	**54.70%**	**100%**	**100%**	**100%**	**100%**	**100%**	**0.75**	**0.74**
		Remain	10.13%	12.66%	7.90%	43.76%	8.59%	2.88%	0.02%	6.91%	6.07%	0.48	0.60
54	afr	**Afrikaans**	0.19%	0.17%	0.08%	59.75%	0.11%	0.15%	0.10%	0.17%	0.13%	0.79	0.73
102	aka	**Akan**	0.06%	0.09%	0.02%	38.80%	0.05%	0.00%	0.01%	0.05%	0.03%	0.35	0.49
60	amh	**Amharic**	0.21%	0.55%	0.09%	20.57%	0.11%	0.01%	0.12%	0.11%	0.11%	0.19	0.51
8	ara	***Arabic***	3.89%	3.53%	2.30%	60.14%	3.02%	2.05%	4.29%	3.01%	3.09%	0.88	0.80
74	asm	**Assamese**	0.11%	0.15%	0.12%	40.03%	0.08%	0.00%	0.03%	0.09%	0.07%	0.49	0.66
119	awa	**Awadhi**	0.03%	0.04%	0.03%	39.25%	0.02%	0.00%	0.00%	0.03%	0.02%	0.43	0.60
42	aze	***Azerbaijani***	0.31%	0.23%	0.26%	74.76%	0.16%	0.11%	0.17%	0.27%	0.22%	0.94	0.69
106	bal	***Balochi***	0.05%	0.09%	0.06%	30.72%	0.04%	0.00%	0.00%	0.03%	0.03%	0.36	0.63
127	bam	**Bamanankan**	0.03%	0.14%	0.01%	12.94%	0.02%	0.00%	0.00%	0.01%	0.01%	0.10	0.42
53	bar	**Bavarian**	0.22%	0.14%	0.10%	87.68%	0.17%	0.00%	0.00%	0.33%	0.14%	0.97	0.61
94	bel	**Belarusian**	0.06%	0.04%	0.02%	82.27%	0.03%	0.03%	0.03%	0.06%	0.04%	1.00	0.66
15	ben	**Bengali**	1.14%	2.58%	1.22%	24.15%	1.13%	0.26%	0.72%	0.84%	0.88%	0.34	0.78
112	bew	**Betawi**	0.04%	0.05%	0.01%	47.69%	0.05%	0.00%	0.00%	0.04%	0.02%	0.50	0.57
34	bho	**Bhojpuri**	0.37%	0.51%	0.40%	39.85%	0.27%	0.00%	0.03%	0.32%	0.23%	0.46	0.63
118	bik	***Bikol***	0.03%	0.04%	0.01%	43.03%	0.04%	0.00%	0.00%	0.03%	0.02%	0.51	0.65
109	bjj	***Kanauji***	0.04%	0.06%	0.05%	40.00%	0.03%	0.00%	0.00%	0.04%	0.03%	0.45	0.62
116	bug	***Bugis***	0.04%	0.04%	0.01%	47.94%	0.04%	0.00%	0.00%	0.03%	0.02%	0.50	0.57
63	bul	***Bulgarian***	0.10%	0.08%	0.05%	70.34%	0.08%	0.13%	0.08%	0.12%	0.09%	1.18	0.92
69	ceb	**Cebuano**	0.12%	0.15%	0.06%	43.15%	0.19%	0.00%	0.02%	0.11%	0.08%	0.54	0.69
38	ces	**Czech**	0.19%	0.13%	0.07%	81.37%	0.13%	0.50%	0.18%	0.25%	0.22%	1.70	1.14
55	dan	**Danish**	0.10%	0.05%	0.04%	97.82%	0.08%	0.26%	0.08%	0.16%	0.12%	2.19	1.22
9	deu	**German**	2.09%	1.30%	1.32%	87.65%	1.95%	5.84%	2.97%	2.98%	2.86%	2.19	1.37
123	doi	***Dogri***	0.03%	0.04%	0.03%	40.00%	0.02%	0.00%	0.00%	0.02%	0.02%	0.46	0.63
107	dyu	**Jula**	0.07%	0.12%	0.02%	30.85%	0.04%	0.00%	0.00%	0.04%	0.03%	0.24	0.43
37	ell	**Greek**	0.18%	0.13%	0.21%	77.71%	0.17%	0.37%	0.19%	0.24%	0.22%	1.75	1.23
1	eng	**English**	15.30%	13.01%	37.4 %	64.33%	27.9%	38.61%	21.73%	17.87%	26.48%	2.04	1.73
125	ewe	**Éwé**	0.03%	0.05%	0.01%	31.78%	0.02%	0.00%	0.00%	0.02%	0.01%	0.26	0.45
19	fas	**Persian**	0.95%	0.81%	0.55%	64.58%	0.39%	0.74%	0.75%	0.81%	0.70%	0.87	0.73
44	fin	**Finnish**	0.09%	0.06%	0.04%	89.67%	0.06%	0.74%	0.08%	0.14%	0.19%	3.42	2.09
4	fra	**French**	3.00%	2.58%	2.64%	63.67%	3.75%	5.40%	4.26%	3.21%	3.71%	1.44	1.24
70	ful	***Fulfulde***	0.19%	0.31%	0.07%	33.16%	0.09%	0.00%	0.00%	0.12%	0.08%	0.25	0.42
89	grn	***Guaraní***	0.08%	0.06%	0.03%	68.83%	0.06%	0.00%	0.01%	0.07%	0.04%	0.64	0.51
73	gsw	**German. Swiss**	0.10%	0.06%	0.08%	91.56%	0.09%	0.00%	0.01%	0.17%	0.08%	1.21	0.72
28	guj	**Gujarati**	0.44%	0.60%	0.53%	40.49%	0.35%	0.05%	0.24%	0.39%	0.34%	0.56	0.76
91	hat	**Haitian Creole**	0.05%	0.08%	0.06%	38.59%	0.06%	0.00%	0.03%	0.03%	0.04%	0.50	0.70
45	hau	**Hausa**	0.43%	0.72%	0.16%	32.61%	0.16%	0.00%	0.10%	0.28%	0.19%	0.26	0.44
20	hbs	***Serbo-Croatian***	0.27%	0.19%	0.14%	77.78%	0.21%	2.49%	0.22%	0.31%	0.61%	3.14	2.21
26	heb	**Hebrew**	0.14%	0.09%	0.08%	85.46%	0.11%	2.20%	0.13%	0.19%	0.47%	5.24	3.35
103	hil	**Hiligaynon**	0.05%	0.06%	0.02%	43.08%	0.07%	0.00%	0.00%	0.04%	0.03%	0.51	0.65
5	hin	**Hindi**	4.26%	5.80%	4.81%	40.18%	3.16%	0.28%	4.03%	3.71%	3.38%	0.58	0.79
82	hmn	***Hmong***	0.09%	0.07%	0.06%	64.80%	0.05%	0.00%	0.03%	0.09%	0.05%	0.72	0.61
75	hne	**Chhattisgarhi**	0.12%	0.16%	0.13%	40.00%	0.08%	0.00%	0.00%	0.10%	0.07%	0.45	0.62
41	hun	**Hungarian**	0.18%	0.12%	0.08%	79.92%	0.15%	0.57%	0.13%	0.20%	0.22%	1.79	1.22

83	hye	**Armenian**	0.05%	0.04%	0.02%	69.86%	0.03%	0.14%	0.02%	0.05%	0.05%	1.41	1.11
101	ibb	**Ibibio**	0.08%	0.10%	0.03%	41.98%	0.03%	0.00%	0.00%	0.06%	0.03%	0.31	0.41
62	ibo	**Igbo**	0.22%	0.28%	0.08%	42.02%	0.08%	0.00%	0.05%	0.16%	0.10%	0.35	0.45
97	ilo	**Ilocano**	0.05%	0.06%	0.03%	43.82%	0.08%	0.00%	0.00%	0.05%	0.03%	0.56	0.69
12	ita	**Italian**	0.91%	0.66%	0.51%	75.65%	0.97%	3.39%	1.22%	1.20%	1.37%	2.09	1.51
27	jav	**Javanese**	0.58%	0.66%	0.20%	47.74%	0.69%	0.00%	0.14%	0.51%	0.35%	0.53	0.61
10	jpn	**Japanese**	2.07%	1.22%	1.98%	92.62%	1.76%	3.55%	2.77%	3.01%	2.52%	2.07	1.22
93	kab	**Amazigh**	0.07%	0.07%	0.04%	62.12%	0.05%	0.00%	0.00%	0.06%	0.04%	0.58	0.51
30	kan	**Kannada**	0.42%	0.57%	0.47%	40.12%	0.31%	0.08%	0.23%	0.36%	0.31%	0.55	0.75
104	kas	**Kashmiri**	0.05%	0.07%	0.06%	38.84%	0.04%	0.00%	0.00%	0.04%	0.03%	0.45	0.63
110	kau	***Kanuri***	0.06%	0.09%	0.02%	39.21%	0.02%	0.00%	0.00%	0.04%	0.02%	0.29	0.40
56	kaz	**Kazakh**	0.18%	0.13%	0.07%	76.98%	0.10%	0.07%	0.10%	0.17%	0.11%	0.90	0.64
64	khm	**Khmer**	0.14%	0.17%	0.07%	43.40%	0.16%	0.02%	0.08%	0.09%	0.09%	0.53	0.66
121	kik	**Gikuyu**	0.03%	0.08%	0.01%	22.57%	0.03%	0.00%	0.01%	0.03%	0.02%	0.22	0.53
111	kin	**Kinyarwanda**	0.06%	0.13%	0.02%	24.69%	0.02%	0.00%	0.01%	0.04%	0.02%	0.19	0.42
132	kln	***Kalenjin***	0.02%	0.04%	0.01%	22.62%	0.02%	0.00%	0.00%	0.01%	0.01%	0.21	0.50
137	kmb	**Kimbundu**	0.00%	0.02%	0.00%	16.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.14	0.48
108	kok	***Konkani***	0.04%	0.06%	0.05%	39.76%	0.03%	0.00%	0.00%	0.04%	0.03%	0.46	0.63
130	kon	***Kongo***	0.02%	0.12%	0.01%	11.62%	0.02%	0.00%	0.00%	0.01%	0.01%	0.09	0.44
14	kor	**Korean**	0.93%	0.79%	0.93%	64.73%	0.99%	0.85%	1.10%	0.95%	0.96%	1.22	1.03
136	ktu	**Kituba**	0.01%	0.05%	0.00%	10.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.07	0.39
40	kur	***Kurdish***	0.32%	0.24%	0.20%	73.02%	0.28%	0.04%	0.15%	0.29%	0.22%	0.89	0.67
39	lah	***Lahnda***	0.31%	0.96%	0.41%	17.43%	0.26%	0.01%	0.15%	0.18%	0.22%	0.23	0.71
134	lua	**Luba-Kasai**	0.01%	0.07%	0.00%	10.05%	0.01%	0.00%	0.00%	0.00%	0.01%	0.07	0.40
117	lug	**Ganda**	0.05%	0.11%	0.01%	25.01%	0.02%	0.00%	0.00%	0.03%	0.02%	0.18	0.39
133	luy	***Luyia***	0.01%	0.03%	0.00%	22.98%	0.01%	0.00%	0.00%	0.01%	0.01%	0.20	0.48
95	mad	**Madura**	0.07%	0.08%	0.02%	47.70%	0.08%	0.00%	0.00%	0.06%	0.04%	0.50	0.57
65	mag	**Magahi**	0.15%	0.20%	0.16%	39.99%	0.11%	0.00%	0.00%	0.13%	0.09%	0.45	0.62
51	mai	**Maithili**	0.24%	0.33%	0.25%	39.28%	0.18%	0.00%	0.02%	0.20%	0.15%	0.44	0.62
35	mal	**Malayalam**	0.28%	0.37%	0.35%	42.54%	0.26%	0.04%	0.18%	0.25%	0.23%	0.62	0.80
120	man	***Mandingo***	0.04%	0.08%	0.01%	26.96%	0.03%	0.00%	0.00%	0.02%	0.02%	0.20	0.42
23	mar	**Marathi**	0.70%	0.96%	0.79%	40.06%	0.52%	0.06%	0.44%	0.61%	0.52%	0.54	0.74
99	mey	**Hassaniyya**	0.07%	0.09%	0.03%	43.68%	0.05%	0.00%	0.00%	0.05%	0.03%	0.35	0.44
77	mlg	***Malagasy***	0.03%	0.18%	0.01%	9.79%	0.03%	0.32%	0.01%	0.01%	0.07%	0.40	2.21
92	mon	***Mongolian***	0.06%	0.06%	0.03%	58.99%	0.04%	0.01%	0.02%	0.06%	0.04%	0.65	0.61
126	mos	**Mòoré**	0.03%	0.08%	0.01%	23.19%	0.02%	0.00%	0.00%	0.02%	0.01%	0.18	0.42
11	msa	***Malay***	2.20%	2.36%	0.89%	51.00%	2.79%	0.79%	1.91%	1.99%	1.76%	0.75	0.80
67	mwr	***Marwari***	0.14%	0.20%	0.16%	39.81%	0.11%	0.00%	0.00%	0.13%	0.09%	0.45	0.62
52	mya	**Burmese**	0.24%	0.41%	0.08%	31.85%	0.25%	0.03%	0.11%	0.14%	0.14%	0.35	0.60
86	nap	**Napoletano-Cal.**	0.07%	0.06%	0.03%	74.39%	0.08%	0.00%	0.00%	0.10%	0.05%	0.84	0.62
58	nep	***Nepali***	0.16%	0.25%	0.09%	35.70%	0.14%	0.03%	0.14%	0.11%	0.11%	0.45	0.69
22	nld	**Dutch**	0.40%	0.24%	0.19%	92.02%	0.42%	1.13%	0.47%	0.60%	0.53%	2.26	1.34
90	nod	**Thai. Northern**	0.07%	0.06%	0.03%	66.47%	0.08%	0.00%	0.00%	0.07%	0.04%	0.70	0.57
122	nya	**Chichewa**	0.04%	0.14%	0.01%	15.87%	0.02%	0.00%	0.01%	0.02%	0.02%	0.12	0.42
43	ori	***Oriya***	0.30%	0.41%	0.33%	39.96%	0.22%	0.01%	0.14%	0.26%	0.21%	0.51	0.70
84	orm	***Oromo***	0.13%	0.36%	0.04%	20.07%	0.06%	0.00%	0.01%	0.07%	0.05%	0.14	0.39
36	pan	**Punjabi. Eastern**	0.33%	0.50%	0.44%	35.80%	0.30%	0.00%	0.03%	0.27%	0.23%	0.45	0.69
17	pol	**Polish**	0.58%	0.39%	0.31%	81.17%	0.53%	1.57%	0.69%	0.73%	0.74%	1.88	1.26

6	por	**Portuguese**	3.05%	2.49%	1.42%	67.16%	5.53%	3.30%	3.85%	2.92%	3.35%	1.35	1.10
57	pus	***Pashto***	0.16%	0.51%	0.20%	17.49%	0.16%	0.00%	0.06%	0.09%	0.11%	0.22	0.69
85	que	***Quechua***	0.07%	0.07%	0.04%	56.82%	0.09%	0.00%	0.01%	0.07%	0.05%	0.66	0.64
78	raj	**Rajasthani**	0.11%	0.16%	0.13%	38.99%	0.08%	0.00%	0.00%	0.10%	0.07%	0.44	0.62
32	ron	**Romanian**	0.32%	0.23%	0.15%	75.66%	0.26%	0.25%	0.30%	0.35%	0.27%	1.18	0.86
135	run	**Rundi**	0.01%	0.11%	0.00%	4.67%	0.01%	0.00%	0.00%	0.00%	0.00%	0.04	0.42
7	rus	**Russian**	3.51%	2.49%	1.81%	77.20%	2.28%	3.38%	3.88%	3.78%	3.11%	1.25	0.88
100	sat	**Santhali**	0.05%	0.07%	0.06%	39.17%	0.04%	0.00%	0.00%	0.05%	0.03%	0.44	0.62
68	sin	**Sinhala**	0.12%	0.17%	0.06%	39.46%	0.11%	0.09%	0.05%	0.11%	0.09%	0.53	0.73
66	slk	**Slovak**	0.11%	0.07%	0.04%	82.47%	0.07%	0.12%	0.08%	0.13%	0.09%	1.30	0.86
114	sna	**Shona**	0.05%	0.09%	0.02%	30.31%	0.03%	0.00%	0.02%	0.03%	0.02%	0.26	0.46
72	snd	**Sindhi**	0.11%	0.32%	0.15%	18.73%	0.10%	0.01%	0.03%	0.06%	0.08%	0.24	0.70
98	som	**Somali**	0.06%	0.21%	0.04%	15.24%	0.06%	0.00%	0.02%	0.03%	0.03%	0.16	0.57
79	sot	**Sotho. Southern**	0.13%	0.13%	0.06%	56.47%	0.08%	0.00%	0.01%	0.12%	0.07%	0.51	0.49
105	sou	**Thai. Southern**	0.05%	0.04%	0.02%	66.68%	0.06%	0.00%	0.00%	0.05%	0.03%	0.70	0.57
3	spa	**Spanish**	7.00%	5.24%	10.7 %	73.08%	11.7%	5.42%	9.94%	7.59%	8.73%	1.67	1.25
80	sqi	***Albanian***	0.08%	0.06%	0.05%	75.48%	0.08%	0.06%	0.03%	0.08%	0.06%	1.12	0.81
124	suk	**Sukuma**	0.04%	0.08%	0.01%	25.00%	0.02%	0.00%	0.00%	0.02%	0.01%	0.18	0.40
47	sun	**Sunda**	0.27%	0.31%	0.09%	47.69%	0.33%	0.01%	0.06%	0.24%	0.17%	0.54	0.62
46	swa	***Swahili***	0.32%	0.78%	0.12%	22.84%	0.21%	0.01%	0.20%	0.20%	0.18%	0.23	0.55
29	swe	**Swedish**	0.22%	0.13%	0.09%	93.49%	0.23%	0.87%	0.24%	0.34%	0.33%	2.61	1.53
25	tam	**Tamil**	0.62%	0.82%	0.71%	41.35%	0.51%	0.19%	0.39%	0.55%	0.50%	0.60	0.80
87	tat	**Tatar**	0.07%	0.05%	0.03%	78.05%	0.04%	0.01%	0.03%	0.08%	0.04%	0.87	0.61
24	tel	**Telugu**	0.69%	0.92%	0.80%	40.71%	0.53%	0.07%	0.38%	0.60%	0.51%	0.55	0.74
113	tgk	**Tajik**	0.05%	0.08%	0.02%	32.22%	0.03%	0.00%	0.01%	0.03%	0.02%	0.29	0.49
33	tgl	**Tagalog**	0.24%	0.25%	0.33%	53.60%	0.43%	0.06%	0.15%	0.24%	0.24%	0.98	1.00
21	tha	**Thai**	0.72%	0.59%	0.29%	66.85%	0.82%	0.33%	0.62%	0.67%	0.57%	0.98	0.80
129	tir	**Tigrigna**	0.03%	0.10%	0.01%	15.68%	0.02%	0.00%	0.00%	0.01%	0.01%	0.12	0.41
76	tsn	**Setswana**	0.14%	0.13%	0.06%	58.16%	0.09%	0.00%	0.01%	0.13%	0.07%	0.53	0.50
96	tso	**Tsonga**	0.08%	0.10%	0.03%	43.30%	0.04%	0.00%	0.01%	0.06%	0.04%	0.38	0.48
61	tts	**Thai. NorthEast**	0.18%	0.14%	0.07%	66.65%	0.20%	0.00%	0.00%	0.17%	0.10%	0.70	0.57
115	tuk	**Turkmen**	0.04%	0.07%	0.02%	31.48%	0.02%	0.02%	0.01%	0.02%	0.02%	0.32	0.55
13	tur	**Turkish**	1.21%	0.85%	1.03%	77.98%	1.59%	0.94%	1.43%	1.22%	1.24%	1.46	1.02
81	uig	**Uyghur**	0.12%	0.10%	0.04%	64.75%	0.03%	0.00%	0.03%	0.13%	0.06%	0.58	0.49
31	ukr	**Ukrainian**	0.37%	0.32%	0.17%	63.96%	0.25%	0.26%	0.33%	0.40%	0.30%	0.92	0.79
131	umb	**Umbundu**	0.02%	0.07%	0.01%	16.00%	0.01%	0.00%	0.00%	0.01%	0.01%	0.14	0.48
18	urd	**Urdu**	0.98%	2.22%	1.33%	24.12%	0.82%	0.03%	0.54%	0.65%	0.72%	0.33	0.74
49	uzb	***Uzbek***	0.27%	0.32%	0.10%	45.90%	0.13%	0.06%	0.13%	0.20%	0.15%	0.46	0.54
16	vie	**Vietnamese**	0.94%	0.74%	0.58%	69.04%	1.15%	0.46%	0.81%	0.83%	0.79%	1.07	0.85
128	vls	**West Flemish**	0.02%	0.01%	0.01%	90.43%	0.02%	0.00%	0.00%	0.03%	0.01%	1.12	0.68
88	wol	**Wolof**	0.10%	0.12%	0.03%	46.09%	0.05%	0.00%	0.00%	0.07%	0.04%	0.36	0.43
59	xho	**Xhosa**	0.20%	0.19%	0.09%	59.96%	0.12%	0.02%	0.05%	0.19%	0.11%	0.59	0.54
50	yor	**Yoruba**	0.32%	0.42%	0.11%	41.74%	0.12%	0.00%	0.10%	0.23%	0.15%	0.36	0.47
71	zha	***Zhuang***	0.17%	0.14%	0.06%	64.67%	0.04%	0.01%	0.00%	0.18%	0.08%	0.54	0.45
2	zho	***Chinese***	17.65%	14.72%	7.79%	65.59%	5.47%	8.18%	25.07%	19.38%	13.92%	0.95	0.79
48	zul	**Zulu**	0.29%	0.27%	0.13%	59.57%	0.17%	0.03%	0.09%	0.27%	0.16%	0.60	0.55