Proceedings of SIGUL2022 @LREC2022, pages 83–91
Marseille, 24-25 June 2022

Resource: Indicators on the Presence of Languages in Internet

Daniel Pimienta
pimienta@funredes.org
Observatory of Linguistic and Cultural Diversity in the Internet
http://funredes.org/lc
Resource link: http://funredes.org/lc2022

Abstract

Reliable and maintained indicators of the space of languages on the Internet are required to support appropriate public policies and well-informed linguistic studies. Current sources are scarce and often strongly biased. The model to produce indicators on the presence of languages in the Internet, launched by the Observatory in 2017, has reached a sensible level of maturity and its data products are shared in CC-BY-SA 4.0 license. It reaches now 329 languages (L1 speakers > one million) and all the biases associated with the model have been controlled to an acceptable threshold, giving trust to the data, within an estimated confidence interval of +-20%. Some of the indicators (mainly the percentage of L1+L2 speakers connected to the Internet per language and derivates) rely on 2021 Ethnologue Global Dataset #24 for demo-linguistic data and ITU, completed by World Bank, for the percentage of persons connected to the Internet by country. The rest of indicators relies on the previous sources plus a large combination of hundreds of different sources for data related to Web contents per language. This research poster focuses the description of the new linguistic resources created. Methodological considerations are only exposed briefly and will be developed in another paper.

Keywords: Linguistic Resource, Languages, Internet, Indicators, Multilingualism

1. Introduction

The Observatory of Linguistic and Cultural Diversity in the Internet [1] has been working with alternative methods for measuring indicators of the presence of languages in the Internet since 1996. The standard method for computing the percentage of Web contents per language is logically to apply a language recognition algorithm to all the existing webpages and count. The huge extension of the Web makes this approach unpractical, except for targeting smaller subsets, as it was done efficiently by the Language Observatory Project, before the project faded out (Mikami et al., 2005). Attempts to use that approach by applying it to a target with a limited number of Webpages supposed to represent faithfully the whole Web, are prone to huge biases, as shown for the method defined by Alis Technologies in 1997[2] and reused) by OCLC (Lavoie and O'Neil, 1999) and (O'Neil et al., 2003. Eight thousand websites were randomly selected by IP numbers and conclusions were derived from a one-shot measurement, instead of a repetitive series treated statistically as a random variable.

Since 2011, W3Techs [3], indeed an excellent and reliable provider of statistics for Web technologies, has been practically the unique source available for Web contents per language, providing daily updated results thru the application of a language recognition algorithm to the home pages of the 10 million of websites classified as the most visited by Alexa.com[4]. The method is analogous to the one used for the other 25 Web technologies that are surveyed by this company, providing extremely interesting results. However, languages are a kind of Web technology quite different from Java Script Libraries or Web servers and processing web content's languages the same way may lead to huge errors. The issue starts by focusing on the home pages of the selection of websites: if you plan counting web contents you need to focus on webpages in order to avoid giving the same weight to a website of ten webpages compared to a website of ten thousand webpages. Furthermore, non-English websites quite often include English words inside their home page (either to introduce the site in English, either because few English words such as *copyright, abstract* or navigation buttons in English are present); this may cause errors to the algorithm. However, the bulk of the error is caused by the lack of consideration to **multilingualism** which makes the algorithm counts as English many websites which offer tenth of language's option in their interfaces. Quite often the website sets the language option automatically, according to user's preference, a practice more and more common, especially for the top sites in the global market (Facebook.com is just one example) and theW3Techs' algorithm is counting only one language per home page, English in those cases. No wonder then why, since 2011, the percentage of English in the Web is kept stable and even growing by

[1] http://funredes.org/lc
[2] https://web.archive.org/web/20010730164601/http://alis.isoc.org/palmares.en.html
[3] http://W3Techs.com
[4] A Web traffic collection and analytics sites belonging to Amazon corporation, about to be retired from the market.

W3Techs, in spite of evidences telling the Internet have changed drastically in the last decade, with Chinese becoming the first language in terms of users, and most Asian languages and Arabic booming. The Web is today probably **more multilingual than the humanity**. According to Ethnologue 2021 data, the ratio of L1+L2 speakers over L1 speakers is 10 361 716 756 / 7 231 699 136 = 1.43. No one shall be surprised then that more than 50% of websites exhibit pages in more than a unique language. Not paying due attention to multilingualism is therefore becoming an inacceptable bias for such studies. W3Techs could, without changing its current selection of websites and core program, fix its biases, with some reworks such as :

- Analyze the language options offered on the homepage and count each language option as well as the English version.
- Find a method to obtain an approximate estimate of the number of pages and multiply each linguistic version by that number in order to count webpages instead of websites.
- When the algorithm reports more than one language on the homepage, as a precaution, do not count the website as English, but rather the second language.

The new results will then be drastically different…

The worrying problem is that, because of the uniqueness of the source, the proven quality of the rest of its surveys, its long-term history and efficient marketing, a large percentage of the linguistic research community (and public policy makers) is taking W3Techs data as reliable inputs. Unfortunately, good theories fed by wrong numbers can hardly provide correct outcomes.

The most symptomatic example of the situation is given by the statistic's aggregator Statista[5] which titles its 2022 announcement about languages in the Internet[6] with a statement which reads as a hard fact: *English Is the Internet's Universal Language*, supported by W3techs data, where English web contents represent 63.7% of the total while Chinese only 1.3%.

At the same time, the Observatory of Linguistic and Cultural Diversity in the Internet computes English and Chinese at the same percentage together, around 20%, while Hindi, with its 224 millions of Internet users, reaches 3.8% (38 times more than the 0.1% measured by W3Techs) and concludes its last announcement with that sentence: *The transition of the Internet between the domination of European languages, English in the lead, towards Asian languages and Arabic, Chinese in the lead, is well advanced and* ***the winner is multilingualism****, but African languages are slow to take their place*.

One, at least, of the two sources shall be extremely wrong and researchers should exercise caution and check the biases of a method before drawing conclusions from its produced data…

2. Alternative Methods

Back in 1998-2007, the alternative method of the Observatory, which provided coherent series for a decade, was limited to English, German and the 5 Latin Languages (French, Italian, Spanish, Portuguese and Romanian). It used Search Engines to count *a comparable vocabulary*[7] for each language (Pimienta, et al. 2009). After 2007, the "marketing evolution" of Search Engines made the method obsolete as their reports of number of occurrences of a searched word become unreliable.

In 2017, the first version of a new Observatory's approach computes 138 languages, those with L1 speakers over 5 million, a limitation adopted to avoid too strong biases as consequence of the working hypothesis of the approach: *all language's speakers in the same country are computed with the same percentage of persons connected to the Internet, the national figure provided by ITU/World Bank.* This hypothesis forbids to compare languages within a country and is hardly applicable to language with low number of speakers. Additionally, it tends to bias positively immigration languages in developing countries (which may be less connected than the average) and to bias negatively European languages in developing countries (which tend to be better connected than the average). Today, the limitation has been extended to L1 > 1M, allowing 329 languages[8] to be processed.

This approach, which has reached maturity in its last version, is an **indirect approximation** to contents, based on the experimental observation that the ratio between world percentage of contents to world percentage of connected speakers has always remained between 0.5 and 1.5 for languages with full digital existence.

There is some kind of *natural economic law* suggested, which would link, for each language, the **offer** (web

[5] http://statista.com Along the line, I will not miss the opportunity to question the ethics of two emerging phenomena which could be correlated. 1) Too many lazy researchers cite Statista as a source of data instead of the very source. 2) Statista offers some data in free access but the identification of the source of that data is only accessible by paid customers.

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

[7] An "equivalent" set of words is selected for each language, with a lot of linguistic precautions (both syntactic and semantic), whose occurrences is counted by Search Engines allowing statistical processing.

[8] Including indigenous languages responding this criterion (for example for languages of the Americas: Aymara, Guarani, Q'eqchi', Kiche and Quechua).

contents and applications) to the **demand** (speakers connected to the Internet). When the number of connected persons increases, the number of webpages logically increases together, in more or less the same proportion. This happens because governments, businesses, educative institutions, etc., and some individuals create contents to respond that demand.

Furthermore, surveys and studies have been consistently reporting that the average Internet users prefer to use their mother tongue and also take opportunity to use, as second option, their second language(s)[9].

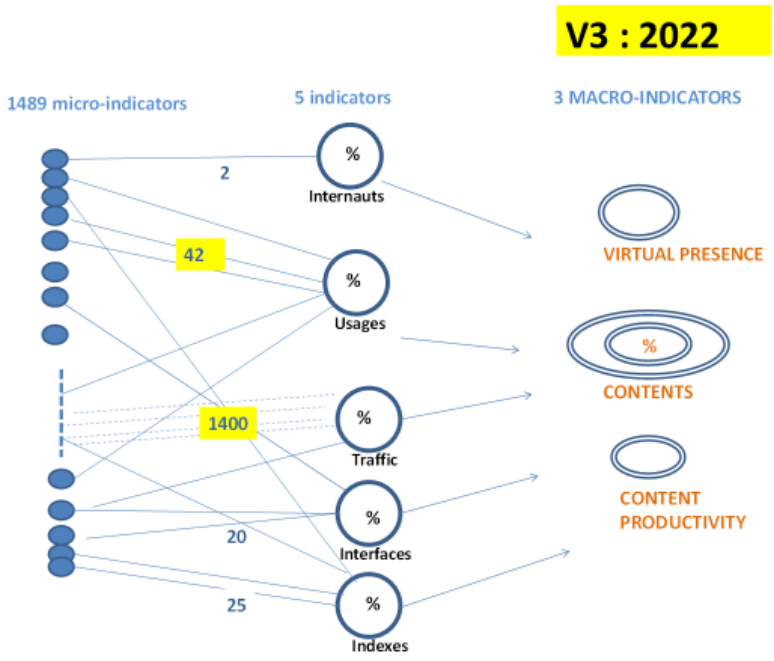
Thus, depending of each language, there is some kind of modulation of the mentioned ratio, to make it above or below one. This would mean that some languages have more content production than others, depending on a set of factors related to languages in their country context, such as :

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

But also:

- The proportion of Internet **traffic** depending of country's tariff, cultural or educational context.
- The number of **subscriptions** to social networks and other Internet applications.
- The digital technological support of the language and its presence in application's **interfaces** and translation programs which would make easier or not the content production.
- The level of submersion of the country where the speaker lives in terms of **Information Society facilities** (e-commerce, government applications to pay taxes and so on).

Then, if it was possible to collect various indicators about each of the mentioned characteristics, one would approximate the fluctuation of the modulation of web contents around one and deduce somehow the contents proportion. This is the core of the method and it is synthetized in the following diagram which shows all the indicators which are processed for each language and the corresponding quantity of sources the model is using. The first and second version of the methodology are fully documented, including the analysis of all identifies biases, see for a lead (Pimienta, 2019). The version 3 detailed methodological description is on the way.



Figure

1: Diagram for indicators creation

This diagram has evolved, from version 1 to version 3, in terms of number of sources and also in terms of indicators, along the hard task of chasing the biases. The computation of the quite complex established

[9] See for instance Union European survey report in https://ec.europa.eu/commission/presscorner/detail/en/IP_1_556 or, for the challenging case of India, this report: https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

model relies extensively in a variety of **weighting operations** to perform the task, with, most of the time, the *vector of percentages of connected persons per country*, which is *the mathematical core* of the process. The source of indicators per language available are scarce; the majority of indicators are obtained per country and most of them only cover a subset of countries. The data source is therefore extrapolated to all countries, weighting with the core data, and the transforming of per country data into per language data is obtained by weighting with the demo-linguistic data (quantity of speakers of each language in each country).

3. Produced Indicators

For each of the 329 languages processed, the model is producing the following indicators per language (note that all world percentages are based on L1+L2 figures and represents the share corresponding for each language).

Intermediary indicators (all are world L1+L2 percentages):
Internauts: speakers connected to the Internet
Usages: relation between users and applications
Traffic: traffic reported to the applications
Interfaces and translation programs: proportion of applications and translation program supported
Indexes: rating of countries in Information Society parameters weighted into language ratings

Model outputs (also called macro-indicators):
Connected speakers : percentage from the total world L1+L2 speakers of those connected to the Internet
Contents : percentage of Web contents (computed as the average of the 5 intermediary indicators)
Content productivity: ratio Contents/Internauts
Virtual presence: ratio Contents/ Speakers

More advanced indicators:
Cyber-geography of languages: repartition of model outputs summed up by language families (European, Asian, Arabic, American, African)
Cyber-Globalization Indicator
CGI (L) = (L1 +L2)/L1(L) x S(L) x C(L)
Where:

L1+L2/L1(L) is the ratio of multilingualism of language L
S(L) is the percentage of world countries which holds speakers of language L
C(L) is the % of speakers of language L connected to the Internet.

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

Additionally, for some languages, it has been displayed the list of countries which hold the major percentages of connected speakers.

The Excel files with the final results can be downloaded from http://funredes.org/lc2022.

A data base access to the results, with the possibility to query by language name or iso code, is in project. The plan is to update yearly the model.

4. Examples

Hereafter some examples of produced data are presented, limited, for the majority of the case, to the top results. The same data is available for any of the 329 processed language. The figure 2 inverted pyramid shall be read as an expression of the confidence interval: Chinese (or English) percentage of Web contents is between 16% and 24%, all the remaining languages together represent between 18% and 26% of the total.

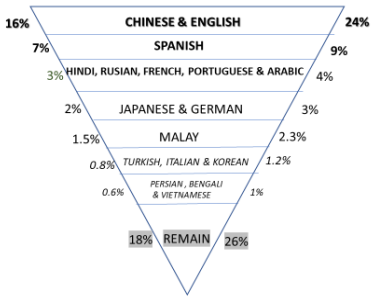


Figure 2: Percentage of contents windows for top languages

. Rank Contents			INTERNAUTS	World Population	Connected Speakers	Contents	Virtual Presence	Content Productivity
L1+L2	**ISO**	**LANGUAGES**	**L1+L2**	**L1+L2**	**L1+L2**	**L1+L2**	**L1+L2**	**L1+L2**
1	**zho**	***Chinese***	18.46%	14.72%	71.38%	21.60%	1.47	1.17
2	**eng**	**English**	14.83%	13.01%	64.86%	19.60%	1.51	1.32
3	**spa**	**Spanish**	6.79%	5.24%	73.72%	7.85%	1.50	1.16
4	**hin**	**Hindi**	4.19%	5.80%	41.16%	3.76%	0.65	0.90
5	**rus**	**Russian**	3.51%	2.49%	80.32%	3.76%	1.51	1.07
6	**fra**	**French**	2.98%	2.58%	65.80%	3.33%	1.29	1.12
7	**por**	**Portuguese**	2.99%	2.49%	68.43%	3.13%	1.26	1.05
8	**ara**	***Arabic***	3.97%	3.53%	63.99%	3.09%	0.87	0.78
9	**jpn**	**Japanese**	1.99%	1.22%	92.63%	2.66%	2.18	1.34
10	**deu**	**German**	2.04%	1.30%	89.17%	2.37%	1.82	1.16
11	**msa**	***Malay***	2.36%	2.36%	56.93%	1.96%	0.83	0.83
12	**tur**	**Turkish**	1.17%	0.85%	78.05%	1.14%	1.35	0.98
13	**ita**	**Italian**	0.87%	0.66%	75.83%	1.00%	1.53	1.14
14	**kor**	**Korean**	0.90%	0.79%	65.16%	0.98%	1.24	1.09
15	**fas**	***Persian***	1.08%	0.81%	75.91%	0.88%	1.09	0.82
16	**ben**	**Bengali**	1.11%	2.58%	24.55%	0.88%	0.34	0.79
17	**vie**	**Vietnamese**	0.92%	0.74%	70.96%	0.85%	1.15	0.92
18	**urd**	**Urdu**	0.95%	2.22%	24.38%	0.66%	0.30	0.70
19	**tha**	**Thai**	0.80%	0.59%	77.95%	0.65%	1.12	0.82
20	**pol**	**Polish**	0.60%	0.39%	87.09%	0.63%	1.59	1.04
21	**mar**	**Marathi**	0.69%	0.96%	41.06%	0.58%	0.60	0.83
22	**tel**	**Telugu**	0.68%	0.92%	41.69%	0.56%	0.60	0.82
23	**tam**	**Tamil**	0.61%	0.82%	42.15%	0.51%	0.62	0.83
24	**jav**	**Javanese**	0.62%	0.66%	53.76%	0.44%	0.66	0.70
25	**nld**	**Dutch**	0.38%	0.24%	91.14%	0.41%	1.73	1.08
26	**guj**	**Gujarati**	0.44%	0.60%	41.47%	0.36%	0.61	0.83
27	**ukr**	**Ukrainian**	0.40%	0.32%	71.02%	0.35%	1.09	0.88
28	**kan**	**Kannada**	0.41%	0.57%	41.11%	0.33%	0.59	0.82
29	**ron**	**Romanian**	0.32%	0.23%	79.57%	0.30%	1.29	0.93
30	**aze**	***Azerbaijani***	0.33%	0.23%	81.54%	0.28%	1.21	0.85
		REMAIN	**22.60%**	**30.10%**		**15.13%**		
		TOTAL	**100%**	**100%**		**100%**		

Table 1: Main indicators for 30 top languages in content's percentage

Table 1 shall be read that way: English represents 14.8% of the Internet connected population and 13% of the L1+L2 world population; 64.9% of English L1+L2 speakers are connected to the Internet; 19.6% of the Web contents is in English; the virtual presence coefficient of English is 1.5, meaning that English contents are over-represented in a factor higher than 50%; the content productivity of English is 1.32, the higher after Japanese.

Note that the *macro languages* are mentioned in italics.

The following tables 2, 3, and 4 expose the top languages for each of the output indicators of the model, respectively:

- Percentage of connected speakers.
- Virtual presence (a value normalized to 1).
- Contents productivity (a value normalized to 1).

Table 5 exposes the Cyber-Geography of languages.
Table 6 exposes the Cyber Globalization Indicator.
Tables 7 and 8 expose respectively the first countries in terms of connected speakers for Chinese and Hindi.
When appropriate explanations are provided below the tables.

LANGUAGE	CONNECTED SPEAKERS
Norwegian	**96.89%**
Danish	**96.42%**
Swedish	**93.94%**
Catalan	**92.88%**
Japanese	**92.63%**
Finnish	**92.07%**
German, Swiss	**91.55%**
Limburgish	**91.42%**
West Flemish	**91.30%**
Dutch	**91.14%**
Galician	**91.07%**
Saxon, Upper	**89.81%**
Estonian	**89.26%**
German. Standard	**89.17%**
Latvian	**89.04%**
Bavarian	**88.24%**

Table 2: Top languages in connected speakers

LANGUAGE	VIRTUAL PRESENCE
Japanese	**2.18**
Norwegian	**1.88**
German, Standard	**1.82**
Swedish	**1.82**
Danish	**1.78**
Dutch	**1.73**
Finnish	**1.69**
Catalan	**1.68**
German, Swiss	**1.63**
Polish	**1.59**
Italian	**1.53**
Estonian	**1.51**
Russian	**1.51**
English	**1.51**
Hebrew	**1.50**
Greek	**1.50**
Spanish	**1.50**
Chinese	**1.47**
Latvian	**1.46**

Table 3: Top languages in virtual presence

LANGUAGE	CONTENTS PROD.
Japanese	**1.34**
English	**1.32**
Chinese	**1.17**
German, Standard	**1.16**
Spanish	**1.16**
Italian	**1.14**
French	**1.12**
Norwegian	**1.10**
Swedish	**1.10**
Korean	**1.09**
Dutch	**1.08**
Russian	**1.07**
Greek	**1.07**
Kabuverdianu	**1.05**
Danish	**1.05**
Portuguese	**1.05**
Finnish	**1.04**
Polish	**1.04**
Catalan	**1.03**
German, Swiss	**1.02**
Hebrew	**1.00**

Table 4: Top languages in contents productivity

LANG. FROM (*)	AFRICA	AMERICAS	ARAB WORLD	ASIA	EUROPE	PACIFIC (**)
Internauts %	29.8%	56.7%	64.0%	49.3%	82.6%	
Contents %	2.89%	0.22%	3.09%	44.77%	45.39%	
POP.L1+L2 %	9.15%	0.31%	3.53%	48.21%	30.91%	
POP. CONN. %	5.18%	0.32%	3.89%	44.60%	39.51%	
Virtual. Pres.	0.28	0.68	0.87	0.65	1.39	
Cont. Prod.	0.51	0.68	0.78	0.72	0.95	
NUMBER OF LANGUAGES	138	8	1	135	47	0

Table 5: Cyber-geography of languages

(*) It has to be understood as native languages.
**) No languages from Pacific are included as none have more than 1 million L1 speakers.

The reading is done that way : African language's L1+L2 speakers have an average connectivity rate of 30% and represent together 3% of Web contents while representing together 9% of world L1+L2 speakers' population and 5% of L1+L2 connected speakers. They have an average virtual presence of 0.3 and a content productivity of 0.5, both indicators quite below the other categories. Note that 138 African languages are processed in the model, a figure slightly higher than the number of Asian languages.

LANGUAGE	CGI	CGI%
English	1.61	14.24%
French	1.09	9.66%
German	0.42	3.75%
Russian	0.31	2.76%
Spanish	0.27	2.40%
Arabic	0.18	1.56%
Malay	0.17	1.51%
Italian	0.17	1.50%
Chinese	0.16	1.46%
Portuguese	0.15	1.37%
Thai	0.15	1.37%
Romani	0.15	1.35%
Turkish	0.15	1.34%
Greek	0.15	1.31%
Ukrainian	0.15	1.31%
Polish	0.13	1.15%
Persian	0.12	1.10%
Rumanian	0.12	1.06%
Hindi	0.12	1.04%

Table 6: Cyber Globalization Indicator

The second column is computed by dividing the CGI value by the total of CGIs for all processed languages. It is mentioned as a way to measure the relative weight[10].

[10] Note that the relative weight of the two first positions, English and French, is close to 25% of the total, showing their strategical advantage. This is coherent with the huge demographic prospects for

CHINESE	L1+L2	%CONN.	CONNECTED	% FROM CONN.
TOTAL	1 525 335 340	71.38%	1 088 735 519	100%
China	1 448 870 000	70.64%	1 023 512 815	94.01%
China–Taiwan	37 320 000	88.82%	33 148 541	3.04%
China–Hong Kong	10 942 800	92.41%	10 112 585	0.93%
Malaysia	7 838 700	89.56%	7 019 949	0.64%
Singapore	4 026 000	75.88%	3 054 766	0.28%
United States	2 894 390	88.50%	2 561 503	0.24%
Viet Nam	2 500 000	70.64%	1 766 054	0.16%
Indonesia	2 054 000	53.73%	1 103 542	0.10%
Thailand	1 729 000	77.84%	1 345 918	0.12%
Canada	1 212 600	97.00%	1 176 222	0.11%
Philippines	1 010 280	43.03%	434 689	0.04%
REST	4 937 570	71.04%	3 507 738	0.32%

Table 7: Repartition of connected Chinese speakers per main countries

HINDI	L1+L2	%CONN.	CONNECTED	% FROM CONN.
TOTAL	600 800 970	41.15%	247 258 401	100%
India	596 000 000	41.00%	244 360 000	98.87%
Kuwait	700 000	98.60%	690 200	0.28%
United States	643 000	88.50%	569 048	0.23%
Nepal	1 307 600	25.00%	326 900	0.13%
South Africa	463 000	68.00%	314 840	0.13%
Saudi Arabia	171 000	97.86%	167 345	0.07%
Australia	160 000	86.54%	138 472	0.06%
Canada	111 000	97.00%	107 670	0.04%
Yemen	316 000	30.00%	94 800	0.04%
REST	929 370	52.63%	489 127	0.20%

Table 8: Repartition of connected Hindi speakers per main countries

5. Bibliographical References

Ethnologue Global Dataset (2022). https://www.ethnologue.com/product/ethnologue-global-dataset-0

Lavoie B.F., O'Neill E. T. (1999). How "World Wide" is the Web? *Annual review of OCLC Research,* https://web.archive.org/web/20031006155123/http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003496

Mikami Y., et al. (2005). The Language Observatory Project (LOP), In *Poster Proceedings of the Fourteenth International World Wide Web Conference*, pp. 990-991, May 2005, Japan

O'Neill E.T., Lavoie B.F., Bennett R. (2003). Trend in the Evolution of the Public Web: 1998 – 2002. *D-Lib Magazine*, 9.4 http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

OIF (2022). Le français dans le monde, Gallimard,

Africa towards 2050: will the African digital divide be overcome those two European languages with higher presence in Africa, will benefit from this phenomenon which will place the African languages which are localized in good situation.

ISBN : 9782072976865. Synthèse en ligne: https://francophonie.org/sites/default/files/2022-03/Synthèse_La_langue_française_dans_le_monde_2022.pdf

Pimienta, D., Prado D., Blanco A. (2009). Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, in *UNESCO Publications for the World Summit on the Information Society*, CI.2009/WS/1

6. Acknowledgements

The version 3 studies were funded by Organisation Internationale de la Francophonie and the results fed the Internet Chapter of (OIF, 2022).

The idea to use various sources of data by country and transform them into data by language was first conceived by Daniel Prado in 2012.

http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

Pimienta D. (2019). Indicators of Languages in the Internet, in Proceedings of International Conference Language Technologies for All (LT4All), 4-6 December 2019, UNESCO, Paris; PP 315-319 https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.79.pdf