Proceedings of the Language Technologies for All (LT4All) , pages 315–319
Paris, UNESCO Headquarters, 5-6 December, 2019.

Indicators of Languages in the Internet

Daniel Pimienta
Observatory of Languages and Culture in the Internet
http://funredes.org/lc
World Network for Linguistic Diversity
http://maaya.org

Abstract

The availability of indicators of the space of languages on the Internet is required to support appropriate public policies. Current sources are scarce and strongly biased. This approach computes indicators for the 140 languages with more than 5 million L1 speakers. It relies on the collection of a large set of micro-indicators measuring languages or countries in various Internet spaces or applications. Statistical methods are applied to produce 6 indicators: Internet users, traffic, use, contents, societal indexes and interfaces, from which 4 macro-indicators are deduced: power, capacity, gradient and content productivity. Some results are presented and the biases of existing methods are analyzed.

Keywords: Languages, Internet, Indicators, Biases

Résumé

Il est nécessaire de disposer d'indicateurs de la place des langues dans l'Internet pour pouvoir conduire des politiques publiques. Les sources disponibles son rares et fortement biaisées. Cette approche calcule des indicateurs pour les 140 langues de plus de 5 millions de locuteurs L1. Elle s'appuie sur la collecte d'une large série de micro-indicateurs mesurant les langues ou les pays d'une variété d'espace ou d'applications de l'Internet. Des méthodes statistiques sont appliquées pour produire 6 indicateurs : utilisateurs de l'Internet, trafic, usages, contenus, index sociétaux et interfaces, à partir desquels 4 macro-indicateurs sont déduits: puissance, capacité, gradient et productivité de contenus. Quelques résultats sont présentés et les biais des méthodes existantes sont analysés.

1. Introduction

During the period 1998-2007, the Observatory of Languages and Cultures in the Internet[1] has been a project of Networks & Development Foundation (FUNREDES[2]) and has collaborated with Union Latine[3] for the design of methods for measurement of language's in the Internet which could provide reproducible and reliable indicators; at the same time other initiatives [4] existed with the same objectives. (Pimienta, 2009). From 2007, changes in the size of the Web and search engines behaviors has rendered obsolete the methods and created a vacuum in the production of indicators of languages in the Internet. A new *artisanal* method, based on the observation of language's behavior in a wide variety of spaces and applications of the Internet was proposed in 2012 and opened new studies of the Observatory, under the World Network for Linguistic Diversity[5] institutional hat and with the support of OIF[6]. Two early studies provide results in terms of rankings for French in the Internet. The second, conducted in 2013, fed the Internet chapter of the 2014 report "Le français dans le monde" (OIF, 2014) and was followed by a similar study of Spanish in the Internet (Pimienta D., Prado D., 2016). The latest OIF funded study, more ambitious, which inspires this article, managed, by the application of a statistical approach, authorized by the increased number of sources, to achieve results in terms of language indicators in the Internet for a wide range of languages.

The method is based on collecting quantitative information about language use in as many as possible applications and Internet spaces. The statistical process of sources enable the measurement of the presence of languages in the Internet and put the results into perspective by building a series of indicators of the share of languages in the Internet. A synthesis is extracted in the form of a series of macro-indicators which combine all indicators. The methodological framework is to use sources either directly when figures concerning languages are available, which is unfortunately rare, or indirectly, using figures per country and transforming them into figures per language. This transformation makes this method an unprecedented approach with the ability to handle the language data quest, in a context where language indicators have become, at best, highly unreliable, but mostly and usually nonexistent.

[1] http://funredes.org/lc
[2] http://funredes.org
[3] http://unilat.org
[4] In particular the ambitious Language Observatory Project (Mikami et al. 2006)
[5] http://maaya.org
[6] http://francophonie.org

This approach is supported by implicit assumptions that need to be made explicit and evaluated to ensure consistency, reliability and expose the corresponding biases. The results are compared with the 2 existing sources: (W3Techs, 2019) and (InternetWorldStats, 2019) and the notable differences are analyzed under the focus of the respective biases.

The details of the methodology, a compilation of the results and the complete biases analysis of the 3 existing methods can be consulted in (Pimienta, 2017) and in http://funredes.org/lc2017.

2. Indicators

The following diagram shows all the indicators which are processed for each language and the corresponding quantity of sources..

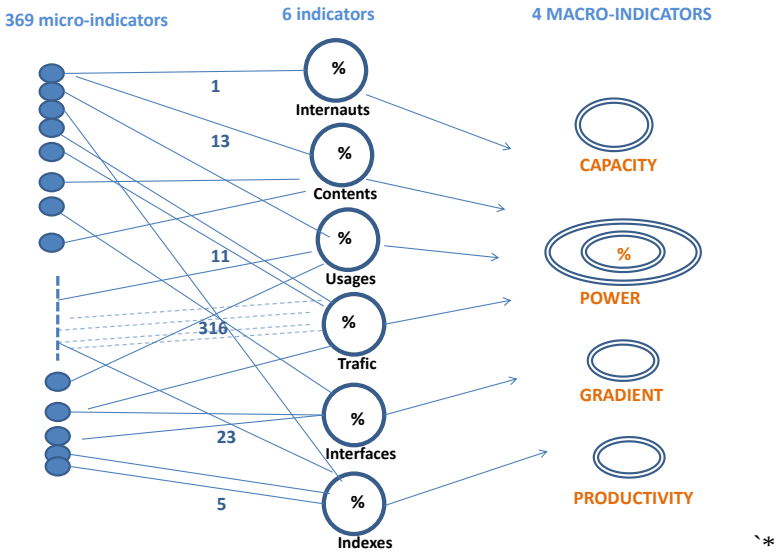

Figure 1 : Indicators diagram

The following table shows for each indicator its sources and how it is computed. All the indicators are expressed in terms of world share, based on the total population of L1+L2 speakers.

INDICATOR	DEFINITION	PROCESS	RELIABILITY
A: INTERNAUTS	Single indicator from ITU: world % of people connected per country.	Weighting country -> language	Very strong Only marginal bias
B: USES	Includes 11 micro indicators: Telephone lines; e.commerce market; OpenOffice download; Social networks users+2021 projection; various social networks subscribers and projections	Weighting C-> L Extrapolation by proportion Truncated mean at 20%	Strong reliability. Low bias. But the number of micro-indicators would need to be extended to give more sense to the mean.
C: TRAFFIC	Alexa.com measured traffic from a selection of 316 websites.	Weighting C-> L Extrapolation proportion. Truncated mean to 20%	Relatively good But huge occidental bias of Alexa
D: INFORMATION SOCIETY	Includes 5 indexes from WebIndex for the following criteria: E.gov, Universal Access, E.participation...	Weighting C -> L, mean Transform to world % by weighting with ITU data	Good (subjective data by competent body). Should be extended.
E: CONTENT (Wikipedia and books)	Includes 13 micro indicators Number of books at Amazon; W3Techs; 11 language indicators from Wikimedia	Direct use of figures per language. Truncated mean to 20%	Very strong. But strong negative bias for Asian languages. Need to be extended
F: **INTERFACE** (and translation languages)	23 binary micro indicators 12 interfaces, 1 content language, 10 translation applications	Presence % transformed in word % by weighting with ITU figures.	Perfect.

Table 1 : Indicators description

The following table shows the defined macro-indicators.

POWER	Measures the global share of the language in the Internet	Mean of the 6 indicators (world L1+L2 %).
CAPACITY	Measures the strength of the language in the Internet regardless of its number of speakers.	Ratio of power vs. world % of speakers. No dimension, normalized to 1.
GRADIENT	Measures the strength of the connected speakers regardless of their number.	Ratio of power vs. world % of connected speakers. No dimension, normalized to 1.
PRODUCTIVITY	Measures the propensity of the connected speakers to produce content in their language.	Ratio of % of contents vs. % of connected speakers. No dimension, normalized to 1.

Table 2 : Macro-indicators description

3. Computations

The model stands on a 3 categories of data : 1) the large list of Internet related sources for language or country 2) Demo-linguistic data 3) ITU data the percentage of people connected to the Internet per country (ITU, 2009).

1) All micro-indicators are expressed into world percentage. The transformation from country to language is realized by weighting with the number of speakers of each language in each country. The sources rarely cover all countries in the world then some extrapolation techniques are used, either in proportion of the percentage of people connected by country or using the method of quartiles. Whenever the extrapolation lacks sense the micro-indicator is rejected.
2) Two sources exists as of today providing the matrix of quantity of L1 speakers of each language for each country: the Joshua project (free of charge) and Ethnologue (fee required). The first edition of the model used Joshua. As for L2 speakers Ethnologue data is used and future measurements will try to use Ethnologue for L1 as well.
3) ITU data, regarded as both reliable and essential to the method, is updated free of charge each year.

The LOC_1 matrix meets the followings definition, for all selected languages and all selected countries: $LOC_1(i,j)$ = Number of L1 speakers for the language i in country j.

The source provides figures for 7500 languages but only a subset will be processed. The estimated number of languages present in the Internet is around 500. One possibility is to target them. Another possibility is to select the languages for which Wikipedia offers statistics (close to 300). After several tests the choice finally settled on the list of the 140 languages with more than 5 million speakers The decision was made in order to reduce the biases resulting from the implicit assumptions.

As for L2, the first priority is that of taking coherent account of multilingualism. The persons computed in L2 obviously have also a first language and therefore the set of L1+L2 speakers includes the same persons more than once. The evidence says that figures must necessarily be based on the total language speakers in the world and not in relation to the world population. This evidence is unfortunately ignored by many sources and provokes errors. In the scenario that is adopted the world share will be calculated on the basis of 125% of the world population (figure computed from the demo-linguistic inputs). This notion is equally applicable to all concepts: users, traffic, usage, content, interfaces and indexes (for instance websites can be made in several languages, the same for the flow of emails). The ideal method to treat the case of L2 would obviously be to produce, as for L1, a matrix $LOC_2(i, j)$ = number of L1+L2 speakers of language i in country j. Unfortunately, this data is unavailable. It is then proposed another approach which simple principle consists, for each language, to get a number that represents the increase to be applied to L1 quantities to get the value L1+L2 and use a linear approach for the results. The global rate of increase (1.25) is the result of the following weighting operation:

$$Rg = \sum_{j=1}^{j=L} L_1(j) \times R_{12}(j)$$

where L is the total number of languages, $L_1(j)$ the number of L1 speakers for language j and $R_{12}(j)$ the rate of increase from L1 to L1+L2 for the language j. The value of micro indicators for L1+ L2 are calculated this way from the value for L1:

$$M_{L1+L2}(i) = Rg \times M_{L1}(i) / \sum_{j=1}^{j=L} M_{L1}(j) \times R_{12}(j)$$

The L1+L2 method is applied to all indicators except Index, Content and Interface which are by nature meant to apply directly to L1+L2. This

method is less accurate than a solution that could work at the country level and generate some biases.

As for computing the indicators, only those expressed by countries require computation. The principle to converts figures expressed in percentages per country into percentages per language is the matrix product between the LOC matrix and, the vector MCn containing the source figures per country for micro-indicator n. The micro-indicator expressed in percentage per language (MLn) is then:

$$MLn(i) = \sum_{j=1}^{j=P} LOC(i, j) \times MCn(j)$$

where P is the total number of countries, LOC(i, j) is the number of speakers of language i in country j and MCn(j) is the measured value for the micro-indicator n in country j.

The matrix product ML = LOC +. x MC in APL[7] notation or ML=SumProduct (LOC;MC) in Excel notation, is a weighting operation of the values of the micro-indicator in each country with the presence of each language in each country. The MLn's totals are the same as those of MPn but this time the distribution is made per language instead of per country. As most of the computations are based on weighting it is useful to identify the different types used in the process and make explicit the simplifying assumptions underlying the validity of the results obtained by these weightings, assumptions which will guide the understanding of biases.

	Demo-linguistic	L2	Users
TYPE	C ---> L	L1 ---> L1+L2	Criterion % -> world %
APPLICATION	Data by C	L1 Results	% by criteria
RESULT	Data by L	L1+L2 Results	% worldwide
WEIGHTING	LOC matrix	L1+L2/L1 per L	IUT data
SCOPE	All sources by country	Users, traffic and usage	Index and interfaces
IMPLICIT ASSUMPTION	Identical connection rate for all L1 in the same C	Identical connection rate for all L2 as for L1	Modulation according to Internet connection rate

Table 3: Different weightings applied

[7] APL, "A Programming Language", a mathematical formalism and programming language.

4. Results

The following table shows the bias corrected results for the first 10 languages for contents and allows comparison with the two other existing sources, showing strong discrepancies which are well understood when the biases are analyzed carefully (Pimienta, 2017).

	CONTENTS	W3TECH	INTERNAUTS	IWS
English	**32,0%**	51,9%	20,4%	26,3%
Chinese	**18,0%**	2,0%	20,0%	20,8%
Spanish	**8,0%**	5,1%	9,1%	7,7%
French	**6,5%**	4,1%	4,9%	2,8%
German	**3,8%**	5,5%	2,7%	2,3%
Portuguese	**3,5%**	2,6%	4,1%	4,3%
Japanese	**3,5%**	5,6%	4,5%	3,2%
Russian	**3,5%**	6,5%	4,9%	2,9%
Hindi	**3,0%**	< 0,1%	4,6%	n.a.
Arabic	**3,0%**	0,7%	3,0%	4,7%
Remaining	***40,2%***	*15,9%*	*46,6%*	*25,0%*
TOTAL	125,0%	100,0%	125,0%	100,0%

Table 4: Ten top languages content & comparisons

The following table shows the ranking for capacity and it is not surprising to see the languages of countries with strong policies for Information Society.

	Capacity	Ranking Power	% connected
Hebrew	5.40	35	76.05
Finnish	5.40	38	92.30
Dutch	4.81	19	92.27
Swedish	4.46	28	90.54
English	3.72	1	78.05
German	3.40	6	86.43
Danish	3.30	49	95.67
Italian	3.16	12	64.20
Czech	3.13	27	81.17
French	2.96	4	81.09

Table 5 : Ten top languages for capacity

And finally the table, sorted by gradient highlights the dynamism of people connected.

	Gradient	Ranking Power
Hebrew	2.62	35
Finnish	2.16	38
Dutch	1.93	19
Swedish	1.81	28
English	1.76	1
Czech	1.42	27
English	1.76	1
Italian	1.73	12
Serbo-Croatian	1.54	22

Table 6: Nine top languages for gradient

5 Bibliographical References

Ethnologue, (2019). Languages of the World. https://www.ethnologue.com

Internet World Stats, (2019), Internet world users per language, top 10 languages. https://www.internetworldstats.com/stats7.htm

ITU, (2019), Percentage of individuals using the Internet per country. https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2019/Individuals_Internet_2000-2018_Jun2019.xls

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

OIF, (2014), Le français dans l'Internet, *Rapport 2014 "La langue française dans le monde"*, pp. 501-541, Nathan. http://francophonie.org/Rapports-Publications.html

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

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

Pimienta, D., Prado D. et al, (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 http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

W3Techs, (2019), Usage of content languages for websites. https://w3techs.com/technologies/overview/content_language/all

6. Acknowledgements

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

The study was funded by OIF.