World Language Barometer

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

This work was carried out by Messrs. Alain Calvet and Louis-Jean Calvet in 2022 with the support of the General Delegation for the French language and the languages of France (Ministry of Culture)

The weight of languages in the world

Summary

1. Introduction

In April 2010, we put online on the Union Latine website a "barometer of , which the reader languages of the [1]will profit from consulting in order to understand the method we world" used to assign a "score" to the different languages taken into account and determine their "weight".

It may seem pointless to "classify" languages in this way and many people think that the most important language in the world is their mother tongue, the language in which they studied or the language they use to communicate with their relatives. In fact, these three functions (mother tongue, language of schooling, family language) can sometimes be fulfilled by different languages in the same individual, and it is precisely by starting from the different roles and the different uses of languages in society that we had produced this barometer. Each language is characterized therein by factors whose value can be continuous or discrete, each factor being able to be taken into account, discarded, or assigned an attenuating coefficient which will reduce its relative importance in relation to the other factors. Each user can thus create his own ranking according to the point of view that interests him and/or his personal assessment of the importance or the merits of the factors proposed.

A second edition of the barometer was put online in 20122 . It allowed the classification of a larger number of languages and used one more factor to carry out this classification.

A third edition was produced in 2017 with a greater number of languages and factors.

We are now putting the fourth edition of our barometer online. We now let's use thirteen factors and classify 634 languages.

The purpose of this document is to provide whoever consults the site with sufficient information to fully understand how we have built it and how they can use it to appreciate the relative importance of languages in the world, to appreciate *their weight*

The first difficulty is to determine what language are we speaking? Twenty-seven variants of Nahuatl as well as Malay, eleven of Malagasy, three Tonga which have no connection between them, three Yaka, two Ndebele, two Sotho, two Azeri, two Punjabi and many other disputed cases have been listed. . The list is too long to be exhaustive. This situation creates a difficulty because many data concerning the languages do not specify which variant it is. The typical case is Arabic. There is a so-called "standard" Arabic and thirty-four so-called dialectal Arabic which are in fact the mother tongues of all or almost all Arabic speakers. If for example we look for the presence of Arabic on Wikipedia we find

[1] http://portalingua.info/en/weight-of-languages/

[2] http://wikilf.culture.fr/barometer2012/

more than 500,000 pages in "Arabic" without further specification and 17,000 in "Egyptian" Arabic. Similarly, if we look at the number of translations from Arabic, about 12,000 relate to Arabic, 5 to Chadian Arabic, 3 to Moroccan Arabic and 1 to Arabic dialects. It is difficult to take these numbers as reflecting reality. To deal with this problem we use the ISO 693-3 standard which assigns a three-letter code to each language, standard Arabic is designated by [arb], Moroccan Arabic by [ary], Egyptian Arabic by [arz] etc.

As we will see further on, it seems difficult to classify between them languages whose characteristics are as different as, for example, Mandarin Chinese, spoken by nearly nine hundred million people, Spanish by about four hundred and thirty million, standard Arabic, the official language of more than twenty states but the mother tongue of a very small number of individuals, if not almost zero, Norwegian, the language of a sparsely populated but rich and cultivated country in the western sense of the term, Swahili spoken by a few million people but the vehicular language of a large part of the African continent. And what about families or groups of languages. For example, in the S30 group of Bantu languages classified by Mr. Guthrie (Sesotho, Sepedi and Tswana) should they be considered as one language or three different languages. Should the complex of Fulani languages be grouped into a single language which would bring together around twenty million speakers or should it remain split between around ten languages (dialects?). The same question can be asked for many other groups of neighboring languages and presenting a high level of inter-intelligibility. What to think of the concept of macrolanguage proposed by the SIL? So many questions whose answers vary from one author to another, from one point of view to another. The ranking necessarily depends on these choices which can obviously be criticized.

We have therefore strictly adhered to the so-called ISO 693-3 classification. [3] which, although imperfect, has the advantage of being consistent and of ensuring a one-to-one correspondence between a language and a three-letter code. It is thus possible to resolve most of the contentious cases.
Using a unique code clarifies the situation. Unfortunately not all data compilations use this nomenclature. Thus certain sources distinguish Tagalog [tgl] from Filipino [fil] and assign between twenty and twenty-five million speakers to each of these languages, which is in disagreement with, for example, Jacques Leclerc's site "Linguistic planning in the world"4 which makes it a single language with approximately twenty-five million speakers. Sometimes a source cites, for example, Guarani, Quichua or Quechua without further details. Gold Ethnologue describes eight varieties of guarani, twelve of quichua and about thirty quechua!

We therefore have to decide which languages we will examine and for this find a reliable compilation of the languages of the world. To our knowledge there are at least three accessible on the web, Joshua6 and People Groups7 . The problem is that ethnologist [5], these three sites mix a will

[3] http://www.ethnologue.com/web.asp

[4] http://www.axl.cefan.ulaval.ca/

[5] https://www.ethnologue.com/

[6] http://www.joshuaproject.net/

[7] http://www.peoplegroups.org/

scientific compilation of languages and their speakers and a more or less present religious aspect. In our first two editions we used Ethnologue as our data source, unfortunately SIL policy has changed and the data is no longer freely available.
So we turned to the Joshua site and put together a database of 17432
recordings. Joshua cites the sources of these data, they are Ethnologue 19th edition in 17398 cases, 34 records not documenting their source. Finally we obtained a base of 6155 different languages taking into account a little less than 7.8 billion speakers. We have eliminated the lines relating to the different sign languages and those indicating “unknown language”. Finally, our working file initially contains 6102 languages with approximately 7.7 billion speakers.

We must then keep in mind that the data we handle is dynamic, it changes, transforms, disappears. Indeed, what about Senna today (iso code 693-3 [sej], a language of Papua New Guinea) for which 10 speakers were listed in 1978? Or Berakou ([bxc], Chad) which had 2 speakers in 1995? What will remain in a few years, or even what remains today of the 360 speakers of Nunggubuyu ([nuy], Australian census of 1996), that the Atlas of languages in danger in the world published by the UNESCO in 20108 reported as "seriously endangered"? Conversely, the existence of Creoles, pidgins of various forms of French in African countries, Filipino, Bahasa Indonesia shows that new languages are appearing.

In the current state of the databases it is impossible to obtain a complete and coherent state of the situation on a given date, therefore to answer these questions with precision, one can only make the best use of the available data, in one word: "make do with what you have", even if it's not very satisfying.

To establish a classification of these 6102 languages the only difficulty consists in providing the amount of work necessary to exploit the available data. Declaring then that this classification makes sense is another problem. Classifying and therefore comparing to others languages whose existence we have no way of verifying in real time is a purely theoretical exercise. We did, but we don't think it has absolute meaning. We will see that the classifications that we can establish are very dependent on the context in which we place ourselves and it therefore seems more reasonable to us to establish partial classifications by filtering the data on one or more criteria, such as the number of speakers, the HDI of the countries in which the language is spoken or the internet equipment etc. This allows comparisons to be made that have more meaning in better defined contexts. We will come back to this below.

The name of the languages sometimes poses a problem. We use the French name of the language as much as possible in accordance with the "Dictionary of languages" [9] But this does not change the list of languages, which are uniquely defined thanks to the three-character code, ISO 639-3, which we make extensive use of.

[8] Atlas of Endangered Languages in the World, Publisher Christopher Moseley, UNESCO Publishing, 2nd edition 2010 and website http://www.unesco.org/culture/languages-atlas/index.php?hl=fr&page=atlasmap

[9] Dictionary of languages, E. Bonvini et alii, Quadrige PUF 2011.

The problems as we can see are numerous and we will detail how we solved them. We will first propose a certain number of factors allowing a language to be described in a "quantitative" way. These factors are not of a *linguistic nature,* but make it possible to appreciate the importance, the *weight,* of a language according to different points of view. Then we will describe how to treat these factors to make them comparable with each other. Finally we will propose various methods to combine these descriptors and arrive at classifications based on the methodology used in our Calvet Barometer of the languages of the world.

2. Factors describing the weight of a language

The factors that we propose are therefore not of a purely linguistic nature and can be separated into two categories

First we will describe those that relate to a language and alone we will call them *intrinsic factors.* The number of speakers is of course the first of these factors but it is possible to imagine others which we describe below.

But the languages live in an environment which influences their importance and their development, this is why we will then consider factors describing the countries in which the languages are spoken, they are then theoretically common to the various languages spoken in the same country *and* a same language spoken in several countries benefits from the contribution of each of them. We will qualify them as *contextual factors* and will retain three of them.

2.A Factors intrinsic to the language 2.A.1 The number of speakers

These are first language speakers, as listed in our database, which we have seen above sometimes poses a problem. In addition, these data relate to L1 speakers (first language), which has the disadvantage of concealing the vehicular character which is one of our factors. For example, Swahili has around three million first language speakers but several tens of millions second language speakers making it a very important language throughout East Africa. Another problem is that the number of inhabitants of the countries does not cross-check with the other sources of demographic data. Moreover the sum of the first language speakers of the languages of the country is sometimes different from the number of the inhabitants of the country, as we have said it is difficult to have exact data.

There is another source of approximation, nobody knows precisely the number of inhabitants in the world and some geographers consider that 25% of births and deaths in the world do not give rise to a declaration to state services. civil. It therefore seems illusory to want to seek precision in these data, they are only approximate.

2.A.2 Entropy

Entropy is a concept that was introduced into statistical thermodynamics by Ludwig Botlzmann in the 19th century. It made it possible to understand the transition from a reversible microscopic dynamic to an irreversible macroscopic evolution. It was then used in theory to information, in linguistics the [10] and today in the theories of prediction of the evolution of universe. It is a subtle concept which, depending on the context, can be interpreted as a quantification of disorder, unpredictability or inaccessible information.

We use it here to differentiate a language spoken in a single country from a language spoken in several countries. Mandarin Chinese or Hindi are spoken in L1 by hundreds of millions of speakers but 98% of them live in a single country, while Spanish is spoken in 72 countries, the largest of which has only 27% of its speakers and eight more than 5%. From the “entropic” point of view, the distribution of Spanish speakers is more “disordered”. The conclusion that we draw is that the usefulness, the international influence of Spanish are superior to those of Mandarin Chinese or Hindi.

We will call pi the proportion of speakers of a given language living in each of the countries concerned.

Classically the mathematical expression of entropy is the following:

Entropy = -ÿ(pi x Log(pi))

in which pi is the probability for a system to be in a given state and Log(pi) the natural logarithm of this probability, the symbol ÿ indicates that we are summing up all the possible states pi . In our case we obviously use pi as defined above, the proportion of speakers of the language considered in each of the countries where it is spoken. The minimum value for this function is zero, when the language in question is only spoken in one country, and there is no defined maximum value.

Let us consider a language spoken overwhelmingly (98%) in a country and of which some speakers live in a second, the entropy will be:

(0.98 x Log(0.98) +0.02 x Log(0.02)) = 0.098

A language whose speakers are evenly distributed over three countries will have an entropy of:

(0.33 x Log(0.33) + 0.33 x Log(0.33) + 0.34 x Log(0.34)) = 1.099

Let us now see in table 1 below some real examples, those of Russian, Japanese, English, Spanish, Standard Arabic and Mandarin Chinese:

Language	Russian	Japanese	English	Spanish	Arab standard	Chinese Mandarin
Entropy	0.667	0.116	1.159	2.543	2.7?	0.160
L1 speakers	136M	124M	362M	443M	?	921M

TABLE 1. ENTROPY AND NUMBER OF L1 SPEAKERS OF SOME LANGUAGES

[10] http://unesdoc.unesco.org/images/0014/001421/142186e.pdf

Russian and Japanese have similar values in terms of the number of speakers, but Japanese is little spoken outside of Japan, while Russian-speaking communities exist in the countries of the former Soviet Union. Russian retains an "imperial language" character and its entropy is greater than that of Japanese. English and Spanish are comparable in terms of the number of their speakers. Spanish is the first language of many medium-sized Latin American countries while the majority of English speakers are concentrated in two countries, the United States and the United Kingdom, the entropy of Spanish is therefore much higher. Standard Arabic has practically no first language speakers but is considered a language present in all Arab-Muslim countries, its entropy, impossible to calculate with accuracy, would be high.
Mandarin Chinese is the most widely spoken language in the world but a very small proportion of its L1 speakers live outside of China, hence its low entropy. We understand then that entropy quantifies the “disorder”, the diversity of the distribution of speakers or even the tendency towards the “universality” of a language.

Entropy has nothing to do with the overall number of speakers of a language, but rather with the way these speakers are distributed in the area or areas in which this language is spoken. It is calculated from the population data described above.

2.A.3 The vehicle factor

Let us first define two concepts. Most often referred to as "mother tongue" first language acquired by an individual, but this appellation is erroneous because in certain plurilingual families this "mother" language may be that of the father (and it would then be necessary to speak of the "paternal" language), it is in any case the language spoken at the house in early childhood or the one in which an individual thinks and expresses himself most naturally. We will use the qualifier of first language which we will note L1. Moreover, it is common for people to study other languages in their school career (or speak of a "foreign language"), use a language that is not their L1 every day in their social or professional practices (we speak then “second language”) or informally acquire in their daily life the rudiments of the languages present in their environment. The situations here are extremely varied. You can learn one or two "foreign" languages at school and speak them more or less well (this is the case of France), you can also acquire at school the official language of the country that you will use. daily (this is the case of French in French-speaking Africa, of English in English-speaking Africa, etc.), one can learn an "identity" L2 language, this is the case of Irish Gaelic and one can finally learn about the many different languages that are only used in limited areas, for example commercial (this is the case with certain traders, in the souks of Marrakech or in the bazaar of Istanbul), etc. While knowing that these situations are different and deserve to be treated in a specific way, we will speak here in a general way of second language, noted L2 for all the situations in which a language other than the L1 is commonly used in life outside close family life.

The number of speakers who have a given language for L1 is obviously an important factor in determining the weight of this language. But just as important are the speakers who speak it as L2, the latter possibly even being higher than the former. The number of L1 speakers of Swahili, we said above, could make believe that it is about a minor language and yet Swahili is a major language of communication in East Africa, spoken by several tens of millions of individuals who have another language for L1.

To quantify this phenomenon it is possible to imagine several methods. We could identify modern language teachers in schools, colleges, high schools and universities around the world or identify pupils and students. But such approaches would be limited to the major languages recognized by the ministries of national education and would leave aside what we wish to take into account: the *vehicle* function of certain languages, this is a fact of society, which does not result from any governmental, administrative or academic decision, but is evident.

It is this fact that we want to quantify by introducing the notion of "vehicularity rate" which we will define as the ratio of the number of speakers using this language as a second language to the total number of speakers.

$$\text{Vehicularity rate} = \frac{2}{1+ 2}$$

This rate varies between 0, for a language which only has speakers in L1 and 1, for a language of which all the speakers would speak it like L2. Presented in this way, things may seem simple, but they very quickly become more complex when we approach the problem of data, both because states are sometimes reluctant to recognize their linguistic diversity and because the sources are often imprecise if not is non-existent.

We first turned to the sources from which we had extracted the number of speakers in L1, then to a few others more specific to second and vehicular languages, but they are generally poorly documented.

- Thus, Ethnologue uses the expression L2 and in its 20th edition indicates for a large number of languages the number of speakers in L1 and L2.

- Ethnologue sometimes indicates that the language considered is used in L2 by speakers of one or more other languages. This is the case for example of the mòoré whose Ethnologue indicates: *"used as L2 by…* " followed by a dozen ethnic groups. In such a case *and when it makes sense,* we retain half of the speakers of these indicated languages to calculate the vehicle factor. But sometimes that doesn't make sense. Let us explain with two examples:

  For Sardinian logudorese ([src]) Ethnologue indicates that it is used in L2 by Catalan ([cat]). This is of course not all Catalan speakers worldwide, they number in the millions, but those living in Sardinia in the Alghero region, they are 23,000.

  Still according to Ethnologue, Uyghur ([uig], L1 speakers 12.5 million) is used in L2, among others, by speakers of peripheral Mongolian ([mvf], L1 2.8 million L1 speakers) and Russian ([rus], L1 speakers 136 million). It is clear that Uyghur is not used in L2 by 68 million Russian speakers but only by Russians living in Xinjiang whose number must be known or estimated.

The situations analogous to these last two are quite numerous, we deal with them by looking for published data on the number of L1 speakers of the considered language living in the vicinity of the one used as L2. Or by estimating it when possible. In the absence of this information, we do not consider that an A language whose number of L1 speakers is significantly higher than that of a B language uses the latter as an L2 language.

There must therefore be a " *hierarchy* " between the L1 and L2 languages. Figure 1 below, of course incomplete, shows the L1/L2 relationships based on a language from Côte d'Ivoire, Anyin. This notion of hierarchy appears clearly there. L1 speakers of Anyin use Dioula, French and/or Akan as their L2 language. But Anyin is used in L2 by speakers of several languages, and English is the “ultimate” L2 language.

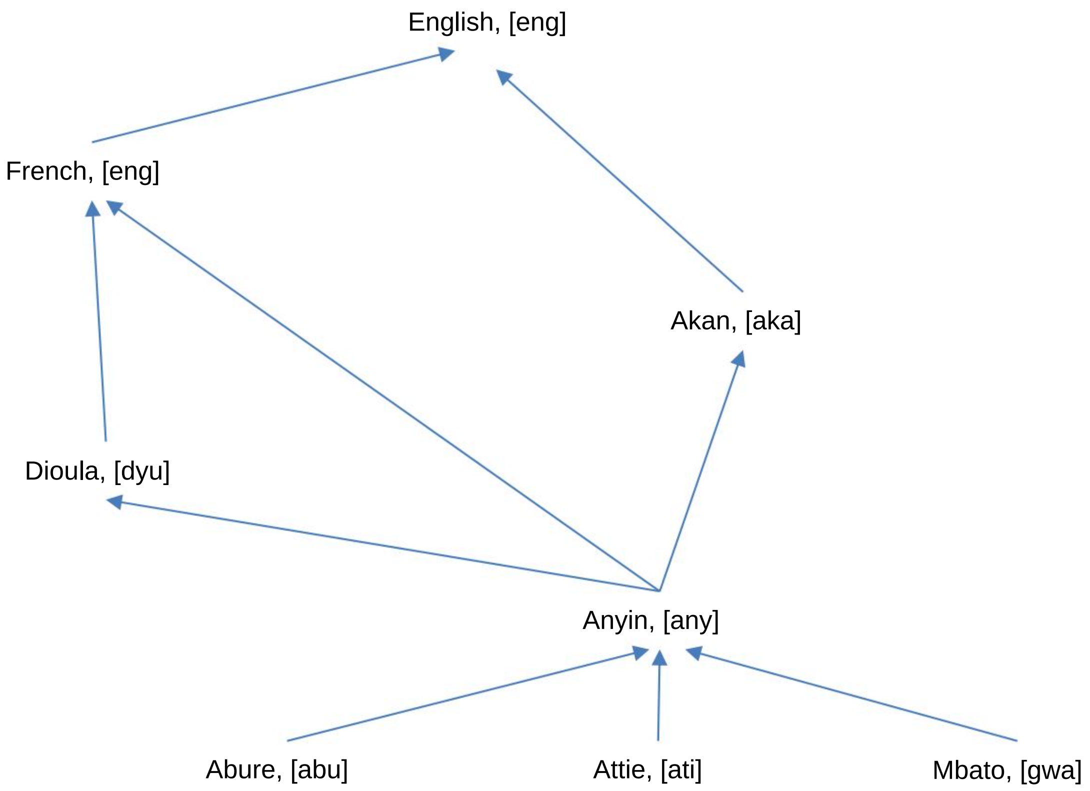


Figure 1. Simplified diagram of L2/L1/L2 relations of Anyin with other languages

- In some cases it is impossible to know the number of L2 speakers. For ebira ([igb]) Ethnologue indicates: *“Other language speakers use ebira to communicate with ebira people”* without any other indication. In this case we consider the language as non-vehicular.

-The data is sometimes at the limit of the contradictory. Thus in Turkey Ethnologue indicates that Southern Zazaki is used as an L2 language by Northern Kurdish and that Northern Kurdish is used as an L2 language by Southern Zazaki! There are 1.7 million Zazaki speakers in Turkey, more than 15 million Kurdish speakers. But no data on their use in L2. What to do in such a situation? In this particular case we have considered that Zazaki, which some consider to be a dialect of Kurdish, is not a vehicular language.

- The country studies of the Laval site are also useful in this area. They often use the term "vehicular" cite the languages concerned but not always the number of speakers. For example we find in the article on Benin the following sentence:

Most Beninese use French, Fon, Yoruba or Bariba as one of the vehicular languages.

but no indication is given as to the number of speakers.

- Some university sites[11,12] provide notices describing the languages and sometimes the number of speakers in L1 and L2.

- The use of keywords such as "secondary speakers" or others in a search engine leads to sites listing the most important languages (> 3,000,000 speakers) and indicating, where appropriate, a number of secondary speakers and an original reference for this data. However, you have to be careful and make sure that the term secondary speakers does indeed correspond to L2 speakers.

- Government sites relating to referendum results are also useful in cases where first and second languages are documented[13] .

The information has therefore been extracted from these different sources but we are often found confronted with problems of definition and probably of "linguistic nationalism". Thus L2 English is spoken by one hundred and sixty-seven million people according to one source[14] and more than six hundred million according to another[15]. French is spoken as a second language by approximately fifty million[16] , or one hundred and fifty-three million. These various sources obviously not talking about the same thing, or not using the same criteria, which number should we retain? Moreover, "minor" languages are completely left out. For example, if it is possible to find an estimate of the number of secondary users of Hiri Motu, Tok Pisin and English in Papua New Guinea, the census of L2 speakers, if any, of the languages of "lower" level among the more than eight hundred spoken in the country is insurmountably difficult.

As we can see, there is no coherent and complete source concerning the vehicular languages and the number of their first and second language speakers. We had to collect the available data and then build the most reasonable set possible.

[11] http://www.lmp.ucla.edu/Profile.aspx?menu=004

[12] http://nalrc.wisc.edu/

[13] http://censusindia.gov.in/2011-common/censusdataonline.html

[14] http://www.nationsonline.org/oneworld/most_spoken_languages.htm

[15] http://en.wikipedia.org/wiki/World_language and cited references

[16] http://www.nationsonline.org/oneworld/most_spoken_languages.htm

[17] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

The examples grouped in Table 2 below give an overview of the method we applied.

For the first, Amharic, there is no agreement between the various sources but the Laval site gives values for the number of speakers in L1 and L2. In such a situation, we have retained the data from this site, which seems to us to offer the best scientific level.

The second example concerns English for which we observe an excellent concordance for the number of L1 speakers. The number of L2 speakers ranges from 167 to 612 million. We retained this last value.

Finally for Tamil a consensus emerges for the number of L1 and L2 speakers, we retained 68 and 8 million respectively.

Language	Code ISO	L1	L2	Sources
Amharic [amh]		25M	5M	http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
		17M		http://www.lmp.ucla.edu/Profile.aspx?LangID=7&menu=004
		21M	21M	http://www.tlfq.ulaval.ca/axl/afrique/ethiopia.htm
		17M	?	http://www.nalrc.indiana.edu/brochures/amharic.pdf
		32 M?		http://www.plc.sas.upenn.edu/languages/amharic.html
English	[eng]	341M	167M	http://www.nationsonline.org/oneworld/most_spoken_languages.htm
		371M	611M	https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers
		340M	170M	http://www.vistawide.com/languages/top_30_languages.htm
		372M	612M	Ethnologue 20th edition
Tamil	[tam]	68M	8M	http://www.ethnologue.org/show_language.asp?code=tam
		67M	8M	http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
		68M	9M	http://www.vistawide.com/languages/top_30_languages.htm
		68M	8M	http://en.wikipedia.org/wiki/World_language
		66M	M's	http://www.lmp.ucla.edu/Profile.aspx?LangID=99&menu=004
		52M	M's	http://www.plc.sas.upenn.edu/languages/tamil.html

TABLE 2. DIFFERENT SOURCES OF L1 AND L2 SPEAKERS

Even if it is neither completely exact nor completely complete, this approach has the merit, in our view, of introducing into the barometer a fundamental factor for evaluating the weight of languages. It adds weight to the dominant hyper-central (English) super-central international languages (French, Spanish, Chinese, etc.), other languages of international communication (Swahili, Hausa) as well as national languages (Kituba, Lingala) and sometimes regional (Gudjarati, Kannada) central or peripheral. At the same time, it underlines the serious gaps in the knowledge of sociolinguistic situations and, on this point, we can only hope that precise studies will multiply.

2.A.4 The status of the language

This factor accounts for the degree of recognition of the languages by the political authorities of the countries in which they are spoken. As we will see below, it goes far beyond the simple notion of the country's official language. Our main source of information here is the site "Linguistic planning in the world" of Laval University[18] .

Let's start by looking at the definitions given there:

"The status of "official language" being a more or less ambiguous concept, it should be understood that, in this site, an official language is recognized by law (de ***jure)*** *or in fact* ***(de facto)*** *by a State (sovereign or non-sovereign), over the whole of the territory or part of it. In all cases, this State must have an assembly, an executive and a public function, which excludes the official languages of an indigenous territory ("reserve"), of an administrative region, of a commune or of a municipality. A State can recognize two, three or four official languages on its territory. This is called a State bilingual, trilingual or quadrilingual.*

A good example of the distinction between *de facto* and *de jure* status can be found in the United States of America, where the constitution does not recognize any official language, but where the de facto official status of English is indisputable.

An ambiguity is sometimes introduced by what we will designate as “co-official languages of convenience” or *second rank.* In Algeria, for example, Tamazight has the status of an official language which is described as "restrictive" by the CEFAN[19] website, the Algerian government has no intention of using this language in communication with citizens, Algeria is not a bilingual state. Should we consider that classical Arabic and Tamazight have the same status? With regard to Djibouti, Jacques Leclerc concludes that French is “more official” there than Arabic[20]. We will find an analogous situation in countries where classical Arabic, spoken by practically no one, is official only for religious reasons. Another case is that of ex-colonial countries which have kept the language of the colonizer as their official language but whose neighbors, with whom their relations are important, are countries which have kept another colonial language.

In Guinea-Bissau, a neighbor of Senegal and Guinea-Conakry and a member of the Francophonie, French has such an important place that Jacques Leclerc considers it a co-official language[21] .

[18] http://www.tlfq.ulaval.ca/axl/

[19] http://www.axl.cefan.ulaval.ca/afrique/algerie-1demo.htm

[20] http://www.axl.cefan.ulaval.ca/afrique/djibouti.htm

[21] http://www.axl.cefan.ulaval.ca/afrique/Guinee-Bissau.htm

Similar situation in Equatorial Guinea, surrounded by French-speaking countries, where the pressure of French, co-official language, to supplant Spanish, first official language, is strong.

If the concept of sovereign state, equivalent to that of country, is clear, that of non-sovereign state must be clarified. Here is how the Laval team defines it:

"Whether they are called State (India or United States), province (Canada), autonomous region (Italy), autonomous community (Spain), territorial community or overseas territory (France), canton (Switzerland), participating government (Francophonie), Free Associated State (Åland, Puerto Rico, Guam), etc., non-sovereign states have varying degrees of legislative, executive and (often) judicial powers.They usually have their own constitution, parliament and their legislation, administration, finances, etc. They enjoy all the attributes of a state, without political sovereignty, but are hierarchically subordinate to another government — the central government.

—, although, in some cases, the fields of jurisdiction are exclusive and exercised autonomously, even sovereignly.

According to this definition, Spain has seventeen autonomous communities and two autonomous cities (Ceuta and Melilla), Hong Kong and Macao are special administrative regions of the People's Republic of China, Easter Island is a special territory of the department of Valparaiso, Mariana Islands is a state loosely associated with the United States of America etc.

We will therefore adopt this definition but will nevertheless distinguish two cases:

- An official or co-official language of a non-sovereign state is the same as that of the sovereign state on which it depends. The islands of the Antilles, the Indian Ocean or the Pacific dependent on France, the United Kingdom, the United States, New Zealand, Australia or the Netherlands have French, English or Dutch as an official or co-official language. Greenland, Gibraltar are in a similar situation. This status most often corresponds not to the importance of the language in the country but to an administrative convenience: English is not spoken much in American Samoa but rather Samoan, in the same way Gilbertin is spoken in Kiribati, Chamorro in the Mariannes, Marshallese in the Marshall Islands, Creole in Guadeloupe and Martinique, Papiamento in the Netherlands Antilles. We will not count these situations and will also consider in the same way certain ambiguous cases of little importance, for example: Saint Helena, the Falklands or Saint Pierre and Miquelon, although no "indigenous" language is opposable there to the official language.

- An official or co-official language of a non-sovereign state is different from the language of the sovereign state on which it depends. Portuguese in Macao, Inuktitut in Greenland, English in Hong Kong, Danish in Schleswig-Holstein, Galician in the autonomous community of Galicia are in this situation. These cases are counted as official language.

On the other hand we will consider this level of official status in non-sovereign states as lower than the previous one. Let us specify by taking the example of Greenland, a non-sovereign state, benefiting from a certain level of political autonomy in relation to Denmark. Greenland has two official languages, Danish and Greenlandic (Greenlandic Inuktitut). We explained above that Danish, the official language of the sovereign state on which Greenland depends, is not counted. Considering Inuktitut as an official language with a status equivalent to that of Danish in Denmark seems abnormal to us, this language is used purely internally and, for example, there is probably no international body providing for simultaneous translations from or into Inuktitut. We will then attribute to the languages of non-sovereign states a lower "value" than that of sovereign states. This point will be clarified below.

There is a third way to distinguish one or more languages among all those spoken in a country. Constitutions and language laws often grant a particular status to one language or another: languages admitted in parliamentary debates, in the administration, in the courts or in the various levels of education. These laws may correspond to a state of affairs, a real desire to promote certain languages, a political or even populist choice, a refusal to choose or any other reason. We will call a language with such a status a "privileged language". We are looking here for a real desire to promote a language and try to avoid declarations of principle not followed by action. This deserves some clarification:

Cases in which a few languages are declared official or national, Senegal for example (six languages) will be considered in this category. But there are extreme cases like Bolivia where article 5 of the 2009 constitution cites three dozen languages by name, Venezuela (2008) which cites about forty, or Peru for which article 48 of the constitution of 1993 indicates in its third paragraph that in addition to Castilian, Quechua and Amayra, "the other languages" are official.

Bolivia:

> ***Article 5***
>
> ***I.*** *Son idiomas oficiales del Estado el castellano y all los idiomas de las naciones y pueblos indígena originario campesinos, que son el aymara, araona, baure, bésiro, canichana, cavineño, cayubaba, chácobo, chimán, ese ejja, guaraní, guarasu' we, guarayu, itonama, leco, machajuyai-kallawaya, machineri, maropa, mojeño trinitario, mojeño-ignaciano, moré, mosetén, movima, pacawara, puquina, quechua, sirionó, tacana, tapiete, toromona, uru-chipaya, weenhayek, yaminawa , yuki, yuracaré and zamuco*

Peru:

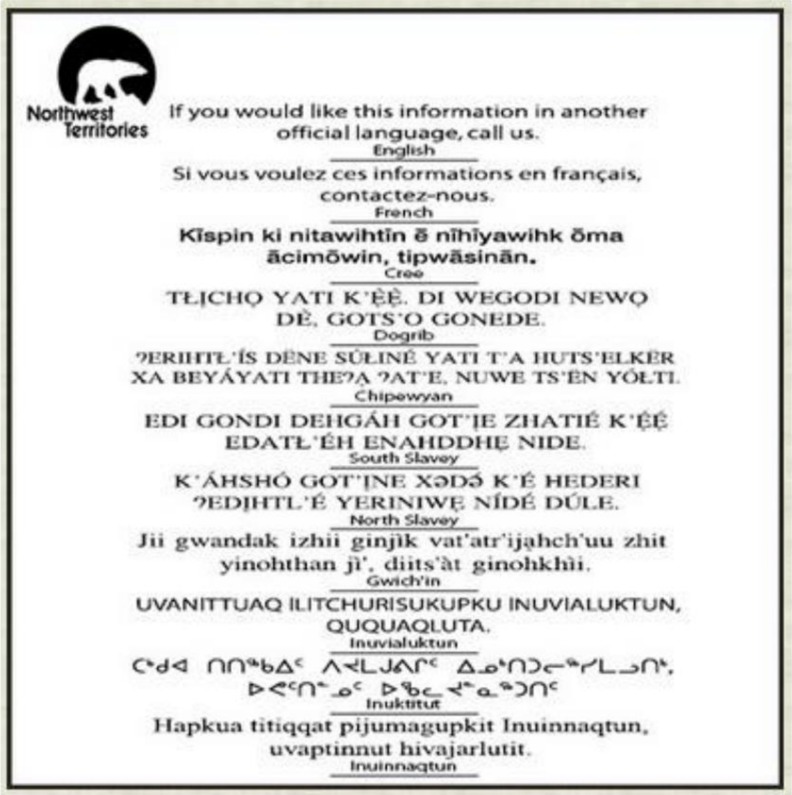
Article 48 [1993]

His official idioms el castellano y, en las zonas donde predominant, también lo son el quechua, el aimara y las demás aboriginal language, según la ley.

These extreme cases will be ignored because they do not correspond to a real voluntarist policy of promotion of such or such language. For example, before a Bolivian court of justice, the documents presented must be written in Spanish and not in one of the thirty-six other so-called official languages, which seems to invalidate article 5 quoted above. Note, however, that the article of the code of civil procedure dates from 1975 and the constitution from 2009.

Sometimes the special treatment granted to languages is reduced to a city. Let us quote Laval University again on the subject of Canada : *"The legislation does not apply to municipalities, but certain municipalities offer services in other languages on an ad hoc basis. The city of Fort-Smith is the only one to have officially declared multilingual services in English, French, Chipewyan, Cree and North Slavey."*

And the Government of the Northwest Territories (of Canada), claims on its website to offer services in eleven languages: English, French, Cree, Dogrib, Chipewyan, South Slavey, North Slavey, Gwich'in, Inuvialuktun, Inuktitut and Inuinnaqtun22. Under the index "Official Languages" this website offers the advertisement below:



Northwest Territories

If you would like this information in another official language, call us.
English

Si vous voulez ces informations en français, contactez-nous.
French

Kīspin ki nitawihtīn ē nīhīyawihk ōma ācimōwin, tipwāsinān.
Cree

TŁĮCHǪ YATI K'ĘĘ. DI WEGODI NEWǪ DÈ, GOTS'O GONEDE.
Dogrib

ʔERIHTŁ'ÍS DËNE SÜŁINÉ YATI T'A HUTS'ELKËR XA BEYÁYATI THEʔĄ ʔAT'E, NUWE TS'ËN YÓŁTI.
Chipewyan

EDI GONDI DEHGÁH GOT'ĮE ZHATIÉ K'ĘĘ EDATŁ'ÉH ENAHDDHĘ NIDE.
South Slavey

K'ÁHSHÓ GOT'ĮNE XƏDƏ́ K'É HEDERI ʔEDĮHTL'É YERINIWĘ NÍDÉ DÚLE.
North Slavey

Jii gwandak izhii ginjìk vat'atr'ijąhch'uu zhit yinohthan jì', diits'àt ginohkhìi.
Gwich'in

UVANITTUAQ ILITCHURISUKUPKU INUVIALUKTUN, QUQUAQLUTA.
Inuvialuktun

ᑖᑯᐊ ᑎᑎᖅᑲᐃᑦ ᐱᔪᒪᒍᕕᑦ ᐃᓄᒃᑎᑐᓕᕐᓯᒪᓗᑎᒃ, ᐅᕙᑦᑎᓐᓄᑦ ᐅᖄᓚᔪᓐᓇᖅᑐᑎᑦ
Inuktitut

Hapkua titiqqat pijumagupkit Inuinnaqtun, uvaptinnut hivajarlutit.
Inuinnaqtun

Let's cite another example of official languages at the local level, it comes to us from the city of São Gabriel da Cachoeira, in the state of Amazonas in Brazil :

Article 1

Portuguese is the official language of the Federal Republic of Brazil

Sole Paragraph *It is established that the Municipality of Saint Gabriel de Cachoeira / State of Amazonas adopts three co-official languages, Nheengatu, Tukano and Baniwa.*

Section 2

The status of co-official language granted by this article obliges the Municipality: ***1°*** *To ensure basic public services of participation to the public in the public distributions in the official language and in the three co-official languages, orally and in writing.*

2° *To produce public documentation, as well as institutional advertising campaigns in the official language and in the three co-official languages.*

3° *To encourage support for the learning and use of co-official languages in schools and in the means of communication.*

[22] http://www.gov.nt.ca/

It will be noted that by article 2 this municipality (34,000 inhabitants in 2005) imposes a certain number of obligations on itself, and that it is not a question here of a generous declaration of intention only formal but of the recognition of these three languages.

It is difficult to arrive at such a level of detail and to identify all the particular situations in the world. We therefore do not claim that our compilation is exhaustive: the origin of the data is found in the constitutions, the linguistic laws, the decrees and regulations enacted at the various levels of the administrations, and it is difficult to consult them all. However, this compilation makes it possible to distinguish languages which have, even at a local level, obtained a certain recognition of their importance, of their "weight".

Another source of confusion is the existence of neighboring languages. Thus in Mali the authorities have recognized 13 national languages. Article 1 of Decree 159 PG-RM of 19 July 1982 cites the following languages[23]:

the ***bambara*** *(or bamanankan), the* ***bobo*** *(bomu), the* ***bozo,*** *the* ***dogon*** *(dogo-so), the* ***peul*** *(fulfulde), the* ***soninké*** *(soninke), the* ***songoy*** *(songaï), the* ***sénoufo-minianka*** *(syenara-mamara and* ***Tamasheq*** *(tamalayt). But other languages are also recognized:* ***Hasanya (Arabic), Kasonkan,*** *Madenkan and* ***Maninkakan.*** *French , meanwhile, enjoys the status of* ***official language,*** *but Bambara serves, in several regions, as the main* ***vehicular language.*** *It is not rare that, in the villages of the South, the children are bilingual (local language + Bambara), even trilingual. At school, French is often taught as a fourth language.*

The problem here is that according to Ethnologue data, there are four varieties of Bozo in Mali, three of which are of similar importance from the point of view of the number of speakers, fourteen varieties of Dogon but no Dogon "dogo-so" , two manikakan and three songai, as for the madenkan, we do not find it anywhere.

The consequence of all this is that the *effective* status of languages in many countries is very vague, the conformity between the texts when they exist and the reality on the ground often being illusory. We will therefore have to be satisfied with a certain degree of approximation in our data, as in the case of the number of speakers.

A fourth level will group together the languages which are accepted as more important *de facto* without a written, constitutional, legislative, regulatory or other text enshrining this pre-eminence.
The existence of schools where a language is accepted as a language and not as a subject of instruction in the elementary levels, the existence of newspapers, radio or television stations internal to the country considered broadcasting in this language, etc.

From the point of view of "weight", we have already indicated that we do not attribute the same value to the different levels of "official, national, constitutional, admitted, privileged" languages.

We will therefore apply the following rules:

a) We will therefore identify in each entity, whether sovereign or not, the languages corresponding to the four levels described and will then combine them by assigning a coefficient of 1 to the

[23] http://www.axl.cefan.ulaval.ca/afrique/mali.htm

official languages of sovereign states, 0.75 to official languages "of convenience", 0.5 to those of non-sovereign states as well as *de facto* national languages and 0.25 to other languages distinguished for any reason whatsoever. It should be noted here that the population of the state is not taken into account, the increment to the score brought by Dominica (75,000 inhabitants) to English is equal to that brought by China (1.4 billion inhabitants). inhabitants, i.e. 20,000 times more populated) to Mandarin.

b) With regard to the languages of the sovereign states we will retain as raw data the number of sovereign states in which the language is official. For Malay for example we get 4, Brunei, Malaysia, Indonesia and Singapore.

c) A language cannot be retained twice in the same sovereign state.

c.1) If in the same sovereign state a language is cited at two or three levels, we will retain it only once, at the higher level. This case is very common in federal states. The official Hindi in India and in ten states of the union is in this case. It will only be retained once as an official language of the Indian Union. Likewise for non-sovereign states, for example Curaçao being formally under Dutch sovereignty, Dutch will not be retained as an official language.

c.2) If a language is official in several non-sovereign states of the same sovereign state, we will retain it only once. Xhosa, official in four states of South Africa, Zulu in three are examples, they will only be counted once each, as the official language of non-states.

Kings.

d) If a language is official in one sovereign state and part of another sovereign state it is considered for both statuses. Swati, official in Swaziland and in the province of Mpumalanga in South Africa is an example, it will be counted twice, national official language in Swaziland and provincial in South Africa.

2.A.5 The number of translations from the language

Here we use data from the Index Translationium24 found on the UNESCO website. The index publishes the number of translations made by language since 1979. The data can be analyzed by country in which the translation took place, by year in which it took place and by subject. The translations are classified into nine categories: general and bibliography; philosophy and psychology; religion and theology; law, social sciences and education; exact and natural sciences; applied sciences; arts, games and sports; literature and finally history, geography and biography. The site does not systematically use the ISO 693-3 codes but with the name of the language indicated in code which is close enough to it so that the user most often has no problem in identifying the language concerned. However, we sometimes encounter ambiguous cases, such as the attribution of a translation to a dialect variety not described by Ethnologue. For example

[24] http://www.unesco.org/xtrans/bsstatlist.aspx

the Translationium index indicates as translation source " Southeastern Ijo dialects", whose code for the index (IJS-DI) does not exist in the ISO 693-3 standard. The ISO code [ijs] actually corresponding to the ijo of the South-East, we assign the translations of the dialects to the ijo of the South-East. Sometimes Translationium indicates a language when there are two or more varieties recognized by Ethnologue. This is the case of Albanian, Azeri, Punjabi. We then make the choice according to the following criteria: the country in which the translations are made give an indication, the ratio of the number of speakers between the varieties is important, one of the varieties is an official language in a country and not the others. Thus data relating to Albanian are assigned to Albanian tosk [als], those relating to Azeri to the northern variety [azl], and those relating to Punjabi to western Punjabi [pnb]. There are other examples of this situation.

Another problem is posed by languages that for political reasons no longer exist. If Hindi and Urdu diverged before the start of the translationium compilation, it is not the same for Serbo-Croatian today split between Serbian [srp], Croatian [hrv] and Bosnian [bos]. We have attributed the data from Serbo-Croatian, which therefore no longer exists, to the three other languages in proportion to their number of respective translations compiled since the "creation" of the languages derived from Serbo-Croatian. Fortunately, there is no Montenegrin yet.

Arabic is another problem. The index reported, as of October 22, 2017, 12,410 translations from Arabic (ARA). In addition, 3 translations were reported from Moroccan Arabic (ARY) and 5 from Chadian Arabic (SHU) and 1 from Arabic dialects (ARA-DI) without further details. In the ISO nomenclature, the code [ara] is that of a *macrolanguage* grouping together all the varieties. On the other hand [arb], [ary] and [shu] are indeed the codes for standard Arabic, Moroccan Arabic and Chadian Arabic (arabe shuwa), and there is obviously no code for the dialects of Arabic. Table 3 summarizes these data. We encounter a similar problem with the compilation of articles in Wikipedia, more than 540,000 articles in standard Arabic, 17,000 in Egyptian Arabic and none in the other Arabic dialects. All this shows that the different *spoken Arabics* are not *written languages.*

The problem probably lies in data collection. The site mentions a certain number of partnerships, national libraries, institutes, universities and experts but it seems that the collection of data is done on a declarative basis which would allow delays, approximate, incomplete or absent declarations. We assigned what Translationium describes as ARA, ARY and SHU to the languages coded [arb], [ary] and [shu] respectively.

	Arab			
	Standard	Moroccan	Chadian Dialects	
Code Translationium ISO 693-3 code	macaw [arb]	ARY [ary]	SHU [shu]	ARA-DI ?
Without indication.	2	0	0	
Arts, Games, Sports	89	0	0	
Exact and natural sciences	68	0	0	
Law, Social Sciences, Education 1072		0	0	1
General, Bibliography...	28	0	0	

History, Geography, Biography	693	0	0	
Literature	4958	3	5	
Philosophy, Psychology	319	0	0	
Religion, Theology	4985	0	0	
Applied Science	196	0	0	

TABLE 3. TRANSLATIONS FROM STANDARD AND DIALECTAL ARABIC

2.A.6 The number of translations into the language

Here we use data from the Index Translationium. We refer to the paragraph previous for more details.

2.A.7 International literary prizes

The purpose of this factor is to take into account the recognition of the culture conveyed by a language through international literary prizes obtained by the writers who have used it.

The first prize that comes to mind is naturally the most prestigious of them all, the Nobel Prize for Literature[25]. However, it is possible to argue that it has several biases. The first of these is to note that most of the prizes have been awarded to authors speaking a language originating from Western Europe. About 60% of the prizes have been awarded to English, French, German or Spanish, the Nobel Committee is "Eurocentric". In the same vein, it should be noted that Sweden alone collected as many prizes as the whole of Asia, eight Swedish prizes against two Japanese and only one Chinese, Bengali, Turkish, Hebrew or Arab. But we must qualify this judgment. Thus the Tagore award distinguishes Bengali from other important languages of the Indian subcontinent. Likewise, languages such as Arabic, Mandarin Chinese, Finnish, Hebrew, Hungarian, Icelandic, Serbian, Czech, Turkish and Yiddish receive recognition of the culture they convey. It should also be noted that Spanish and Portuguese are today more South American than Western languages and in the case of Spanish, the contribution of South American culture is recognized with the prizes awarded to Miguel Angel Asturias, Pablo Neruda, Gabriel Garcia Marques, Octavio Paz and Mario Vargas Llosa.

The second type of controversy is political in nature. The Swedish academy is considered to think "left", which would explain that Jorge Luis Borges was not distinguished because of his support for the Argentinian and Chilean dictatorships. Jean Paul Sartre and Pablo Neruda, who did not condemn left-wing dictatorships, were singled out. The academy has also been suspected of favoring Germany and disfavoring Russia. Tolstoy was nominated sixteen times, Merzkovsky eight times, Berdyayev seven times, they were never awarded and Ivan Bunin was nominated eighteen times to finally be honored in 1933.

The list of authors recognized as adults who have never been distinguished is long: Marcel Proust, Ezra Pound, James Joyce, Vladimir Nabokov, Virginia Woolf, Jorge Luis Borges, Gertrude

[25] http://www.nobelprize.org/

Stein, August Strindberg, John Updike, Arthur Miller, Yannis Ritsos and many more. Many authors "unknown" to non-specialists have however been distinguished: among recent prizes we can cite Hertha Müller and Tomas Tranströmer.

It will be understood, the simple Nobel Prize for Literature is not enough to achieve the goal defined at the beginning of this paragraph. This is the reason why we have chosen to consider other international literary prizes and have chosen the Neustadt Prize[26], the Man Booker Prize[27], the Franz Kafka Prize[28], the Ovid Prize[29] , the Jerusalem Prize[30], the American Award in literature[31] and the "Golden Wreath" award. We also retained the Prince of Asturias prize[32] but only from 1999. Before that date, the winners were all Spanish speakers. We also retained the Park Kyung-Ni prize awarded in South Korea, which has existed since 2011 and has an international vocation.

Many other literary prizes exist but are not intended to examine candidates from all over the world[33]. Some are dedicated to Asian literature, others to Arabic literature or even to a single language such as the Nigerian Karaye prize dedicated to works written in Hausa. The Jnanpith[34] prize awarded in India distinguished authors speaking one of the twenty-two constitutional languages ("scheduled languages") of the Indian subcontinent. We have of course looked for international prizes awarded in countries other than those belonging to the "Western world", there are few of them, most of the prizes reward authors writing in the language of the country. These prices distinguish individuals rather than languages, by construction their domain is limited and we therefore do not retain them.

The rules we apply are as follows:

To. For each of these prizes we attribute a point to the language in which the winner speaks.

b. If a prize is shared, the two languages are awarded one point or if the two authors express themselves in the same language, the latter is awarded two points.

vs. If an author has written in two different languages and he is rewarded for his work as a whole, both languages are awarded a point. This is the case of Milan Kundera.

[26] http://www.ou.edu/wlt/neustadt-prize.html

[27] http://www.themanbookerprize.com/prize/man-booker-international

[28] http://www.franzkafka-soc.cz/

[29] http://en.wikipedia.org/wiki/Ovid_Prize

[30] http://www.jerusalembookfair.com/the_jerusalem_prize.html

[31] https://en.wikipedia.org/wiki/America_Award_in_Literature

[32] http://www.fpa.es/premios/

[33] http://en.wikipedia.org/wiki/Man_Asian_Literary_Prize

[34] http://jnanpith.net/index.html

d. If the same author receives several prizes, his language of expression is awarded as many points as the author received prizes. This is for example the case of Amos Oz or Ismail Kadaré.

Even if for the reasons discussed above this factor only imperfectly reflects what we want to quantify, the international recognition of the level of culture of a language introduces into the barometer an important factor for the evaluation of the weight of languages.

2.A.8 Activity in Wikipedia We use here
the data found on the statistics site of Wikipedia[35]. The number we retain is the sum of all articles published in Wikipedia from the origin of the encyclopedia to the most recent update at the time we collect the data.
Note here that Wikipedia does not use the ISO 693-3 code to unambiguously identify languages, which could theoretically pose certain difficulties but did not constitute a major problem in this case. Ambiguities are resolved in a manner similar to that described in Section 2.A.5.

2.A.9. Education at university level

The idea of this factor, introduced in the third edition (2017) of the language barometer, is to quantify the importance of a language through its teaching at university level. The aim is to examine the websites of a sample of universities in all the countries of the world to extract information on the languages taught at the first levels of higher education, the doctoral level (or "post-graduate ") being excluded. Are also taken into account:

-University or para-university organizations devoted to the teaching of "rare languages" or rather rarely taught languages. These are INALCO (Institut National des Langues et Civilizations Orientales, in France), SOAS (School of Oriental and African Studies, in the United Kingdom), NARLC (National African Languages Resource Center, in the United States), CIIL (Central Institute of Indian Languages, India), etc.,

- "Foreign/modern language centers" which allow students of any level to familiarize themselves with a language without it being formally part of their
University course.

There are about 20,000 universities in the world, there is no question of review them comprehensively. The rules we apply are as follows:

1. Number of universities considered

The first rule concerns the number of universities chosen in each country:

[35] http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total

1.a We select at least 10% of the total number of universities in the country under consideration. That is to say that at least one university will be selected out of a total number between 1 and 10, that at least two universities will be selected out of a total number between 11 and 20, etc.
We insist that this is a minimum. Thus in India we considered 32 universities out of a possible total of 15036 .

1.b For countries in which the number of universities is greater than 100 we examine at least 10 universities and do not always respect the 10% rule 1.a. For example it is possible to visit the websites of nearly 2000 universities in the United States, we considered 49 of them. Similarly in China we considered 25 universities out of a possible total of 354. We will explain ourselves below on this point .

2. How to choose universities?

The aim here is to select the most representative universities in the field of foreign language teaching. Several compilations are available, we use them.

2.a The universities included in the ranking of the best universities for the teaching of modern languages. Two hundred universities are included, a number which does not seem sufficient to us.

2.b Universities included in a ranking of the best universities in the world ranked by country38 .

2.c Universities included in the ranking of the best Asian universities in the field "Arts and humanities"39 One hundred universities are included in this ranking.

2.d A site listing “all” universities in “all” countries of the world40 .

2.e Websites of establishments or organizations dedicated to language teaching SOAS42 are also taken into account. The most representative examples are INALCO41 (School of , , Oriental and African Studies), CIIL in India and NALRC43 (National African Languages Resource Center). There may be others.

2.f Any other university site accessible by other means.

[36] http://www.bulter.nl/universities/

[37] http://www.topuniversities.com/university-rankings/world-university-rankings/2011/subject-rankings/arts humanities/modern-languages

[38] http://www.webometrics.info/

[39] http://www.topuniversities.com/university-rankings-articles/asian-university-rankings/top-universities-asia arts-humanities-2014

[40] http://univ.cc/

[41] http://www.inalco.fr/

[42] http://www.soas.ac.uk/academic/

[43] http://www.nalrc.wisc.edu/

The rules defined in paragraphs 1 and 2 above are applied with flexibility and we examine as many universities as necessary to arrive at a reasonable conviction that the information collected in the country considered is representative of reality. The fact that we examine universities in descending order of "quality" if the site consulted offers such a ranking helps to obtain this conviction. The repetitive appearance of the same languages in the various universities considered in the same country is also a convincing element.

In the end, we try to arrive at a situation in which the best universities teaching modern languages in a given country are taken into account and any addition of one or more other universities in this country would create redundancy.

3. What are the possible sources of approximations?

The method of selecting a sample of universities and not all of them obviously introduces a source of approximation into the study. Several causes are possible.

3.a. The university website is non-existent or inaccessible. This case is encountered for example in Mongolia where we have not been able to find the site[44] "School of Foreign Languages and It is difficult to Cultures" of the National University of Mongolia [45] . know whether the inaccessibility is permanent or temporary.

3.b. The site is difficult to read because it is poorly organized. For example, the "Universidad Nacional del Littoral" in Argentina gives the list of departments and their internet addresses but does not describe the content of the education provided[46]. It is then necessary to explore the tree structure of the site to make sure that the information is not available by another way. Another way would of course be to use the e-mail address indicated to request the information sought. We did not do it.

3.c. The information is unclear or fragmented. For example, some African universities in their African languages departments do not clearly indicate the list of languages actually taught on a permanent basis and often indicate nothing more than "african languages".

3.d. The site is in a language that we do not understand. This problem occurs in Indonesia, and in very closed countries like North Korea and Burma.

3.e. Only the portal or the first pages of the university site are accessible in a language that we understand and the pages on which the information sought may be found are not translated. The situation is analogous to that of the previous case, we encountered it for example in Hungary.

3.f One or more "rare" languages are taught in a university that we have not considered by applying the criteria defined above.

3.g. The university site is classified as risky or malicious by our antivirus software.

[44] http://sflc.num.edu.mn/

[45] http://www.num.edu.mn/

[46] http://www.fhuc.unl.edu.ar/

3.h. Official languages at the national or federal level are not considered. German in Austria, English and Afrikaans in South Africa, English and Hindi in the Indian Union, Malay Mandarin, Tamil and English in Singapore are examples. Languages with hybrid status are considered even in the country where they have this status. The constitutional languages in India, the languages of the provinces in South Africa are therefore taken into account, including in these countries. The consequence is that their importance is overestimated.

3.i Finally, the very nature of language is subject to uncertainty. Here are some examples:

-When a site indicates "Arabic" without further precision, we decide that it is standard Arabic [arb].

-On the other hand in the case of Malay there is in Ethnologue a macrolanguage [msa], a standard Malay [zsm] without L1 speakers and a Malay spoken in Malaysia [zlm]. We chose to retain the latter. This choice seems to us to reflect more precisely the importance of Malay spoken in L1.

-In the case of Nepali we made the opposite choice of the macrolanguage [nep] at the expense of Nepali [npi] because the two components of this one are spoken in Nepal and much less in India and Bhutan

-In the case of languages like Azeri or Kurdish for which several variants coexist our choice depends on the context. North Azeri [azj] being an official language in Azerbaijan is retained at the expense of minority South Azeri in Iran, even if the latter has more speakers.

When organizations such as SOAS or INALCO indicate "Berber" for example, we retain several varieties of Berber.

The sources of inaccuracy are therefore numerous, but in most cases they concern "minor" languages and therefore change little in the analysis of the situation.

4. Results

The idea is not to quantify the number of times a language is taught but the proportion of universities that offer it compared to the number of universities that *could* offer it. To be clear we eliminate for each language the universities of the countries in which the language is official. English in the United States, United Kingdom etc., French in France, Belgium Ivory Coast etc. This calculation is made for the whole world

The process that we have just described allowed us to compile 327 different languages (characterized by their ISO 639-3 code) taught in 1142 universities located in 198 countries. Since the total number of universities in the world is around 20,000, we estimate the margin of uncertainty at around 3%. Remember that the universities selected are chosen, when possible, on the basis of their reputation for excellence in the field of foreign language teaching.

2.A.10. Graphics systems

When we compare the first twenty languages of our barometer, table 4 below, we notice that in our three previous editions fourteen of them are still present

(English, French, Spanish, German, Dutch, Russian, Japanese, Swedish, Italian, Mandarin, Polish, Portuguese, Hungarian, Danish), sometimes in different places. They are still present in the 2022 ranking. Three appear only twice (Catalan, Finnish, Norwegian) and five appear only once (Arabic, Hebrew, Swiss German, Greek and Turkish).

	2010	2012	2017	2022
1	English	English	English	English
2	French	Spanish	French	French
3	Spanish	French	Spanish	Spanish
4	German	German	German	German
5	Dutch	Russian **	Russian **	Russian **
6	Japanese **	Japanese **	Italian	Italian
7	Swedish	Dutch	Portuguese	Swedish
8	Arabic **	Italian	Japanese **	Romanian
9	Italian	Portuguese	Dutch	Portuguese
10	Danish	Mandarin **	Swedish	Polish
11	Finnish	Swedish	Mandarin **	Dutch
12	Russian **	Turkish	Polish	Catalan
13	Mandarin**	Norwegian	Czech	Czech
14	Hebrew **	Polish	Croatian	Croatian
15	Polish	Danish	Romanian	Mandarin**
16	Portuguese	Finnish	Serbian	Hungarian
17	Hungarian	Hungarian	Hungarian	Indonesian
18	Swiss German	Romanian	Korean **	Japanese **
19	Greek **	Catalan	Norwegian	Norwegian
20	Catalan	Czech	Danish	Finnish

Table 4 The top twenty languages in the four editions of the barometer

Some languages therefore appear: Turkish, Norwegian, Finnish, Romanian and Czech in the 2012 version, Croatian, Serbian and Korean in the 2017 version. These modifications are explained by the change of certain languages with regard to our factors, by the addition new factors (vehicularity in 2012, language teaching in universities in 2017) and, with regard to Arabic, the fact that in 2010 we took into account only standard Arabic, whereas we then introduced the different national Arabs (Arabic, Egyptian, Algerian Arabic, etc.)

But there was one constant in these three rankings: the domination of the first twenty places by languages using the Latin alphabet. In 2010 six of them used another graphic transcription system (Japanese, Arabic, Russian, Mandarin, Hebrew, Greek), three in 2012 (Russian, Japanese, Mandarin) and four in 2017 (Russian, Japanese, Mandarin, Korean). These languages are marked with two asterisks **. In this new version, which introduces the graphic system as a new factor, Korean has disappeared from the top of the ranking and if Russian remains in fifth place, Mandarin and especially Japanese have fallen back significantly.

This peculiarity raises a question (concerning the correlation between the place of a language in our barometer and its writing system) and led us to the following reflection: can the graphic system used by a language have an influence on its expanding? For example, is it easier for an English speaker to learn to read French or Turkish than Chinese, for an Arabic speaker to learn to read Urdu or Farsi than Hindi, for a Russian speaker to learn to read Serbian or Bulgarian than Urdu? Let's explain:

The linguist Nicolas Tournadre, after reminding us that we had "proposed to speak of the 'weight of languages' and, in this vein, we could propose the notion of the 'weight of writings'[47], detailed what constituted for his eyes this difficulty:

"The difficulty of writing systems is relatively easy to establish based on the number of signs to be memorized and on the internal complexity of these signs. Logographic scripts have a few thousand signs, syllabics have a few hundred, while alphasyllabaries and alphabets generally only have a few dozen. It is therefore indisputable that logographic systems are more difficult to master than syllabaries and that the latter are more complex than alphasyllabaries and alphabets.

[48].

To verify this hypothesis that we share with him, we have therefore decided to add a new factor in this fourth version of our barometer.

As for the 2017 version of the barometer, we have classified 634 languages. Among the 44 graphic systems used for their transcription, 23 are only used for one language, 9 only for 2 languages. Table 5 below shows the most used systems among our 634 languages:

Graphics system Alphabet	Occurrence	Graphics system Alphabet	Occurrence
Latin	379	Sinograms	29
Arab	74	Ge'ez	13
Cyrillic	34	Bengali	10
Devanagari	30	Laotian	5

Table 5. Compilation of graphics systems

The sum is not equal to 634, we have limited this table to systems used by at least five languages. Some systems are language specific and some languages are simply not written.

We see that the Latin alphabet largely dominates, far ahead of the Arabic or Cyrillic alphabets and sinograms. This domination is of course explained by historical reasons. By the fact firstly for the first two (Latin and Arabic) that alphabetic writing was born during the first millennium BC in the Mediterranean basin and for the Cyrillic alphabet that it was created in the 9th century after by monks (Cyrille and Methodius) inspired mainly by the Greek alphabet.

But it is also explained by religious and imperialist expansions: the Latin and Arabic alphabets

[47] N. Tournadre, *The prism of languages,* Paris, L'Asiathèque, 2016, page 101,

[48] Op.cit. page 287

spread both with the Christian or Muslim religions and with Arab or European imperialism.

We have therefore compiled for the 634 languages that we have retained since the 2017 version of our barometer the graphic system(s) used. A language can be unwritten or used depending on the countries in which it is spoken using one of two or three different graphic systems. This may be due to political and/or religious reasons. Let's give the example of Turkmen, table 6:[49]

Graphic systems used by Turkmen	
In Turkmenistan	Cyrillic Official Latin since 1991
In Iran and Afghanistan	arabic alphabet

Table 6. Graphic systems used by Turkmen

The compilation of the systems used gave us a file of 712 lines, the summary given in Table 7.

Graphics system	Number of occurrences	Graphics system	Number of occurrences	Graphics system	Number of occurrences
Latin	421	Khmer	3	Llao	1
Arab	80	nuosu bburma	3	Malayalam	1
Cyrillic	39	Georgian	2	Meitei mayek 1	
Devanagari	30	Gujarati	2	Mongolian 1 bitchig	
Sinograms 30		Hebrew	2	Nushu	1
Language no 17 ecite		Tai-le	2	Oriya	1
Ge'ez	12	Armenian	1	pahawh	1
Bengali	10	Assamese	1	Sinhalese	1
Batak	6	Balinese	1	Cherokee syllabary	1
Laotian	5	chakra	1	Syriac	1
Thai	5	Ge'ez	1	Tamil	1
Tifinagh	5	Hangul	1	Telugu	1
Tibetan	4	Hmong	1	Thaana	1
Burmese	3	Kana	1	Tigalari	1
Greek	3	Lati	1	Utkala Lipi	1
Kannada	3	Lisa	1		

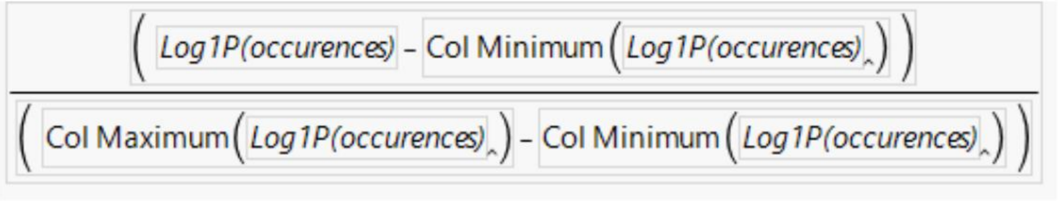
Table 7. Compilation of graphics systems and their use

[49] https://www.axl.cefan.ulaval.ca/asie/turkmenistan-1General.htm

The scores for each of the graphics systems are then calculated as follows:

We assign the value 0 to the absence of a graphics system.

The range of values of the number of occurrences extending over more than two orders of magnitude we perform a logarithmic transformation and calculate the score of a system by applying the formula:



$$\frac{\left(\text{Log1P(occurences)} - \text{Col Minimum}\left(\text{Log1P(occurences)}_{\wedge}\right)\right)}{\left(\text{Col Maximum}\left(\text{Log1P(occurences)}_{\wedge}\right) - \text{Col Minimum}\left(\text{Log1P(occurences)}_{\wedge}\right)\right)}$$

The symbol Log1P(occurrences) means that we take the logarithm of the number of occurrences increased by 1. This makes it possible to have the finite value zero for the absence of a graphics system (the logarithm of zero is not defined). Scores range from 0 (no graphics system) to 1 (Latin alphabet).

Finally for each of the languages we take the average of the scores of the graphic systems used. The scores are still between 0 (unwritten language) to 1 for languages using only the Latin alphabet.

The result of the influence of this new factor on the ranking of languages appears in Figure 2 below. 634 languages are represented there, with rank 2017 on the ordinate, rank 2021 on the abscissa.2 The languages in red below the first diagonal are less well ranked in 2021 than they were in 2017. Their graphic system score is less than 0.6. The languages in green are those which have a graphic system score equal to 1, ie they only use a Latin alphabet. Languages above the first diagonal, ranked higher in 2021 than they were in 2017 are predominantly green.

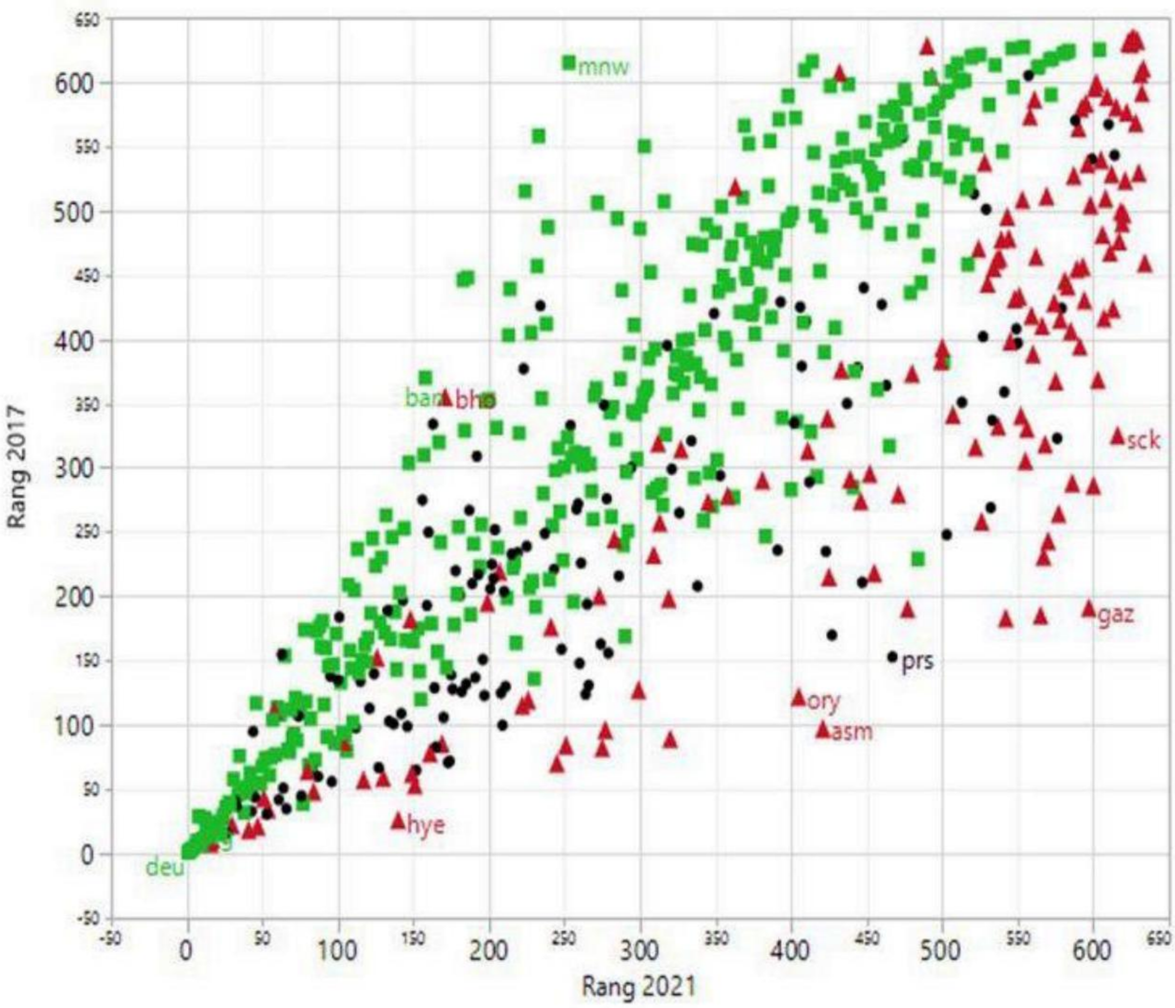


Figure 2. Comparisons of 2017 and 2022 rankings

If we now compare the 2022 ranking (with writing factor) and the 2017 ranking (without writing), still for the top twenty languages, Table 8 below shows us that the languages that drop or disappear in the ranking (in bold) have, with the exception of Portuguese, another graphic system than the Latin alphabet, and those which advance or appear (in red) all use the Latin alphabet.

	2017	2022
1	english	English
2	French	French
3	Spanish	Spanish
4	German	German
5	Russian	Russian
6	Italian	Italian
7	**Portuguese**	Dutch
8	**Japanese**	Swedish

9	Dutch Romanian Swedish Portuguese Mandarin Polish	
10	Polish Catalan Czech Czech Croatian Croatian **Mandarin**	
11	Romanian Serbian Indonesian **Hungarian** Hungarian	
12	Korean **Japanese** Norwegian Norwegian Danish	
13	Finnish **Table 8. Comparison of 2017 and 2022**	
14	**rankings**	
15		
16		
17		
18		
19		
20		

This correlation between changes in the rank of languages and the introduction of the factor graph in our barometer, therefore appears as a causal relationship

2.B Contextual factors

These are the factors that are not specific to a particular language but to the country or countries in which a language is spoken, the context in which it lives.

2.B.1 The human development index

We use here the data found on the site of the United Nations Development Program (UNDP) which publishes an annual report on the state of development of the various countries of the world50. We use the most recent edition of this report, published online in the spring of 2017. The data relating to the HDI are included in table 1 on pages 212 and following. The HDI is a composite index taking into account gross national product per person, life expectancy at birth and level of education. It quantifies the level of development of a country. It is likely that in many countries the human development index is not the same for all regions, for all the ethnic groups living there and therefore for all the languages that these ethnic groups use.

But we do not have more precise data and we assume that the index is constant throughout the country, whatever its heterogeneity may be. To assign a value to each language, we take a weighted average of the index in each of the countries in which the language is spoken. For example, suppose Somali is spoken in Somalia (51% of speakers), Ethiopia (30%), Kenya (16%) and Djibouti (3%). Somali's human development index will be

so calculated as follows:

$$\text{Somali HDI} = 0.51 * \text{IDHSomalia} + 0.30 * \text{HDI Ethiopia} + 0.16 * \text{IDHKenya} + 0.03 * \text{IDHDjibouti}$$

[50] http://hdr.undp.org/en/2016-report

The numbers used above are not strictly exact but they suffice to explain the method used. The UNDP site only provides data for countries affiliated to the UN and for which an index has actually been calculated, which notably excludes countries that are not members of the UN and countries at war. In this case we assign to the undocumented country an estimated index that we decide by analogy with neighboring and/or comparable countries. Thus, for Somalia, a country at war for several years and whose state structures are failing, we assigned a value equal to that of Niger, the lowest value published by the UNDP, i.e. 0.348

2.B.2 The fertility index We use the

same source here as before, the data relating to fertility appear in table 8 on page 238 and following of the report. The total fertility rate is the number of births per woman. It is likely that in many countries the fertility rate is not the same in all regions of the country, for all the ethnic groups living there and therefore for all the languages that these ethnic groups use. Unfortunately we do not have access to more precise data and assume that the fertility rate is constant throughout the country. To assign a value to each language, we take a weighted average of the index in each of the countries in which the language is spoken. For example, Somali is spoken in Somalia (65% of speakers), Ethiopia (30%), Kenya (3%) and Djibouti (2%). The Somali fertility rate will therefore be calculated as follows:

$$\text{Somali fertility} = 0.51 * \text{FertilitySomalia} + 0.30 * \text{FertilityEthiopia} + 0.16 * \text{FertilityKenya} +0.03 * \text{IFertilityDjibouti}$$

As in the case of the HDI, the UNDP site only provides data for countries affiliated to the UN and for which an index has actually been calculated, which here also excludes non-UN member countries and countries at war. In cases where the country is not referenced on the site, we use the same method as above and estimate fertility by analogy with neighboring and/or comparable countries.

Other possible sources for this data are “Fecondity Index by country[51]”, “Index Mundi[52]” and a few others. The data is generally consistent.

2.B.3 Internet network penetration

Here we use data found on the World Stats[53] website which maintains the number of internet users for all countries in the world and from the demographic data calculates a penetration rate as a percentage of the population , a percentage that we took over. The United Nations Development Program also publishes a rate of internet users[54], but, for the sake of consistency with our previous work, we have kept the data from the World Stats website. Here again it is likely that in many

[51] https://worldpopulationreview.com/country-rankings/total-fertility-rate

[52] https://www.indexmundi.com/

[53] http://www.internetworldstats.com/stats.htm

[54] ~~http://hdr.undp.org/en/countries/profiles/~~

country the internet penetration rate is not the same in all regions of the country, for all the ethnic groups living there and therefore for all the languages that these ethnic groups use. And here again we can only consider by hypothesis that the rate is constant throughout the country. To assign a value to each language, we take a weighted average of the rate in each of the countries in which the language is spoken. For example, Somali is spoken in Somalia (65% of speakers), Ethiopia (30%), Kenya (3%) and Djibouti (2%). The penetration rate of Somali will therefore be calculated as follows:

$$\text{Somali rate} = 0.51 \,{}^{*}\, \text{RateSomalia} + 0.30 \,{}^{*}\, \text{RateEthiopia} + 0.16 \,{}^{*}\, \text{RateKenya} + 0.03 \,{}^{*}\, \text{RateDjibouti}$$

The data used are those indicated on the site used in February 2021.

3. Data Processing

3.A Normalization of values

The different factors used, such as those that we could add, do not give us numerical values of the same type. A language is official or not in a certain number of countries, we then obtain a set of discrete values between 0 and the highest number of countries in which a language is official (24 for French). The fertility rate gives us a continuous type value between 1.1 (Macao, Hong Kong) and 6.9 (Niger). The number of speakers can take any value between 0 (a dead language) and 888,000,000 (Mandarin Chinese).

To give each of the factors equal importance, we went from the raw values obtained as described above to standardized values, by carrying out a linear transformation according to the formula:

$$\text{Standard value} = \frac{\text{(Raw value)ÿ(Minimum Raw value)}}{\text{(Maximum Raw Value)ÿ(Minimum Raw Value)}}$$

This transformation assigns the normalized value 1 to the maximum raw value of the factor, the normalized value 0 to the minimum raw value and intermediate values distributed in a linear fashion for the other values. The result is that all the factors vary between 0 and 1, which makes it possible to assign them equal importance in the ranking.

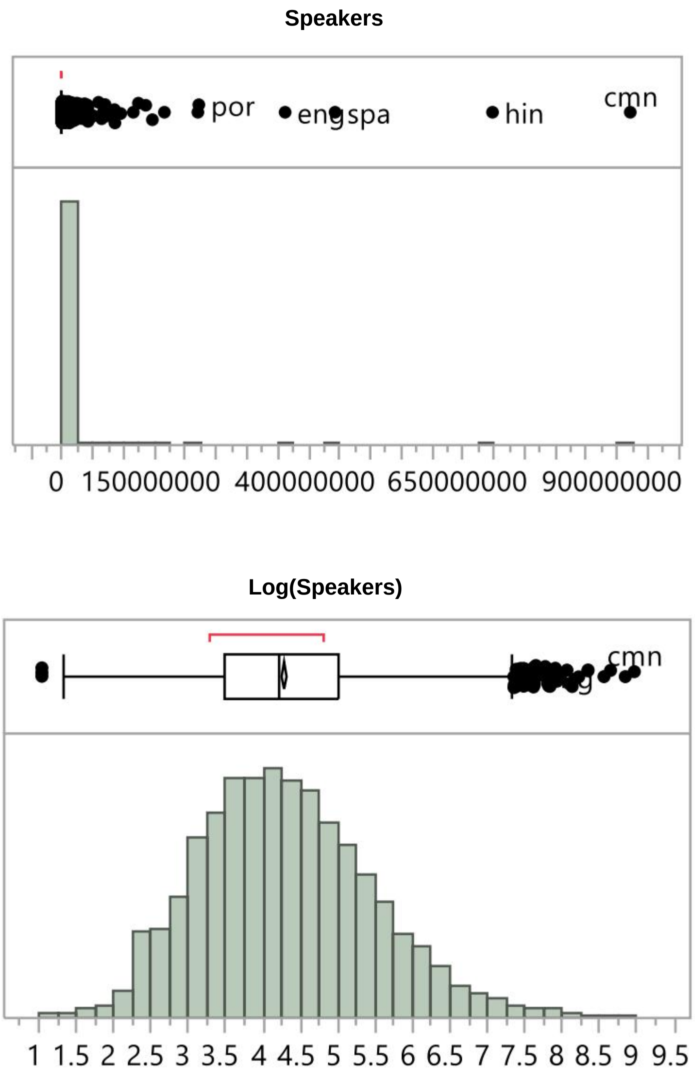
3.B Use of logarithms

For some factors the range of variation is restricted and spans two orders of magnitude or less. Internet penetration is strictly between 0 and 100%, the fertility index between 1.1 and 76.9, the human development index is by construction between 0 and 1. On the other hand, the number of speakers of a language spans nearly nine orders of magnitude (from 0 to 888 million), the number of articles in Wikipedia over six, the translation streams over five, and the number of literary awards over nearly 2. extent of these ranges of variation makes it difficult to distinguish between the lowest values. To get around this difficulty we will use the logarithms of the raw values, which has the effect of bringing the highest values closer together and spreading the lowest values. The normalized values, between 0 and 1 are then calculated as indicated in the previous paragraph.

Figure 3 below makes it possible to clearly understand the interest of such a logarithmic transformation.

The graph in the upper part of the figure represents the distribution of the number of speakers. Mandarin (cmn) has 920 million, Hindi 680, Spanish 440, English 360 and Portuguese 220. But the most important part of this graph is the green vertical bar on the left, it contains 6093 languages with less than 10 million speakers. Suffice to say that this factor differentiates very little or not at all between 0 and 10 million speakers, which is difficult to satisfy. In statistics, such a distribution of values is called positive asymmetry.

It is characterized in a numerical way by the coefficient of asymmetry which, for the amateurs of mathematics, is the centered moment of order 3 normalized by the cube of the standard deviation. In this particular case, this coefficient is equal to 36, which is of course a very high value.


Speakers

Log(Speakers)

FIGURE 3. DISTRIBUTION OF THE NUMBER OF SPEAKERS. LINEAR AND LOGARITHMIC

The second graph (lower part) represents the Log(speakers) factor as well. These are the same languages with the same number of speakers. The distribution is now practically symmetric, and the asymmetry coefficient equal to 0.38. The factor then becomes much more discriminating, it makes a clearer difference between languages with a medium and low number of speakers. It should however be noted that there is a price to pay for this improvement, it is at the top of the scale where the group of languages most spoken in L1 differ less well from Mandarin than previously (group of points on the right).

Table 9 visualizes the question. It represents according to a number of speakers varying from 1000 (for example Lacandon) to 920 million (Mandarin), the logarithm of this number and the standardized values calculated as defined above in paragraph 2.C.1 for raw values and their logarithms. We can clearly see the desired effect, the spread of low and medium values (in red), as well as the price to pay, the compression of high values (in blue).

Language	Speakers	Log (Speakers)	Standard (loc)	Standard (Log(Loc))
Mandarin	920M	20.641	**1,000**	**1.0000**
Spanish	443M	19.908	**0481**	**0.960**
Javanese	116M	18.566	**0.126**	**0.886**
Afrikaans	10M	16.127	**0.011**	**0.753**
Kambaata	1M	6,000	**0.001**	**0.626**
Alago	100000	5,000	0.000	0.500
Zapotec, Ozolotepec	10000	4,000	0.000	0.373
Lacandon	1000	3,000	0.000	0.247

TABLE 9. USING A LOGARITHMIC TRANSFORM

To decide which will be the factors that will undergo this logarithmic transformation on we will base on the ratio between the highest and the lowest value. If this ratio is equal to or greater than 100, an arbitrarily chosen value, we will use the logarithmic transformation of the raw data. Five factors out of thirteen will thus be dealt with: the number of speakers, the two streams of translations, the number of articles in Wikipedia and teaching at the university level.

3.C Statistical independence between data

In all multifactorial problems, care must be taken that the multiplication of factors does not lead to too much redundancy in the data. The statistical approach to this question consists in calculating what is called the "linear correlation coefficient of Pearson", named after the English mathematician who defined it. The mathematical expression of the coefficient is of no interest here, it suffices to know that it varies from -1 to +1. A value of 0 demonstrates the absence of correlation, the independence between two columns of values. This is of course the desired ideal situation. A value of 1 indicates a perfect correlation, the two factors considered are completely equivalent, the redundancy is total and not taking one of the two factors into account causes no information to be lost. A value of -1 also indicates a perfect correlation but in the negative direction.

Usually an intermediate value is obtained. Any value lower than 0.5 (in absolute value) shows a satisfactory independence of the two factors, any value higher than 0.85 shows a significant redundancy, the intermediate values are interpreted according to the circumstances.

To better understand, let's examine the four graphs in figure 4, they represent the internet factor along the abscissa and the ordinate from top to bottom and from left to right the HDI, Wikipedia, entropy and fertility factors. The correlation coefficients are respectively 0.859, 0.460, 0.155 and -0.762.
There is a strong positive correlation between HDI and internet, a weak correlation between Wikipedia and internet, no correlation with entropy and a moderate negative correlation with fertility. The conclusion is that the internet and HDI factors essentially give the same information, whereas internet and entropy are completely independent of each other. We discuss below how to deal with this problem using the attenuator coefficients.

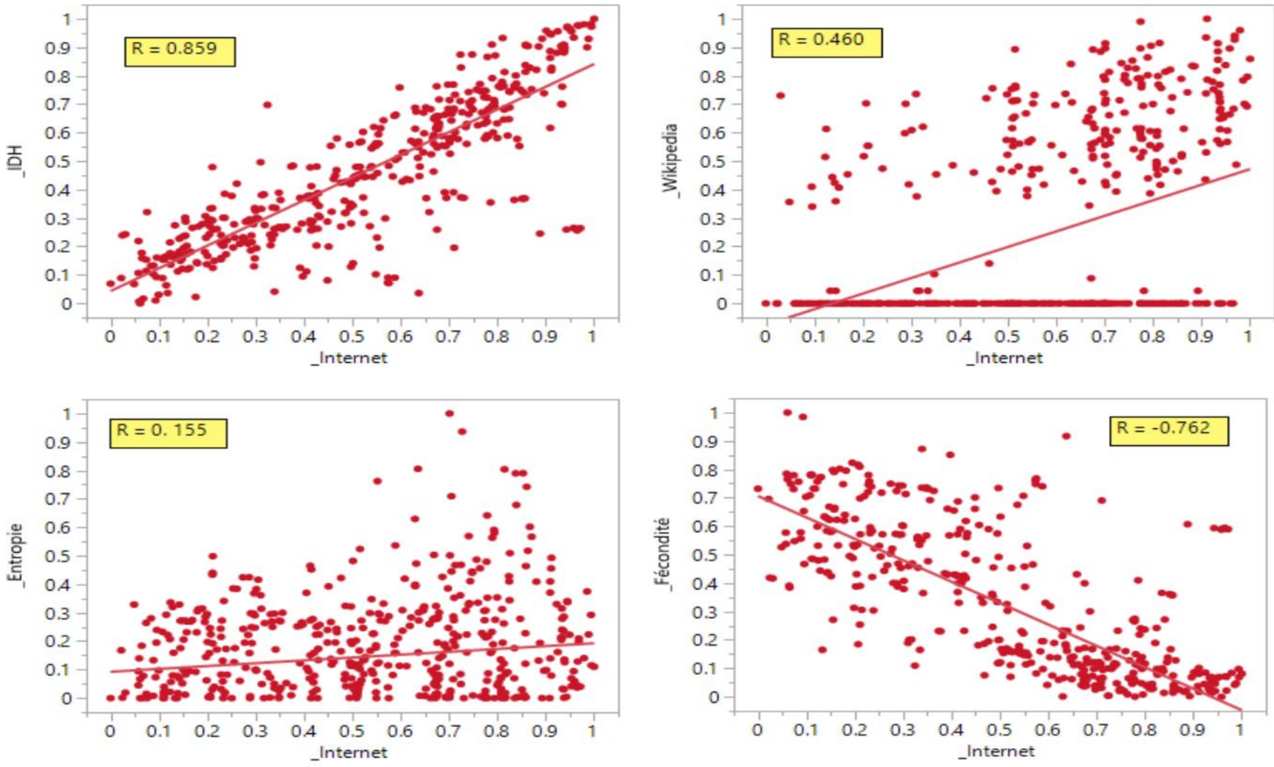
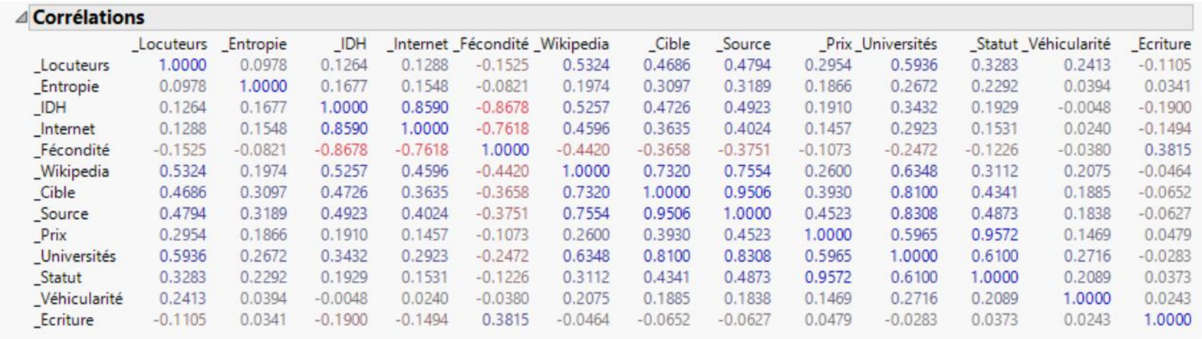


FIGURE 4 CORRELATION BETWEEN SOME FACTORS

Table 10 shows all the correlations between the language parameters shown in the barometer.



Corrélations

	_Locuteurs	_Entropie	_IDH	_Internet	_Fécondité	_Wikipedia	_Cible	_Source	_Prix	_Universités	_Statut	_Véhicularité	_Ecriture
_Locuteurs	1.0000	0.0978	0.1264	0.1288	-0.1525	0.5324	0.4686	0.4794	0.2954	0.5936	0.3283	0.2413	-0.1105
_Entropie	0.0978	1.0000	0.1677	0.1548	-0.0821	0.1974	0.3097	0.3189	0.1866	0.2672	0.2292	0.0394	0.0341
_IDH	0.1264	0.1677	1.0000	0.8590	-0.8678	0.5257	0.4726	0.4923	0.1910	0.3432	0.1929	-0.0048	-0.1900
_Internet	0.1288	0.1548	0.8590	1.0000	-0.7618	0.4596	0.3635	0.4024	0.1457	0.2923	0.1531	0.0240	-0.1494
_Fécondité	-0.1525	-0.0821	-0.8678	-0.7618	1.0000	-0.4420	-0.3658	-0.3751	-0.1073	-0.2472	-0.1226	-0.0380	0.3815
_Wikipedia	0.5324	0.1974	0.5257	0.4596	-0.4420	1.0000	0.7320	0.7554	0.2600	0.6348	0.3112	0.2075	-0.0464
_Cible	0.4686	0.3097	0.4726	0.3635	-0.3658	0.7320	1.0000	0.9506	0.3930	0.8100	0.4341	0.1885	-0.0652
_Source	0.4794	0.3189	0.4923	0.4024	-0.3751	0.7554	0.9506	1.0000	0.4523	0.8308	0.4873	0.1838	-0.0627
_Prix	0.2954	0.1866	0.1910	0.1457	-0.1073	0.2600	0.3930	0.4523	1.0000	0.5965	0.9572	0.1469	0.0479
_Universités	0.5936	0.2672	0.3432	0.2923	-0.2472	0.6348	0.8100	0.8308	0.5965	1.0000	0.6100	0.2716	-0.0283
_Statut	0.3283	0.2292	0.1929	0.1531	-0.1226	0.3112	0.4341	0.4873	0.9572	0.6100	1.0000	0.2089	0.0373
_Véhicularité	0.2413	0.0394	-0.0048	0.0240	-0.0380	0.2075	0.1885	0.1838	0.1469	0.2716	0.2089	1.0000	0.0243
_Ecriture	-0.1105	0.0341	-0.1900	-0.1494	0.3815	-0.0464	-0.0652	-0.0627	0.0479	-0.0283	0.0373	0.0243	1.0000

TABLE 10. CORRELATION COEFFICIENTS, 634 BAROMETER LANGUAGES

The conclusion is that we have a set of reasonable factors three quarters of the correlation coefficients are less than 0.5

3.D Attenuation coefficients

By considering the factors that we have chosen and the classifications that we have made, the user of our work may consider that such and such a factor is not relevant or does not interest him and is too redundant with another. Not taking into account such and such a factor may be considered useful in the case where a particular problem is studied. Thus, if we have to decide in which languages, the menus, tutorials and help program of a new software must be written, we will retain more particularly the factors speakers, internet access and article in Wikipedia, which will make it possible to optimize the number of potential customers. We have therefore decided to use a set of "attenuating" coefficients which will be used as multipliers of the normalized values of the factors. These coefficients take a value between 0 and 1. The value 0 means that the factor is not taken into account, the value 1 that it is considered to be of primary importance. Any intermediate value is possible and is at the choice of the barometer user. The overall score that we will use to classify the languages will therefore be calculated by applying the formula:

$$\text{Score} = \text{ÿ}_{=1} \quad \text{ÿ}$$

in which the sign ÿ indicates that the sum is made over all the factors of the value fi of the ith factor multiplied by the attenuator coefficient wi chosen for this factor. This "global" score can vary continuously between 0, all the products wi* fi are zero) to 12, number of factors used, theoretical situation in which all the wi and all the fi would then be equal to 1.

4 Should all languages be classified?

We therefore have a file containing 6155 languages described by thirteen factors, which allows us to calculate a score and classify them all in relation to each other. Is this reasonable?

When we examine their values for the factors we have chosen, we find that many of them are only spoken in one country (thus their entropy is zero), or have no vehicular function, or have no official status, or have not given rise to any translation listed in the Translationium database, or have not received any literary prize, or have not given rise to any article in Wikipedia and are not taught in any university. More importantly, many languages combine several or almost all of these negative characteristics! Comparing them to each other doesn't make much sense. What is there in common between Mandarin and Faroese (900 million and 58,000 speakers), Hausa, an important vehicular language in the Sahelian strip taught in African and Oriental language schools and a language spoken by 2,000 people in a village in the Niger delta and unknown 20 kilometers away. What sense would there be in declaring that Papamiento, Creole of the Netherlands Antilles is ranked 127th and Hiri Motu, Pidgin of Papua New Guinea 678th ? Of course, we have to make a choice.

But to choose is necessarily to eliminate, to make people unhappy and also to expose oneself to making mistakes. In the two previous editions of the barometer, we relied on the number of speakers, 5 million in 2010 and 500,000 in 2012, which led us to include 137 and then 563 languages in the barometer. These choices seemed judicious to us at the time, our vision changed when we produced the 2017 edition of the barometer and we decided to make our selection criteria more complex.

4.A Choice based on the number of speakers

We explain in this paragraph the choice of the 634 languages retained in 2017. For reasons of consistency and ease of comparison, we have, in this edition, strictly retained the same 634 languages.

Consider Figure 5 which for the 6141 languages (2017 database) relates the total score to the logarithm of the number of speakers. To the right of the vertical red line are all languages with more than 500,000 speakers. The horizontal red line is at the value 2.5 for the overall score. This value is purely arbitrary, but it highlights the fact that a choice of 500,000 speakers "promotes" 434 languages (in blue in the figure) to the detriment of 182 others (in brown) which have a higher score in an overall ranking of all the languages of the world. We wanted to challenge this choice based solely on the number of speakers.

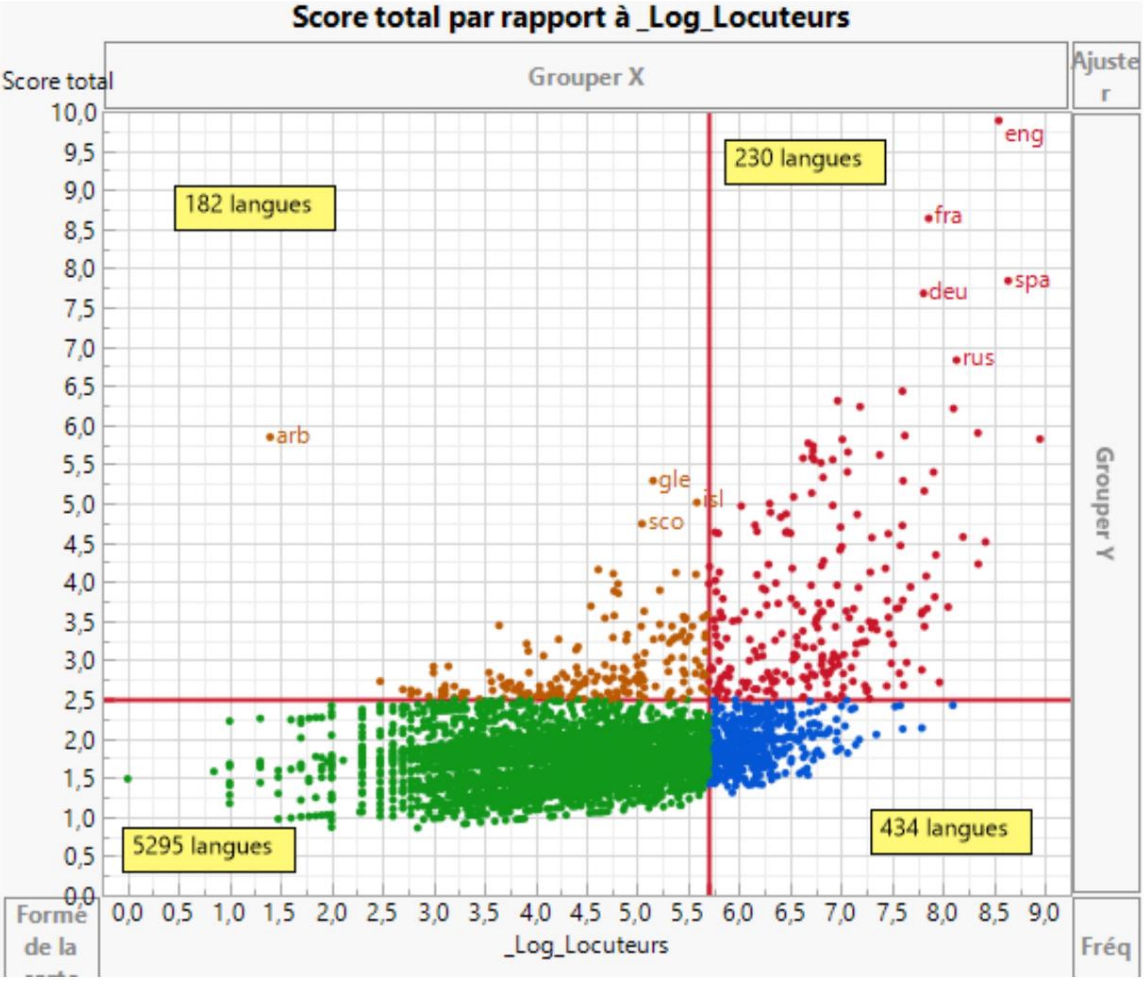


FIGURE 5. BAROMETER LANGUAGE SELECTION BASED ON THE NUMBER OF SPEAKERS

To reduce this inconvenience we had drawn a line that was no longer vertical but oblique joining the points of coordinates {0.6} and {8.0}. This eliminated the “blue” languages with the lowest scores and gave a chance to “brown” languages with high scores to be reinstated.
We added another condition: we only retained languages with more than 300,000 speakers. Figure 7 shows the result of these two criteria.

4.B Choice based on the importance of economic factors

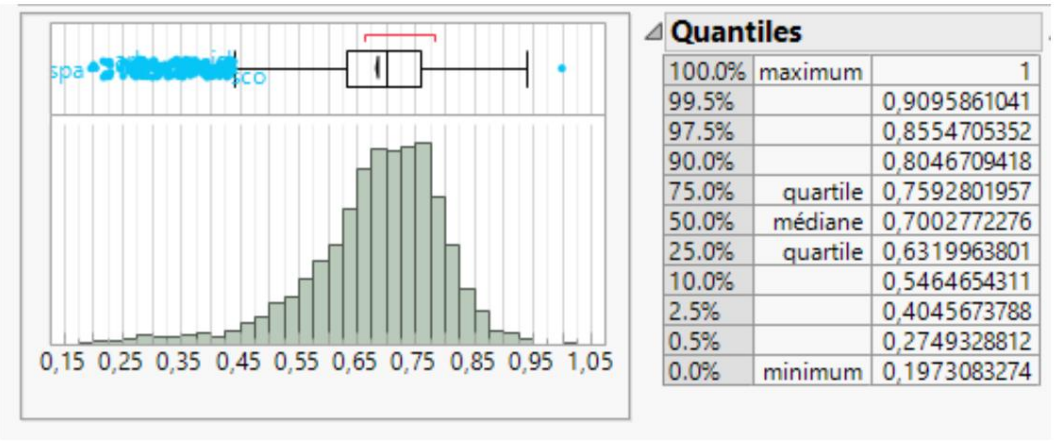
There was another problem that we wanted to deal with: the influence on the total score of the *situational parameters,* HDI, Internet and fertility. The “country” score (HDI + Internet + fertility) only depends on the country or countries in which the language in question is spoken.
Let's relate this country score to the total score. Figure 6 and tables 11 and 12 below lead us to the following thoughts:

The ratio varies between 0.20 and 0.91. The value of the median of this distribution told us that for one out of two languages this ratio is greater than 0.7. Worse still, it was above 0.5 for 5781 languages, or 94% of the total.

Table 11 indicates for example that a number of Australian indigenous languages would have been ranked in the top 600 out of 6141 and that this is not due to the language itself but to Australia which accounts for 88% of the score, it wouldn't have made sense.

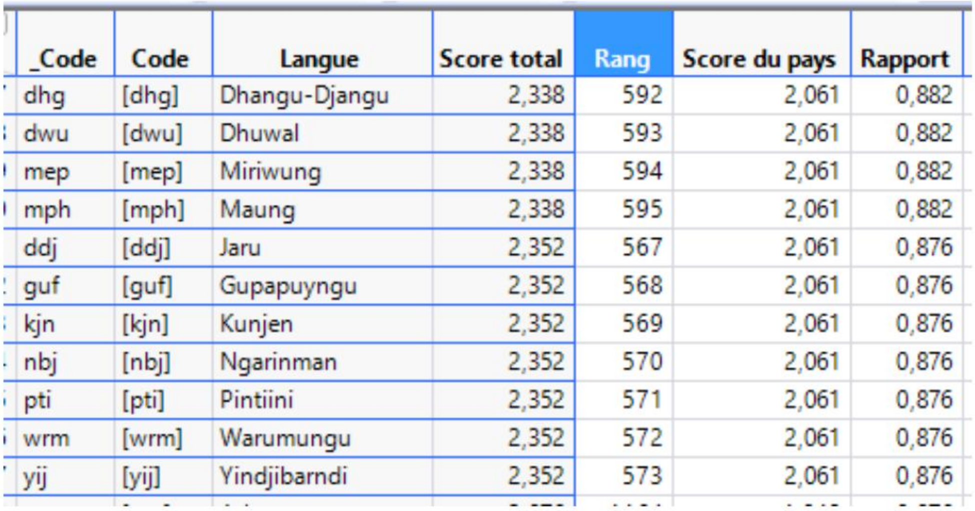
Conversely, Table 12 shows the twenty languages for which the ratio was the lowest. There were ten of the twenty languages best ranked by the barometer, which was satisfactory, the ranking of the most "heavy" languages only partially depended on geographical criteria, parameters that we qualify as contextual.

Our conclusion was that it seemed reasonable to define an upper limit for this ratio, we had chosen the value of 2/3. This eliminated languages that “benefit” from the country in which they are spoken: Aboriginal languages from Australia, Indian languages from Canada, Sami from Norway and many others.



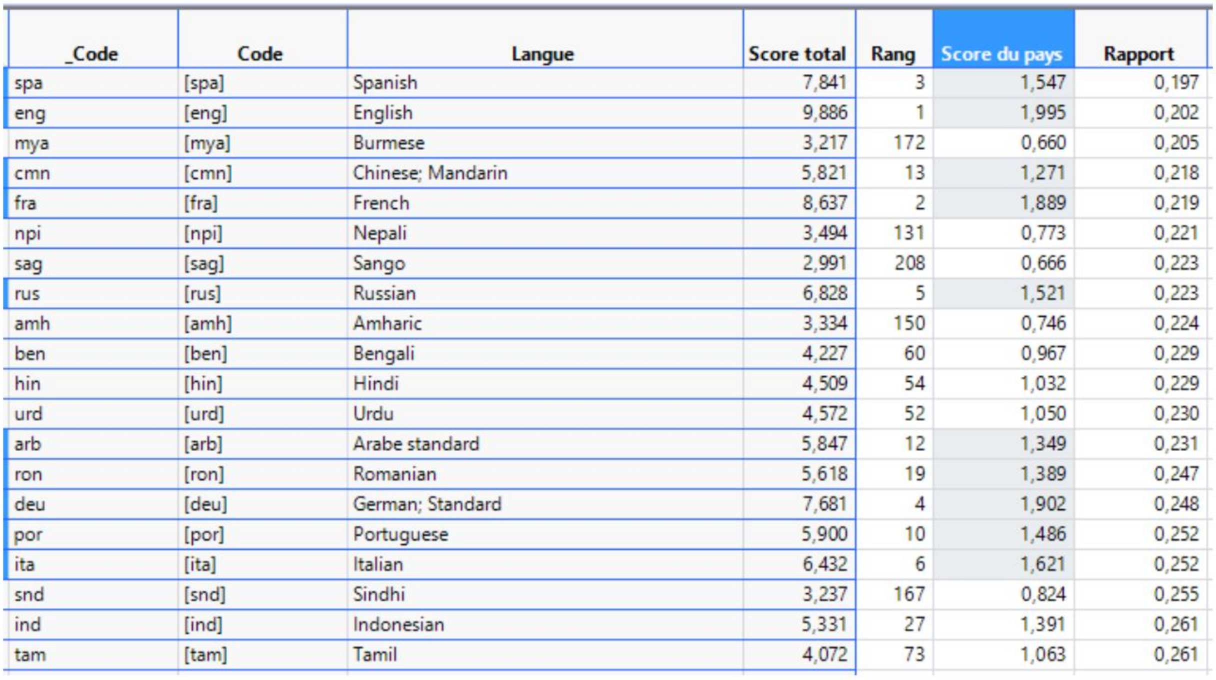
100.0%	maximum	1
99.5%		0,9095861041
97.5%		0,8554705352
90.0%		0,8046709418
75.0%	quartile	0,7592801957
50.0%	médiane	0,7002772276
25.0%	quartile	0,6319963801
10.0%		0,5464654311
2.5%		0,4045673788
0.5%		0,2749328812
0.0%	minimum	0,1973083274

FIGURE 6. DISTRIBUTION OF THE RATIO (CONJUNCTURAL SCORE)/(GLOBAL SCORE)



_Code	Code	Langue	Score total	Rang	Score du pays	Rapport
dhg	[dhg]	Dhangu-Djangu	2,338	592	2,061	0,882
dwu	[dwu]	Dhuwal	2,338	593	2,061	0,882
mep	[mep]	Miriwung	2,338	594	2,061	0,882
mph	[mph]	Maung	2,338	595	2,061	0,882
ddj	[ddj]	Jaru	2,352	567	2,061	0,876
guf	[guf]	Gupapuyngu	2,352	568	2,061	0,876
kjn	[kjn]	Kunjen	2,352	569	2,061	0,876
nbj	[nbj]	Ngarinman	2,352	570	2,061	0,876
pti	[pti]	Pintiini	2,352	571	2,061	0,876
wrm	[wrm]	Warumungu	2,352	572	2,061	0,876
yij	[yij]	Yindjibarndi	2,352	573	2,061	0,876

TABLE 11. HIGH (CONJUNCTURAL SCORE) / (TOTAL SCORE) RATIO



_Code	Code	Langue	Score total	Rang	Score du pays	Rapport
spa	[spa]	Spanish	7,841	3	1,547	0,197
eng	[eng]	English	9,886	1	1,995	0,202
mya	[mya]	Burmese	3,217	172	0,660	0,205
cmn	[cmn]	Chinese; Mandarin	5,821	13	1,271	0,218
fra	[fra]	French	8,637	2	1,889	0,219
npi	[npi]	Nepali	3,494	131	0,773	0,221
sag	[sag]	Sango	2,991	208	0,666	0,223
rus	[rus]	Russian	6,828	5	1,521	0,223
amh	[amh]	Amharic	3,334	150	0,746	0,224
ben	[ben]	Bengali	4,227	60	0,967	0,229
hin	[hin]	Hindi	4,509	54	1,032	0,229
urd	[urd]	Urdu	4,572	52	1,050	0,230
arb	[arb]	Arabe standard	5,847	12	1,349	0,231
ron	[ron]	Romanian	5,618	19	1,389	0,247
deu	[deu]	German; Standard	7,681	4	1,902	0,248
por	[por]	Portuguese	5,900	10	1,486	0,252
ita	[ita]	Italian	6,432	6	1,621	0,252
snd	[snd]	Sindhi	3,237	167	0,824	0,255
ind	[ind]	Indonesian	5,331	27	1,391	0,261
tam	[tam]	Tamil	4,072	73	1,063	0,261

TABLE 12. LOW (CONJUNCTURAL SCORE) / (TOTAL SCORE) RATIO

4.C Final choice of 634 languages

Figure 7 therefore visualized our choice. The selected languages are shown in red, the rejected ones in blue. The logarithm of the number of speakers is plotted along the abscissa. The vertical red line selects languages with more than three hundred thousand speakers. The oblique blue line reduces the inconvenience dealt with in paragraph 4.A. We added to this the need to have a ratio (cyclical score) / (total score) greater than 0.6667. The languages rejected by this last filter are those appearing in blue above and to the right of the two blue and red lines. It will be noted that "standard" Arabic, Gaelic and Scots and many others can be considered sacrificed.

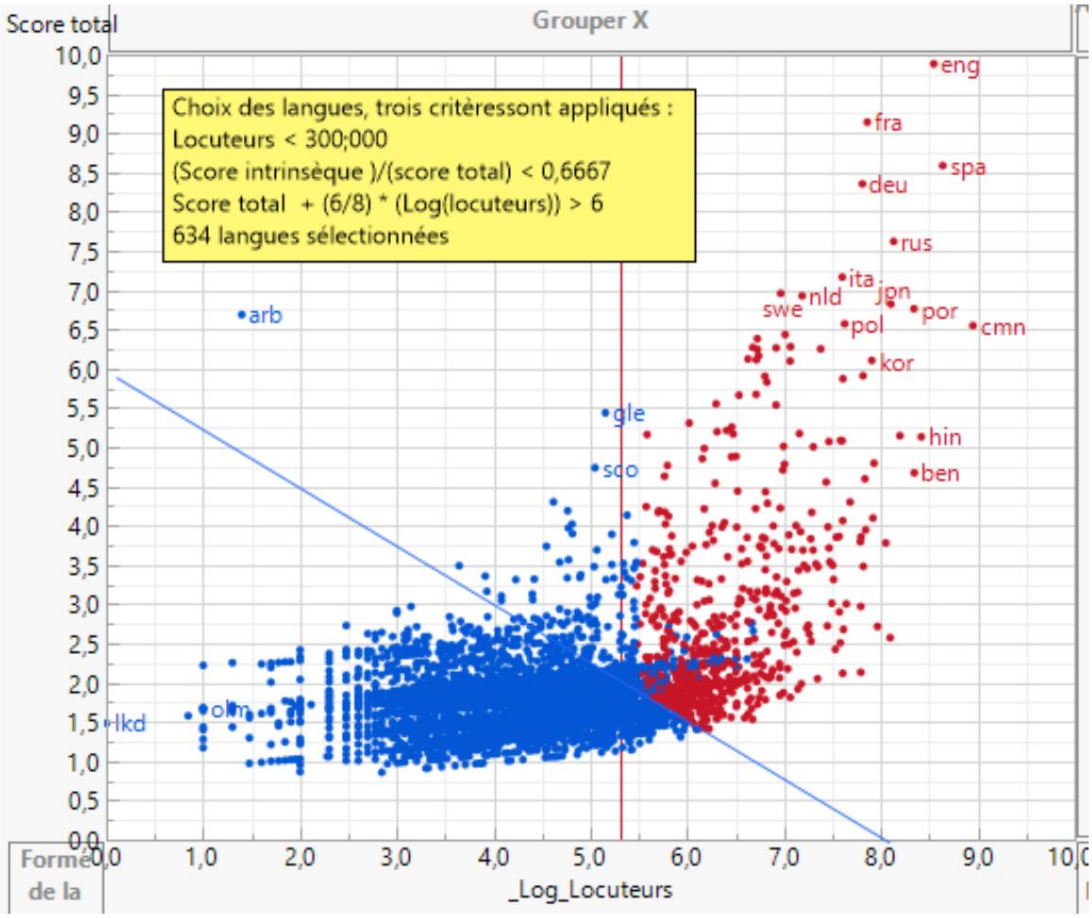


FIGURE 7 SELECTION OF LANGUAGES BASED ON THREE CRITERIA

We were then left with 634 languages, these are the ones that have been classified in our edition of the 2017 barometer. These are the ones that are classified in this current edition.

From this point we redefine the values 0 and 1 of all the factors which now correspond to the minimum and maximum values of the factors of the languages retained. The classification thus becomes a coherent internal classification of these languages and these languages alone.

On the diagram above, Standard Arabic [arb] appears as an outlier, isolated from all the others. Its score in all 6141 languages is very high and ranks it 12th .
However, in all the compilations of languages and their speakers, dialectal Arabic is retained as the L1 language of Arabic speakers and it is impossible to know the number of speakers of standard Arabic, assuming that there are any. The criterion of a minimum of 300,000 speakers therefore leads to its elimination from the final ranking. We are obviously aware of the fact that Arabic is a great language at the world level, official in many international organizations and heir to a great history and a great culture, but how to do it? The situation in Arabic-speaking countries seems somewhat schizophrenic, they speak a language that they don't write and they write a language that they don't speak. We are not responsible for this situation, but it is the reality.

5 Proper use of the barometer

As we said above, our barometer is intended to be flexible and it is always possible, by using the attenuating coefficients attached to the factors, to modulate the score in such a way that it answers the questions that the user asks himself. We now want to provide four examples of scores that we refer to as “standard scores”. They give four different points of view on the ranking of the languages of the world.

The top fifty languages for each of the four scores that we are going to describe are reported in Table 13.

5.A Global score

The score takes into account all the factors that we have defined above by assigning the coefficient 1 to all the factors, this is the maximum possible score that the considered language can reach. It gives an overview of the ranking of languages in the world. Strictly speaking, it is the measure of the weight of languages at the global level.

On examining the top of the ranking (table 8) our attention is drawn to the presence of a high proportion of European languages (cells colored in blue): ten, twenty and thirty-four in the ten, twenty-five and fifty s languages. At the top of the ranking we mainly find the languages of the countries which had constituted a colonial empire, but Japanese and Mandarin are also well placed. All these languages combine a high number of speakers, an abundant or recognized cultural production and they are spoken in countries with significant economic power. The importance of social economic factors is very apparent when considering the ranks of Dutch, Swedish, Norwegian, Finnish and Danish which are all between 9th and 24th place .

This global score is therefore a point of view but it is of course not the only one and we would now like to introduce other ways of judging the weight of languages.

5.B Intrinsic score

As we have seen, thousands of languages have null values on several or even the majority of the intrinsic parameters, those which do not depend on the countries in which these languages are spoken. This means that if they do not have a very high number of speakers most of their total score is due to contextual factors which relate to the countries in which the language is spoken. This phenomenon can be reproduced to a greater or lesser degree for most languages, which is why it seems interesting to consider a classification that only takes into account the ten factors that relate only to the language and no longer to its language. distribution area. We call it the intrinsic score.

When we refer to table 8 of the classifications, we observe a stability of the very first languages classified (green cells). The group of Nordic languages mentioned above is retreating as a whole (grey cells), they have lost the advantage conferred on them by their highly developed countries. We also observe the advance of the languages of the Indian subcontinent as well as Asian languages which progress by one or more ranks.

This classification can be considered as that of the languages of the future or even of the future. Asian and African languages can be considered as “penalized”, when we consider the global score, by the fact of belonging to less advanced countries than “Western” countries. When the countries in which they are spoken have caught up all or part of their economic and/or cultural “delays”, they will progress in the overall ranking.

5.C Demographic score

In developing countries, the economic level and the level of education taken into account by the contextual factors are not the only points which are expected to improve in the medium to long future. long term. It is likely that economic progress will lead to cultural progress taken into account by the translation flow, Wikipedia and literary price factors. To anticipate this evolution, it seems interesting to define a classification taking into account only the possible current strong points of the languages of these developing countries, namely the number of speakers and the vehicularity. We call this score demographic.

The important point here is the majority presence of the languages of the Indian subcontinent as well as Asian languages, we find 8, 19 and 40 of them respectively in the 10, 25 and 50 first languages classified. As a corollary, the European presence is greatly diminished.

5.D Prestige score

Similarly, the languages of developed countries have strengths that distinguish them from the languages of developing countries. The official status of languages like English or French comes from the importance of the former colonial empires and made them the language of the elites in a large number of countries. The high level of education in developed countries has as its corollary the recognition of this culture through the awarding of literary prizes as well as the development of translation flows. This aspect of the weight of languages is highlighted by the prestige score which is the sum of these three factors plus the university factor.

This classification is in a way complementary to the two previous ones and we observe the return of European languages, eight, nineteen and thirty-three in the top ten, twenty-five and fifty respectively.

It should also be noted that the Indonesian (bahasa indonesia) which in the three previous rankings was in 17th , 12th then 4th rank is now found in 36 th position.

To sum up, we can say that the global and prestige scores crown the "established" languages, while the intrinsic and demographic scores give a vision of what the panorama of world languages could be in the future.

The weight of languages in the world

Rank	Classification according to different scores			
	Total	Intrinsic	Demographic	Prestige
1	English	English	English	English
2	French	French	urdu	French
3	spanish	Spanish	French	Spanish
4	german	German	Indonesian	German
5	russian	Russian	yue	Russian
6	italian	Italian	Tagalog	Italian
7	Swedish	Portuguese	Javanese	Japanese
8	romanian	Romanian	Hindi	Mandarin
9	Portuguese	Mandarin	Thai	Portuguese
10	Polish	Polish	dyula	Polish
11	dutch	Swedish	tok pisin	Swedish
12	Catalan	Indonesian	Russian	Czech
13	czech	Catalan	Swahili	Dutch
14	Croatian	Swahili	Cameroonian English Creole	Korean
15	mandarin	Croatian	Mandarin	Danish
16	Hungarian	Czech	oromo, central west	Norwegian
17	indonesian	Dutch	Lingala	Hungarian
18	japanese	Hungarian	igbo	Hebrew
19	norwegian	farsi	bamanankan	Greek
20	finnish	Turkish	amharic	Serbian
21	turkish	Japanese	mòoré	Turkish
22	danish	urdu	Romanian	Finnish
23	Farsi	Finnish	north azeri	Croatian
24	Swahili	Serbian	German	Romanian
25	Slovak	Korean	Spanish	farsi
26	Serbian	Hindi	Zulu	Catalan
27	Tagalog	Norwegian	Hausa	Bulgarian
28	Estonian	Danish	Northern Sotho (Sepedi)	Hindi
29	Lithuanian	Slovak	oromo borana	Slovenian
30	korean	Vietnamese	gogo	Slovak
31	Slovenian	Tagalog	farsi	Estonian
32	ukrainian	Ukrainian	Afrikaans	Ukrainian
33	vietnamese	Bulgarian	north eastern thai	Lithuanian
34	Icelandic	Afrikaans	Xhosa	Icelandic
35	greek	Estonian	Krio	Bengali
36	urdu	Lithuanian	hiligaynon	Indonesian
37	Galician	Greek	tswana	urdu
38	basque	Slovenian	nyanja	Armenian

39	hindi	north azeri	eastern oromo	Macedonian
40	bulgarian	Kazakh	efik	Basque
41	hebrew	bosnian	Nigerian English Creole	Vietnamese
42	Kazakh	Bengali	dari	Northern Uzbek
43	North Azeri	North Uzbek	South Sotho (Sesotho)	tamil
44	Bosnian	Hausa	Armenian	Galician
45	welsh	Thai	wolof	Thai
46	Latvian	Malay	Assamese	Welsh
47	Hausa	Galician	Vietnamese	Belarusian
48	malaysian	Basque	Catalan	tibetan, central
49	Afrikaans	Armenian	asturian	bosnian
50	asturian	Albanian Tosk	Moluccan Malay	Georgian

TABLE 13. TOP RANKING FOR 4 STANDARD SCORES

5.D Personalized scores It is up to
you, the barometer user, to build your own score by adjusting the sliders of the attenuation coefficients. Each of the parameters can be assigned a coefficient that varies continuously between 0 and 1 depending on your vision of languages or the problem you are faced with.

Up to you !