

## Baromètre des langues dans le monde

1



**MINISTÈRE  
DE LA CULTURE**

*Liberté  
Égalité  
Fraternité*

Ce travail a été réalisé par MM. Alain Calvet et Louis-Jean Calvet en 2022 avec le soutien de la Délégation générale à la langue française et aux langues de France (ministère de la Culture)



Ce document est mis à disposition sous licence CC-BY-SA 3.0 : Attribution - Partage dans les Mêmes Conditions 3.0 France. Pour voir une copie de cette licence, visitez

<http://creativecommons.org/licenses/by-sa/3.0/fr/> .

## Sommaire

1. Introduction.....	3
2. Les facteurs décrivant le poids d'une langue .....	6
2.A Les facteurs intrinsèques à la langue.....	6
2.A.1 Le nombre de locuteurs .....	6
2.A.2 L'entropie.....	6
2.A.3 Le facteur véhiculaire .....	8
2.A.4 Le statut de la langue .....	13
2.A.5 Le nombre de traductions à partir de la langue.....	18
2.A.6 Le nombre de traductions vers la langue .....	20
2.A.7 Les prix littéraires internationaux .....	20
2.A.8 L'activité dans Wikipedia.....	22
2.A.9.L'enseignement au niveau universités.....	22
2.A.10. Systèmes graphiques.....	25
2.B Les facteurs contextuels .....	31
2.B.1 L'index de développement humain .....	31
2.B.2 L'indice de fécondité.....	32
2.B.3 La pénétration du réseau internet .....	32
3. Traitement des Données .....	33
3.A Normalisation des valeurs .....	33
3.B Utilisation des logarithmes .....	33
3.C Indépendance statistique entre les données.....	35
3.D Coefficients atténuateurs .....	37
4 Faut-il classer toutes les langues ? .....	37
4.A Choix basé sur le nombre de locuteurs .....	38
4.B Choix basé sur l'importance des facteurs conjoncturels .....	39
4.C Choix final de 634 langues .....	40
5 Du bon usage du baromètre.....	42
5.A Score global.....	42
5.B Score intrinsèque .....	42
5.C Score démographique.....	43
5.D Score prestige .....	43
5.D Scores personnalisés.....	45
6 Les macrolangues .....	<b>Erreur ! Signet non défini.</b>

## 1. Introduction

En avril 2010, nous mettions en ligne sur le site de l'Union Latine un « baromètre des langues du monde »<sup>1</sup>, que le lecteur consultera avec profit afin de comprendre la méthode que nous utilisons pour attribuer un « score » aux différentes langues prises en compte et déterminer leur « poids ».

Il peut paraître vain de "classer" ainsi des langues et beaucoup pensent d'ailleurs que la langue la plus importante au monde est leur langue maternelle, la langue dans laquelle ils ont fait leurs études ou encore la langue qui leur sert à communiquer avec leurs proches. En fait ces trois fonctions (langue maternelle, langue de scolarisation, langue familiale) peuvent parfois être remplies par des langues différentes chez un même individu, et c'est justement en partant des différents rôles et des différents usages des langues dans la société que nous avons réalisé ce baromètre. Chaque langue y est caractérisée par des facteurs dont la valeur peut être continue ou discrète, chaque facteur pouvant être pris en compte, écarté, ou affecté d'un coefficient atténuateur qui diminuera son importance relative par rapport aux autres facteurs. Chaque utilisateur peut ainsi créer son propre classement en fonction du point de vue qui l'intéresse et/ou de son appréciation personnelle de l'importance ou du bien-fondé des facteurs proposés.

Une deuxième édition du baromètre a été mise en ligne en 2012<sup>2</sup>. Elle permettait le classement d'un nombre de langues plus important et utilisait un facteur de plus pour effectuer ce classement.

Une troisième édition a été réalisée en 2017 avec un plus grand nombre de langues et de facteurs.

Nous mettons aujourd'hui en ligne une quatrième édition de notre baromètre. Nous utilisons maintenant treize facteurs et classons 634 langues.

Le propos de ce document est de fournir à qui consultera le site l'information suffisante pour bien comprendre la manière dont nous l'avons construit et dont il peut l'utiliser pour apprécier l'importance relative des langues dans le monde, pour apprécier *leurs poids*

La première difficulté est de déterminer de quelle langue parlons-nous ? Ont été répertoriées vingt-sept variantes de nahuatl ainsi que de malais, onze de malgache, trois tonga qui n'ont aucun rapport entre eux, trois yaka, deux ndébélé, deux sotho, deux azéri, deux panjabi et bien d'autres cas litigieux. La liste est trop longue pour qu'il soit raisonnable de la donner de manière exhaustive. Cette situation crée une difficulté car bien des données concernant les langues ne précisent pas de quelle variante il s'agit. Le cas typique est l'arabe. Il existe un arabe dit « standard » et trente-quatre arabes dits dialectaux qui sont en fait les langues maternelles de tous ou presque tous les arabophones. Si par exemple nous cherchons la présence de l'arabe sur Wikipedia nous trouvons

---

<sup>1</sup> <http://portalingua.info/fr/poids-des-langues/>

<sup>2</sup> <http://wikilf.culture.fr/barometre2012/>

## Le poids des langues dans le monde

---

plus de 500.000 pages en « arabe » sans autre spécification et 17.000 en arabe « égyptien ». De manière similaire si nous nous intéressons au nombre de traductions à partir de l'arabe environ 12.000 concerne l'arabe, 5 l'arabe tchadien, 3 l'arabe marocain et 1 les dialectes arabes. Il est difficile de considérer ces nombres comme reflétant la réalité. Pour traiter ce problème nous utilisons la norme ISO 693-3 qui affecte à chaque langue un code de trois lettres, l'arabe standard y est désigné par [arb], l'arabe marocain par [ary], l'arabe égyptien par [arz] etc.

Ainsi que nous le verrons plus loin il apparaît difficile de classer entre elles des langues dont les caractéristiques sont aussi différentes que par exemple le chinois mandarin, parlé par près de neuf cents millions d'individus, l'espagnol par environ quatre cent trente millions, l'arabe standard, langue officielle de plus de vingt états mais langue maternelle d'un nombre d'individus très faible sinon quasi nul, le norvégien, langue d'un pays peu peuplé mais riche et cultivé au sens occidental du terme, le swahili parlé par quelque millions d'individus mais langue véhiculaire d'une partie importante du continent africain. Et que dire des familles ou groupe de langues. Par exemple, dans le groupe S30 des langues bantoues classées par M. Guthrie les sésotho, sépédi et tswana) doivent-ils être considérés comme une seule langue ou trois langues différentes. Le complexe des langues peut-il être regroupé en une seule langue qui regrouperait environ vingt millions de locuteurs ou bien rester éclaté entre une dizaine de langues (dialectes ?). La même question peut être posée pour de nombreux autres groupes de langues voisines et présentant un haut niveau d'inter-intelligibilité. Que penser de la notion de macrolangue proposée par le SIL ? Autant de questions dont les réponses varient d'un auteur à l'autre, d'un point de vue à l'autre. Le classement dépend forcément de ces choix qui peuvent évidemment être critiqués.

4

Nous nous en sommes donc strictement tenus à la classification dite ISO 693-3<sup>3</sup> qui bien qu'imparfaite présente l'avantage d'être cohérente et d'assurer une correspondance biunivoque entre une langue et un code de trois lettres. Il est ainsi possible de résoudre la plupart des cas litigieux. L'utilisation d'un code unique clarifie la situation. Malheureusement toutes les compilations de données n'utilisent pas cette nomenclature. Ainsi certaines sources distinguent le tagalog [tgl] du filipino [fil] et affectent à chacune de ces langues entre vingt et vingt-cinq millions de locuteurs ce qui est en désaccord avec par exemple le site de Jacques Leclerc "L'aménagement linguistique dans le monde"<sup>4</sup> qui en fait une seule langue avec environ vingt-cinq millions de locuteurs. Parfois une source cite, par exemple, le guarani, le quichua ou le quechua sans plus de précision. Or Ethnologue décrit huit variétés de guarani, douze de quichua et une trentaine de quechua!

Il nous faut donc décider quelles langues nous examinerons et pour cela trouver une compilation fiable des langues du monde. À notre connaissance il en existe au moins trois accessibles sur le web, Ethnologue<sup>5</sup>, Joshua<sup>6</sup> et People Groups<sup>7</sup>. Le problème est que ces trois sites mélangent une volonté

---

<sup>3</sup> <http://www.ethnologue.com/web.asp>

<sup>4</sup> <http://www.axl.cefan.ulaval.ca/>

<sup>5</sup> <https://www.ethnologue.com/>

<sup>6</sup> <http://www.joshuaproject.net/>

<sup>7</sup> <http://www.peoplegroups.org/>

## Le poids des langues dans le monde

---

scientifique de compilation des langues et de leurs locuteurs et un aspect religieux plus ou moins présent. Dans nos deux premières éditions nous avons utilisé Ethnologue comme notre source de données, malheureusement la politique du SIL a changé et les données ne sont plus en accès libre. Nous nous sommes donc tournés vers le site Joshua et avons constitué une base de données de 17432 enregistrements. Joshua cite les sources de ces données, elles sont Ethnologue 19<sup>ème</sup> édition dans 17398 cas, 34 enregistrements ne documentant pas leur source. Finalement nous avons obtenu une base de 6155 langues différentes prenant en compte un peu moins de 7,8 milliards de locuteurs. Nous en avons éliminé les lignes relatives aux différentes langues des signes et celles indiquant « langage inconnu ». Finalement notre fichier de travail contient au départ 6102 langues regroupant environ 7,7 milliards de locuteurs.

Il faut ensuite bien garder à l'esprit que les données que nous manipulons sont dynamiques, elles changent, se transforment, disparaissent. En effet, qu'en est-il aujourd'hui du Séné (code iso 693-3 [sej]), une langue de Papouasie Nouvelle-Guinée) pour laquelle 10 locuteurs ont été répertoriés en 1978 ? Ou encore du Berakou ([bxc], Tchad) qui comptait 2 locuteurs en 1995 ? Que restera-t-il dans quelques années, ou même que reste t'il aujourd'hui des 360 locuteurs du Nunggubuyu ([nuy], recensement australien de 1996), que l'Atlas des langues en danger dans le monde publié par l'UNESCO en 2010<sup>8</sup> signale comme "sérieusement en danger" ? A l'inverse l'existence des créoles, des pidgins des diverses formes de français des pays d'Afrique, du filipino, du bahasa indonesia montre que des langues nouvelles apparaissent.

Dans l'état actuel des bases de données il est impossible d'obtenir un état complet et cohérent de la situation à une date donnée, donc de répondre avec précision à ces questions, on ne peut qu'exploiter au mieux les données disponibles, en un mot : "faire avec ce que l'on a", même si ce n'est pas très satisfaisant.

Pour établir un classement de ces 6102 langues la seule difficulté consiste à fournir la somme de travail nécessaire pour exploiter les données disponibles. Déclarer ensuite que ce classement a un sens est un autre problème. Classer et donc comparer aux autres des langues dont nous n'avons aucun moyen de vérifier l'existence en temps réel est un exercice purement théorique. Nous l'avons fait, mais nous ne pensons pas qu'il ait une signification absolue. Nous verrons que les classements que nous pouvons établir sont très dépendants du contexte dans lequel nous nous plaçons et il nous semble donc plus raisonnable d'établir des classements partiels en filtrant les données sur un ou plusieurs critères, comme le nombre de locuteurs, l'IDH des pays dans lesquels la langue est parlée ou l'équipement internet etc... Ceci permet d'établir des comparaisons ayant plus de sens dans des contextes mieux définis. Nous y reviendrons plus bas.

Le nom des langues pose parfois problème. Nous utilisons le nom français de la langue autant que possible en accord avec Le « Dictionnaire des langues »<sup>9</sup> Mais cela ne change rien à la liste des langues, qui sont définies de manière unique grâce au code à trois caractères, ISO 639-3 dont nous faisons un large usage.

---

<sup>8</sup> Atlas des Langues en danger dans le monde, Editeur Christopher Moseley, Editions UNESCO, 2<sup>ème</sup> édition 2010 et site Internet <http://www.unesco.org/culture/languages-atlas/index.php?hl=fr&page=atlasmap>

<sup>9</sup> Dictionnaire des langues, E. Bonvini et alii, Quadrige P.U.F. 2011.

Les problèmes on le voit sont nombreux et nous détaillerons la manière dont nous les avons résolus. Nous proposerons tout d'abord un certain nombre de facteurs permettant de décrire une langue de manière "quantitative". Ces facteurs n'ont pas de caractère *linguistique*, mais permettent d'apprécier l'importance, le *poids*, d'une langue selon différents points de vue. Ensuite nous décrirons comment traiter ces facteurs pour les rendre comparables entre eux. Enfin nous proposerons diverses méthodes pour combiner ces descripteurs et arriver à des classements fondés sur la méthodologie utilisée dans notre Baromètre Calvet des langues du monde.

## 2. Les facteurs décrivant le poids d'une langue

Les facteurs que nous proposons n'ont donc pas de caractère purement linguistique et peuvent être séparés en deux catégories

En premier lieu nous décrirons ceux qui se rapportent à une langue et à elle seule, nous les appellerons *facteurs intrinsèques*. Le nombre de locuteurs est bien sûr le premier de ces facteurs mais il est possible d'en imaginer d'autres que nous décrivons plus bas.

Mais les langues vivent dans un environnement qui influe sur leur importance et leur développement, c'est pourquoi nous considérerons ensuite des facteurs décrivant les pays dans lesquels les langues sont parlées, ils sont alors *théoriquement* communs aux diverses langues parlées dans un même pays et une même langue parlée dans plusieurs pays bénéficie de l'apport de chacun de ceux-ci. Nous les qualifierons de *facteurs contextuels* et en retiendrons trois.

### 2.A Les facteurs intrinsèques à la langue

#### 2.A.1 Le nombre de locuteurs

Il s'agit des locuteurs de langue première, tels que répertoriés dans notre base de données ce qui nous l'avons vu plus-haut pose parfois problème. En outre ces données concernent les locuteurs L1 (langue première), ce qui a le désavantage d'occulter le caractère véhiculaire qui fait partie de nos facteurs. Par exemple, le swahili a environ trois millions de locuteurs en langue première mais plusieurs dizaines de millions en langue seconde ce qui en fait une langue très importante dans toute l'Afrique de l'Est. Un autre problème est que le nombre d'habitants des pays ne se recoupe pas avec les autres sources de données démographiques. En outre la somme des locuteurs de langue première des langues du pays est parfois différente de nombre des habitants du pays, nous l'avons dit il est difficile d'avoir des données exactes.

Il existe une autre source d'approximation, personne ne connaît avec précision le nombre d'habitants dans le monde et certains géographes considèrent que 25% des naissances et des décès dans le monde ne donnent pas lieu à une déclaration aux services de l'état civil. Il semble donc illusoire de vouloir rechercher une précision dans ces données, elles ne sont qu'approximatives.

#### 2.A.2 L'entropie

L'entropie est une notion qui a été introduite en thermodynamique statistique par Ludwig Boltzmann au 19<sup>ème</sup> siècle. Elle permettait de comprendre le passage d'une dynamique microscopique réversible à une évolution macroscopique irréversible. Elle a ensuite été utilisée en théorie de

# Le poids des langues dans le monde

l'information, en linguistique<sup>10</sup> et aujourd'hui dans les théories de prédiction de l'évolution de l'univers. C'est un concept subtil qui, suivant le contexte où l'on se place, peut être interprété comme une quantification du désordre, de l'imprédictibilité ou de l'information inaccessible.

Nous l'utilisons ici pour différencier une langue parlée dans un seul pays d'une langue parlée dans plusieurs pays. Le chinois mandarin ou l'hindi sont parlés en L1 par des centaines de millions de locuteurs mais 98% d'entre eux vivent dans un seul pays, alors que l'espagnol est parlé dans 72 pays dont le plus important ne regroupe que 27% de ses locuteurs et huit plus de 5%. Du point de vue « entropique » la distribution des locuteurs de l'espagnol est plus « désordonnée ». La conclusion que nous en tirons est que l'utilité, le rayonnement international de l'espagnol sont supérieurs à ceux du chinois mandarin ou de l'hindi.

Nous appellerons  $p_i$  la proportion des locuteurs d'une langue donnée vivant dans chacun des pays concernés.

Classiquement l'expression mathématique de l'entropie est la suivante :

$$\text{Entropie} = -\sum(p_i \times \text{Log}(p_i))$$

dans laquelle  $p_i$  est la probabilité pour un système de se trouver dans un état donné et  $\text{Log}(p_i)$  le logarithme naturel de cette probabilité, le symbole  $\Sigma$  indique que l'on fait la somme de tous les états  $p_i$  possibles. Dans notre cas nous utilisons évidemment  $p_i$  ainsi que défini plus haut, la proportion des locuteurs de la langue considérée dans chacun des pays où elle est parlée. La valeur minimale de cette fonction est zéro, lorsque la langue en question n'est parlée que dans un seul pays, et il n'existe pas de valeur maximale définie.

Considérons une langue parlée très majoritairement (98%) dans un pays et dont quelques locuteurs vivent dans un second, l'entropie sera :

$$(0,98 \times \text{Log}(0,98) + 0,02 \times \text{Log}(0,02)) = 0,098$$

Une langue dont les locuteurs sont répartis de manière égale sur trois pays aura une entropie de :

$$(0,33 \times \text{Log}(0,33) + 0,33 \times \text{Log}(0,33) + 0,34 \times \text{Log}(0,34)) = 1,099$$

Voyons à présent dans le tableau 1 ci-dessous quelques exemples réels, ceux du russe, du japonais, de l'anglais, de l'espagnol, de l'arabe standard et du chinois mandarin :

Langue	Russe	Japonais	Anglais	Espagnol	Arabe standard	Chinois mandarin
Entropie	0.667	0.116	1.159	2.543	2.7 ?	0.160
Locuteurs L1	136 M	124M	362M	443M	?	921M

TABLEAU 1. ENTROPIE ET NOMBRE DE LOCUTEURS L1 DE QUELQUES LANGUES

<sup>10</sup> <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>

## Le poids des langues dans le monde

---

Le russe et le japonais ont des valeurs similaires en ce qui concerne le nombre de locuteurs mais le japonais est peu parlé à l'extérieur du Japon alors que des communautés russophones existent dans les pays de l'ex Union Soviétique. Le russe garde un caractère de « langue impériale » et son entropie est supérieure à celle du japonais. L'anglais et l'espagnol sont comparables en ce qui concerne le nombre de leurs locuteurs. L'espagnol est la langue première de nombreux pays d'Amérique latine de taille moyenne alors que la majorité des locuteurs de l'anglais se concentre dans deux pays, les États-Unis et le Royaume Uni, l'entropie de l'espagnol est donc bien supérieure. L'arabe standard n'a pratiquement pas de locuteurs en langue première mais est considérée comme une langue présente dans tous les pays arabo-musulmans, son entropie, impossible à calculer avec exactitude serait élevée. Le chinois mandarin est la langue la plus parlée dans le monde mais une proportion très faible de ses locuteurs en L1 vit hors de Chine, d'où sa faible entropie. On comprend alors bien que l'entropie quantifie le « désordre », la diversité de la répartition des locuteurs voire la tendance à « l'universalité » d'une langue.

L'entropie n'a rien à voir avec le nombre global de locuteurs d'une langue, mais bien avec la façon dont ces locuteurs sont répartis dans l'aire ou les aires dans lesquelles cette langue est parlée. Elle est calculée à partir des données de population décrites plus haut.

### 2.A.3 Le facteur véhiculaire

Définissons tout d'abord deux notions. On qualifie le plus souvent de « langue maternelle » la première langue acquise par un individu, mais cette appellation est erronée car dans certaines familles plurilingues cette langue « maternelle » peut être celle du père (et il faudrait alors parler de langue « paternelle »), c'est en tout cas la langue parlée à la maison dans la petite enfance ou celle dans laquelle un individu réfléchit et s'exprime le plus naturellement. Nous utiliserons le qualificatif de langue première que nous noterons L1. Par ailleurs il est fréquent que des gens étudient d'autres langues dans leur parcours scolaire (ou parle alors de « langue étrangère »), utilisent tous les jours dans leurs pratiques sociales ou professionnelles une langue qui n'est pas leur L1 (on parle alors de « langue seconde ») ou acquièrent de façon informelle dans leur vie quotidienne des rudiments de langues présentes dans leur environnement. Les situations sont ici extrêmement variées. On peut apprendre à l'école une ou deux langues «étrangères» et les parler plus ou moins bien (c'est le cas de la France), on peut également acquérir à l'école la langue officielle du pays que l'on utilisera quotidiennement (c'est le cas du français en Afrique francophone, de l'anglais en Afrique anglophone, etc.), on peut apprendre une langue L2 « identitaire », c'est le cas du gaélique irlandais et on peut enfin apprendre sur le tas différentes langues que l'on n'utilise que dans des domaines réduits, par exemple commerciaux (c'est le cas de certains commerçants, dans les souks de Marrakech ou dans le bazar d'Istanbul), etc. Tout en sachant que ces situations sont différentes et méritent d'être traitées de façon spécifique, nous parlerons ici de façon générale de langue seconde, notée L2 pour toutes les situations dans lesquelles une langue autre que la L1 est couramment utilisée dans la vie extérieure à la vie familiale proche.

Le nombre de locuteurs qui ont une langue donnée pour L1 est évidemment un facteur important pour déterminer le poids de cette langue. Mais tout aussi important est celui des locuteurs qui la parlent comme L2, ce dernier pouvant même être plus élevé que le premier. Le nombre de locuteurs L1 du swahili, nous l'avons dit plus haut, pourrait faire croire qu'il s'agit d'une langue mineure et pourtant le swahili est une langue de communication majeure en Afrique de l'est, parlée par plusieurs dizaines de millions d'individus qui ont une autre langue pour L1.

# Le poids des langues dans le monde

---

Pour quantifier ce phénomène il est possible d'imaginer plusieurs méthodes. Nous pourrions recenser dans le monde les enseignants en langue vivante dans les écoles, collèges, lycées et universités ou bien recenser les élèves et étudiants. Mais de telles approches seraient limitées aux grandes langues reconnues par les ministères de l'éducation nationale et laisseraient de côté ce que nous souhaitons prendre en compte : la fonction *véhiculaire* de certaines langues, il s'agit là d'un fait de société, qui ne résulte pas d'une quelconque décision gouvernementale, administrative ou universitaire, mais qui se constate.

C'est ce fait que nous voulons quantifier en introduisant la notion de "taux de véhicularité" que nous définirons comme le rapport du nombre de locuteurs utilisant cette langue comme langue seconde au nombre total de locuteurs.

$$\text{Taux de véhicularité} = \frac{L2}{L1+L2}$$

Ce taux varie entre 0, pour une langue qui n'a que des locuteurs en L1 et 1, pour une langue dont tous les locuteurs la parleraient comme L2. Ainsi présentées, les choses peuvent paraître simples, mais elles se complexifient très vite lorsque nous abordons le problème des données, à la fois parce que les états ont parfois certaines réticences à reconnaître leur diversité linguistique et parce que les sources sont souvent imprécises si ce n'est inexistantes.

Nous nous sommes d'abord tournés vers les sources dont nous avons extrait le nombre de locuteurs en L1, puis vers quelques autres plus spécifiques des langues secondes et véhiculaires, mais elles sont en général mal documentées.

- Ainsi, Ethnologue utilise l'expression L2 et dans sa 20<sup>ème</sup> édition indique pour un nombre important de langues le nombre de locuteurs en L1 et en L2.

-Ethnologue indique parfois que la langue considérée est utilisée en L2 par les locuteurs d'une ou plusieurs autres langues. C'est le cas par exemple du mòré dont Ethnologue indique : «*used as L2 by...* » suivi d'une douzaine d'ethnies. Dans un tel cas *et quand cela a un sens*, nous retenons la moitié des locuteurs de ces langues indiquées pour calculer le facteur véhiculaire. Mais cela n'a parfois aucun sens. Expliquons-nous par deux exemples :

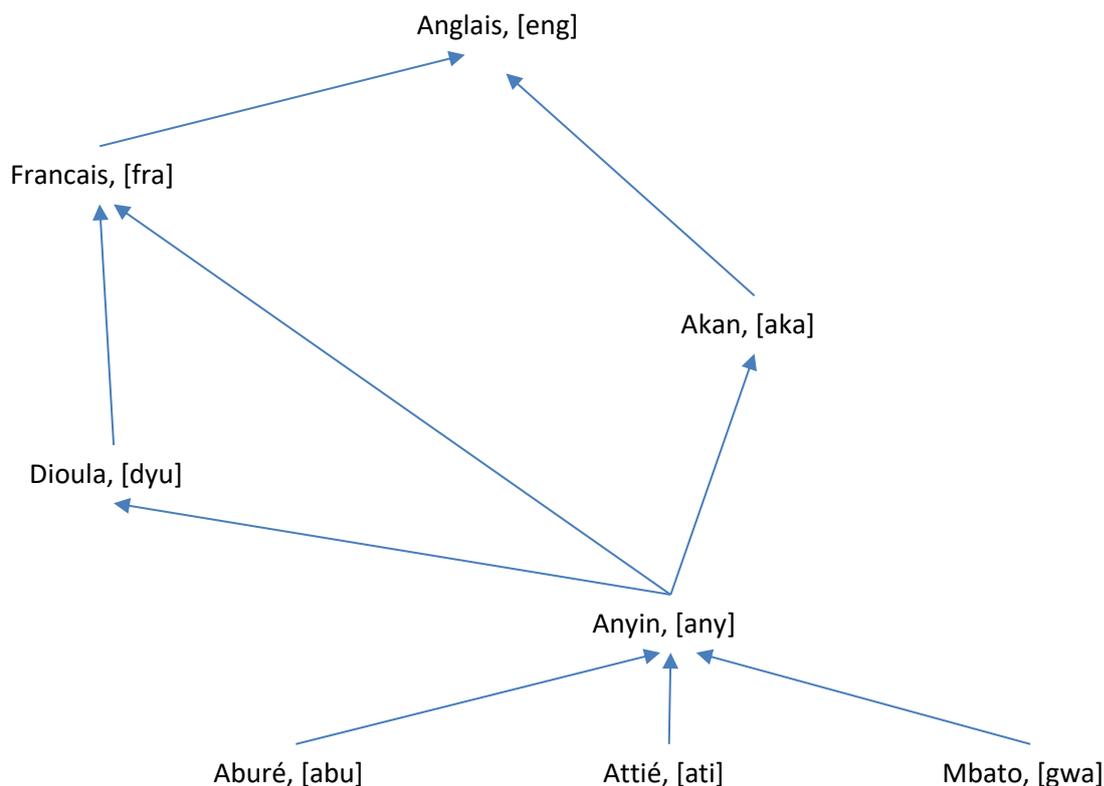
Pour le sarde logudorese ([src]) Ethnologue indique qu'il est utilisé en L2 par le catalan([cat]). Il s'agit bien entendu non pas de l'ensemble des locuteurs du catalan au niveau mondial, ils se comptent en millions, mais bien de ceux vivant en Sardaigne dans la région d'Alghero, ils sont 23000.

Toujours selon Ethnologue l'ouïghour ([uig], locuteurs L1 12,5 millions) est utilisé en L2, entre autres, par les locuteurs du mongol périphérique ([mvf], L1 2,8 millions de locuteurs L1) et du russe ([rus], locuteurs L1 136 millions). Il est clair que l'ouïghour n'est pas utilisé en L2 par 68 millions de locuteurs du russe mais seulement par les russes vivant au Xinjiang dont il faut connaître ou estimer le nombre.

Les situations analogues à ces deux dernières sont assez nombreuses, nous les traitons en cherchant les données publiées sur le nombre de locuteurs L1 de la langue considérée vivant au voisinage de celle utilisée comme L2. Ou en l'estimant lorsque cela est possible. En l'absence de cette information nous ne considérons pas qu'une langue A dont le nombre de locuteurs L1 est nettement supérieur à celui d'une langue B utilise cette dernière comme langue L2.

## Le poids des langues dans le monde

Il faut donc qu'il existe une « hiérarchie » entre les langues L1 et L2. La figure 1 ci-dessous, bien entendu incomplète, montre les relations L1/L2 à partir d'une langue de Côte d'Ivoire, l'anyin. Cette notion de hiérarchie y apparaît clairement. Les locuteurs L1 de l'anyin utilisent le dioula, le français et/ou l'akan comme langue L2. Mais l'anyin est utilisé en L2 par les locuteurs de plusieurs langues et l'anglais est la langue L2 « ultime ».



10

**Figure 1. Schéma simplifié des relations L2/L1/L2 de l'anyin avec d'autres langues**

-Dans certains cas il est impossible de connaître le nombre de locuteurs L2. Pour l'ebira ([igb]) Ethnologue indique : « *Other language speakers use ebira to communicate with ebira people* » sans aucune autre indication. Dans ce cas nous considérons la langue comme non véhiculaire.

-Les données sont parfois à la limite du contradictoire. Ainsi en Turquie Ethnologue indique que le zazaki méridional est utilisé comme langue L2 par le kurde du Nord et que le kurde du Nord est utilisé en langue L2 par le zazaki méridional ! Il existe en Turquie 1,7 million de locuteurs du zazaki plus de 15 millions du kurde. Mais aucune donnée sur leur utilisation en L2. Que faire dans une telle situation ? Dans ce cas particulier nous avons considéré, que le zazaki que certains considèrent comme un dialecte du kurde n'est pas une langue véhiculaire.

# Le poids des langues dans le monde

---

- Les études par pays du site Laval sont également utiles dans ce domaine. Elles utilisent souvent le terme "véhiculaire" citent les langues concernées mais pas toujours le nombre de locuteurs. Par exemple nous trouvons dans l'article sur le Bénin la phrase suivante:

*La plupart des Béninois utilisent le français, le fon, le yorouba ou le bariba comme l'une des langues véhiculaires.*

mais aucune indication n'est donnée quant au nombre de locuteurs.

- Certains sites universitaires<sup>11,12</sup> fournissent des notices décrivant les langues et parfois le nombre de locuteurs en L1 et L2.

- L'utilisation de mot clés tels que "secondary speakers" ou autres dans un moteur de recherche conduit à des sites recensant les langues les plus importantes (> 3000000 de locuteurs) et indiquant le cas échéant un nombre de locuteurs secondaires et une référence originale pour cette donnée. Il faut cependant être prudent et s'assurer que le terme secondary speakers correspond bien à des locuteurs de L2.

- Les sites gouvernementaux se rapportant aux résultats de référendums sont également utiles dans les cas où les langues maternelles et secondes sont documentées<sup>13</sup>.

L'information a donc été extraite de ces différentes sources mais nous sommes souvent trouvés confrontés à des problèmes de définition et probablement de "nationalisme linguistique". Ainsi l'anglais L2 est parlé par cent soixante-sept millions d'individus selon une source<sup>14</sup> et plus de six-cents millions selon une autre<sup>15</sup>. Le français est parlé en seconde langue par environ cinquante millions<sup>16</sup>, ou cent cinquante-trois millions<sup>17</sup>. Ces diverses sources ne parlant évidemment pas de la même chose, ou n'utilisant pas les mêmes critères, quel nombre devons-nous retenir ? Par ailleurs, les langues "mineures" sont totalement laissées de côté. Par exemple, s'il est possible de trouver une estimation du nombre d'utilisateurs secondaire du hiri motu, du tok pisin et de l'anglais en Papouasie Nouvelle-Guinée, le recensement des locuteurs L2, s'il en existe, des langues de niveau "inférieur" parmi les plus de huit cents parlées dans le pays est d'une difficulté insurmontable.

On le voit il n'existe pas de source cohérente et complète concernant les langues véhiculaires et le nombre de leurs locuteurs en langue première et seconde. Il nous a fallu collecter les données disponibles et bâtir ensuite un ensemble le plus raisonnable possible.

---

<sup>11</sup> <http://www.lmp.ucla.edu/Profile.aspx?menu=004>

<sup>12</sup> <http://nalrc.wisc.edu/>

<sup>13</sup> <http://censusindia.gov.in/2011-common/censusdataonline.html>

<sup>14</sup> [http://www.nationsonline.org/oneworld/most\\_spoken\\_languages.htm](http://www.nationsonline.org/oneworld/most_spoken_languages.htm)

<sup>15</sup> [http://en.wikipedia.org/wiki/World\\_language](http://en.wikipedia.org/wiki/World_language) et références citées

<sup>16</sup> [http://www.nationsonline.org/oneworld/most\\_spoken\\_languages.htm](http://www.nationsonline.org/oneworld/most_spoken_languages.htm)

<sup>17</sup> [https://en.wikipedia.org/wiki/List\\_of\\_languages\\_by\\_total\\_number\\_of\\_speakers](https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers)

## Le poids des langues dans le monde

Les exemples regroupés dans le tableau 2 ci-dessous donnent un aperçu de la méthode que nous avons appliquée.

Pour le premier, l'amharique, il n'y a pas d'accord entre les diverses sources mais le site Laval donne des valeurs pour le nombre de locuteurs en L1 et L2. Dans une telle situation nous avons retenu les données de ce site qui nous semble offrir le meilleur niveau scientifique.

Le deuxième exemple concerne l'anglais pour lequel nous observons une excellente concordance pour le nombre de locuteurs L1. Le nombre de locuteurs L2 varie de 167 à 612 millions. Nous avons retenu cette dernière valeur.

Enfin pour le tamoul un consensus se dégage pour le nombre de locuteurs L1 et L2, nous avons retenu 68 et 8 millions respectivement.

Langue	Code ISO	L1	L2	Sources
Amharique	[amh]	25 M	5 M	<a href="http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers">http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers</a>
		17 M		<a href="http://www.lmp.ucla.edu/Profile.aspx?LangID=7&amp;menu=004">http://www.lmp.ucla.edu/Profile.aspx?LangID=7&amp;menu=004</a>
		21 M	21 M	<a href="http://www.tfq.ulaval.ca/axl/afrique/ethiopie.htm">http://www.tfq.ulaval.ca/axl/afrique/ethiopie.htm</a>
		17 M	?	<a href="http://www.nalrc.indiana.edu/brochures/amharic.pdf">http://www.nalrc.indiana.edu/brochures/amharic.pdf</a>
		32 M?		<a href="http://www.plc.sas.upenn.edu/languages/amharic.html">http://www.plc.sas.upenn.edu/languages/amharic.html</a>
Anglais	[eng]	341 M	167 M	<a href="http://www.nationsonline.org/oneworld/most_spoken_languages.htm">http://www.nationsonline.org/oneworld/most_spoken_languages.htm</a>
		371 M	611 M	<a href="https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers">https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers</a>
		340 M	170 M	<a href="http://www.vistawide.com/languages/top_30_languages.htm">http://www.vistawide.com/languages/top_30_languages.htm</a>
		372 M	612 M	<a href="#">Ethnologue 20<sup>ème</sup> édition</a>
Tamoul	[tam]	68 M	8 M	<a href="http://www.ethnologue.org/show_language.asp?code=tam">http://www.ethnologue.org/show_language.asp?code=tam</a>
		67M	8 M	<a href="http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers">http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers</a>
		68 M	9 M	<a href="http://www.vistawide.com/languages/top_30_languages.htm">http://www.vistawide.com/languages/top_30_languages.htm</a>
		68 M	8 M	<a href="http://en.wikipedia.org/wiki/World_language">http://en.wikipedia.org/wiki/World_language</a>
		66 M	M's	<a href="http://www.lmp.ucla.edu/Profile.aspx?LangID=99&amp;menu=004">http://www.lmp.ucla.edu/Profile.aspx?LangID=99&amp;menu=004</a>
		52M	M's	<a href="http://www.plc.sas.upenn.edu/languages/tamil.html">http://www.plc.sas.upenn.edu/languages/tamil.html</a>

12

**TABLEAU 2. DIFFÉRENTES SOURCES DE LOCUTEURS L1 ET L2**

Même si elle n'est ni totalement exacte ni totalement complète cette approche a le mérite à nos yeux d'introduire dans le baromètre un facteur fondamental d'évaluation du poids des langues. Elle ajoute un poids aux langues internationales dominantes hyper-centrale (anglais) super-centrales (français, espagnol, chinois, etc.), aux autres langues de communication internationale (swahili, hausa) ainsi qu'aux langues nationales (kituba, lingala) et parfois régionale (goudjarati, kannada)

centrales ou périphériques. Elle souligne en même temps les graves lacunes dans la connaissance des situations sociolinguistiques et, sur ce point, nous ne pouvons que souhaiter que des études précises se multiplient.

### 2.A.4 Le statut de la langue

Ce facteur rend compte du degré de reconnaissance des langues par les instances politiques des pays dans lesquels elles sont parlées. Comme nous allons le voir ci-dessous il va bien au-delà de la simple notion de langue officielle du pays. Notre principale source d'information est ici le site "L'aménagement linguistique dans le monde" de l'université Laval<sup>18</sup>.

Commençons par nous intéresser aux définitions qui y sont données :

*"Le statut de «langue officielle» étant un concept plus ou moins ambigu, il faut comprendre que, dans ce site, une langue officielle est reconnue par la loi (de jure) ou dans les faits (de facto) par un État (souverain ou non souverain), sur l'ensemble du territoire ou une partie de celui-ci. Dans tous les cas, il faut que cet État dispose d'une assemblée, d'un Exécutif et d'une fonction publique, ce qui exclut les langues officielles d'un territoire autochtone («réserve»), d'une région administrative, d'une commune ou d'une municipalité. Un État peut reconnaître deux, trois ou quatre langues officielles sur son territoire. On parle alors d'État bilingue, trilingue ou quadrilingue."*

Un bon exemple de la distinction entre les statuts *de facto* et *de jure* se trouve aux Etats Unis d'Amérique où la constitution ne reconnaît aucune langue officielle, mais où le statut officiel de fait de l'anglais est incontestable.

Une ambiguïté est parfois introduite par ce que nous désignerons comme des « langues co-officielles de convenance » ou de *second rang*. En Algérie par exemple le tamazight a un statut de langue officielle qui est qualifié de « restrictif » par le site du CEFAN<sup>19</sup>, le gouvernement algérien n'a aucunement l'intention d'utiliser cette langue dans la communication avec les citoyens, l'Algérie n'est pas un état bilingue. Faut-il considérer que l'arabe classique et le tamazight ont le même statut ? En ce qui concerne Djibouti Jacques Leclerc conclue que le français y est « plus officiel » que l'arabe<sup>20</sup>. Nous trouverons une situation analogue dans les pays où l'arabe classique, parlé par pratiquement personne, n'est officiel que pour des raisons religieuses. Un autre cas est celui des pays ex-coloniaux ayant gardé pour langue officielle la langue du colonisateur mais ayant pour voisins, avec lesquels leurs relations sont importantes, des pays ayant conservé une autre langue coloniale. En Guinée-Bissau, voisine du Sénégal et de la Guinée-Conakry et membre de la Francophonie, le français a une place si importante que Jacques Leclerc le considère comme une langue co-officielle<sup>21</sup>.

---

<sup>18</sup> <http://www.tlfq.ulaval.ca/axl/>

<sup>19</sup> <http://www.axl.cefane.ulaval.ca/afrique/algerie-1demo.htm>

<sup>20</sup> <http://www.axl.cefane.ulaval.ca/afrique/djibouti.htm>

<sup>21</sup> <http://www.axl.cefane.ulaval.ca/afrique/Guinee-Bissau.htm>

## Le poids des langues dans le monde

---

Situation analogue en Guinée Équatoriale, entourée de pays francophones, où la pression du français, langue co-officielle, pour supplanter l'espagnol, première langue officielle, est forte.

Si le concept d'état souverain, équivalent à celui de pays, est clair, celui d'état non souverain doit être précisé. Citons ici la manière dont l'équipe de Laval le définit :

*"Qu'on les appelle État (Inde ou États-Unis), province (Canada), région autonome (Italie), communauté autonome (Espagne), collectivité territoriale ou territoire d'outremer (France), canton (Suisse), gouvernement participant (Francophonie), État libre associé (Åland, Porto Rico, Guam), etc., les États non souverains disposent, à des degrés variables, de compétences législatives, exécutives et judiciaires (souvent). Ils ont généralement leur propre constitution, leur parlement et leur législation, leur administration, leurs moyens financiers, etc. Ils jouissent de tous les attributs d'un État, sans la souveraineté politique. Ils sont cependant hiérarchiquement subordonnés à un autre gouvernement — le gouvernement central —, bien que, dans certains cas, les champs de juridiction soient exclusifs et s'exercent de façon autonome, voire souveraine."*

Selon cette définition l'Espagne comporte dix-sept communautés autonomes et deux cités autonomes (Ceuta et Melilla), Hong Kong et Macao sont des régions administratives spéciales de la République Populaire de Chine, l'île de Pâques est un territoire spécial du département de Valparaiso, les Îles Mariannes sont un état librement associé aux États Unis d'Amérique etc..

Nous adopterons donc cette définition mais distinguerons cependant deux cas :

- Une langue officielle ou co-officielle d'un état non souverain est la même que celle de l'état souverain dont il dépend. Les îles des Antilles, de l'océan Indien ou du Pacifique dépendant de la France, du Royaume Uni, des États-Unis, de la Nouvelle Zélande, de l'Australie ou des Pays-Bas ont le français, l'anglais ou le néerlandais comme langue officielle ou co-officielle. Le Groenland, Gibraltar sont dans une situation similaire. Ce statut correspond le plus souvent non pas à l'importance de la langue dans le pays mais à une commodité administrative : on ne parle pas beaucoup anglais aux Samoa américaines mais plutôt samoan, de même le gilbertin est parlé à Kiribati, le chamorro aux îles Mariannes, le Marshallais aux îles Marshall, le créole à la Guadeloupe et à la Martinique, le papiamentu aux Antilles néerlandaises. Nous ne comptabiliserons pas ces situations et considérerons aussi de la même manière certains cas ambigus de peu d'importance, par exemple : Sainte Hélène, les Falklands ou Saint Pierre et Miquelon, bien qu'aucune langue "indigène" n'y soit opposable à la langue officielle.
- Une langue officielle ou co-officielle d'un état non souverain est différente de la langue de l'état souverain dont il dépend. Le portugais à Macao, l'inuktitut au Groenland, l'anglais à Hong-Kong, le danois au Schleswig-Holstein, le galicien dans la communauté autonome de Galice sont dans cette situation. Ces cas sont comptabilisés comme langue officielle.

D'autre part nous considérerons ce niveau de statut officiel dans des états non souverains comme inférieur au précédent. Précisons en prenant l'exemple du Groenland, état non souverain, bénéficiant d'un certain niveau d'autonomie politique par rapport au Danemark. Le Groenland a deux langues officielles, le danois et le groenlandais (inuktitut du Groenland). Nous avons expliqué plus haut que le danois, langue officielle de l'état souverain dont dépend le Groenland n'est pas comptabilisé. Considérer l'Inuktitut comme langue officielle de statut équivalent à celle du danois au Danemark nous semble anormal, cette langue est d'usage purement interne et par exemple, il n'existe probablement aucune instance internationale prévoyant des traductions simultanées à partir

## Le poids des langues dans le monde

---

ou vers l'inuktitut. Nous attribuerons alors aux langues des états non souverains une "valeur" inférieure à celle des états souverains. Ce point sera précisé plus bas.

Il existe une troisième manière de distinguer une ou plusieurs langues parmi toutes celles parlées dans un pays. Les constitutions et les lois linguistiques accordent souvent un statut particulier à telle ou telle langue : langues admises dans les débats parlementaires, dans l'administration, dans les cours de justice ou dans les divers niveaux d'enseignement. Ces lois peuvent correspondre à un état de fait, une réelle volonté de promouvoir certaines langues, un choix politique voire populiste, un refus de choisir ou toute autre raison. Nous qualifierons une langue relevant d'un tel statut de "langue privilégiée". Nous recherchons ici une réelle volonté de promouvoir une langue et essayons d'éviter les déclarations de principe non suivies d'actions. Ceci mérite quelques précisions :

Les cas dans lesquels quelques langues sont déclarées officielles ou nationales, le Sénégal par exemple (six langues) seront considérés dans cette catégorie. Mais il existe des cas extrêmes comme la Bolivie où l'article 5 de la constitution de 2009 cite nommément trois douzaines de langues, le Venezuela (2008) qui en cite une quarantaine, ou le Pérou pour lequel l'article 48 de la constitution de 1993 indique dans son troisième alinéa qu'outre le castillan, le quechua et l'amayra, "les autres langues" sont officielles.

Bolivie :

### **Artículo 5**

*I. Son idiomas oficiales del Estado el castellano y todos los idiomas de las naciones y pueblos indígena originario campesinos, que son el aymara, arawona, 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 y zamuco*

Pérou :

### **Artículo 48 [1993]**

*Son idiomas oficiales el castellano y, en las zonas donde predominen, también lo son el quechua, el aimara y las demás lenguas aborígenes, según la ley.*

Ces cas extrêmes seront ignorés car ils ne correspondent pas à une réelle politique volontariste de promotion de telle ou telle langue. Par exemple devant une cour de justice bolivienne, les documents présentés doivent être rédigés en espagnol et non dans l'une des trente-six autres langues prétendues officielles, ce qui semble invalider l'article 5 cité plus haut. Précisons cependant que l'article du code de procédure civile date de 1975 et la constitution de 2009.

Parfois le traitement particulier accordé aux langues se réduit à une ville. Citons encore l'université Laval au sujet du Canada : *"La législation ne s'applique pas aux municipalités, mais certaines municipalités offrent des services en d'autres langues sur une base ponctuelle. La ville de Fort-Smith est la seule à avoir officiellement déclaré des services multilingues en anglais, en français, en chipewyan, en cri et en slavey du Nord."*

Et le gouvernement des territoires du Nord-Ouest (du Canada), affirme sur son site Internet offrir des services dans onze langues : anglais, français, cri, dogrib, chipewyan, slavey du sud, slavey du Nord,

# Le poids des langues dans le monde

gwich'in, inuvialuktun, inuktitut et inuinnaqtun<sup>22</sup>. Sous l'index "Official Languages" ce site Internet propose l'annonce ci-dessous :



16

Citons un autre exemple de langues officielles au niveau local, il nous vient de la ville de São Gabriel da Cachoeira, dans l'état d'Amazonas au Brésil :

## Article 1<sup>er</sup>

*Le portugais est la langue officielle de la République fédérale du Brésil*

## Paragraphe unique

*Il est établi que la municipalité de Saint-Gabriel de Cachoeira / État d'Amazonas adopte trois langues co-officielles, le nheengatu, le tukano et le baniwa.*

## Article 2

*Le statut de langue co-officielle accordé par le présent article oblige la Municipalité :*

**1°** À assurer des services publics de base de participation au public dans les répartitions publiques dans la langue officielle et dans les trois langues co-officielles, à l'oral et à l'écrit.

**2°** À produire la documentation publique, ainsi que les campagnes publicitaires institutionnelles dans la langue officielle et dans les trois langues co-officielles.

**3°** À encourager à soutenir l'apprentissage et l'usage des langues co-officielles dans les écoles et dans les moyens de communication.

<sup>22</sup> <http://www.gov.nt.ca/>

## Le poids des langues dans le monde

---

On notera que par l'article 2 cette municipalité (34000 habitants en 2005) s'impose un certain nombre d'obligations, et qu'il ne s'agit pas ici d'une généreuse déclaration d'intention uniquement formelle mais bien de la reconnaissance de ces trois langues.

Il est difficile d'en arriver à un tel niveau de détail et de recenser toutes les situations particulières dans le monde. Nous ne prétendons donc pas que notre compilation soit exhaustive : l'origine des données se trouve dans les constitutions, les lois linguistiques, les décrets et règlements édictés aux différents niveaux des administrations, et il est difficile de tous les consulter. Cependant cette compilation permet de distinguer des langues qui ont, même à un niveau local, obtenu une certaine reconnaissance de leur importance, de leur "poids".

Une autre source de confusion est l'existence de langues voisines. Ainsi au Mali les autorités ont reconnu 13 **langues nationales**. L'article 1 du décret 159 PG-RM du 19 juillet 1982 cite les langues suivantes<sup>23</sup>:

*le bambara (ou bamanankan), le bobo (bomu), le bozo, le dogon (dogo-so), le peul (fulfulde), le soninké (soninke), le songoy (songaï), le sénoufo-minianka (syenara-mamara et le tamasheq (tamalayt). Mais d'autres langues sont également reconnue s: le hasanya (arabe), le kasonkan, le madenkan et le maninkakan. Le français, quant à lui, bénéficie du statut de langue officielle, mais le bambara sert, dans plusieurs régions, de principale langue véhiculaire. Il n'est pas rare que, dans les villages du Sud, les enfants soient bilingues (langue locale + bambara), voire trilingues. À l'école, le français est souvent enseigné en tant que quatrième langue.*

Le problème est ici que selon les données d'Ethnologue, il existe au Mali quatre variétés de bozo, dont trois sont d'importance similaire du point de vue du nombre de locuteurs, quatorze variétés de dogon mais pas de dogon "dogo-so", deux maninkakan et trois songaï, quant au madenkan, nous ne le trouvons nulle part.

La conséquence de tout ceci est que le statut *effectif* des langues dans de nombreux pays est très flou, la conformité entre les textes lorsqu'ils existent et la réalité du terrain étant souvent illusoire. Il nous faudra donc, comme dans le cas du nombre de locuteurs, se contenter d'un certain degré d'approximation dans nos données.

Un quatrième niveau regroupera les langues qui sont admises comme plus importantes *de facto* sans qu'un texte écrit, constitutionnel législatif, réglementaire ou autre ne consacre cette prééminence. L'existence d'écoles où une langue est admise comme langue et non comme sujet d'enseignement dans les niveaux élémentaires, l'existence de journaux, stations de radio ou de télévision interne au pays considéré émettant dans cette langue etc.

Du point de vue du "poids", nous avons déjà indiqué que nous n'attribuons pas la même valeur aux différents niveaux de langues "officielles, nationales, constitutionnelles, admises, privilégiées".

Nous appliquerons donc les règles suivantes :

- a) Nous recenserons donc dans chaque entité, souveraine ou non, les langues correspondant aux quatre niveaux décrits et les combinerons ensuite en affectant un coefficient 1 aux

---

<sup>23</sup> <http://www.axl.cefan.ulaval.ca/afrique/mali.htm>

langues officielles des états souverains, 0.75 aux langues officielles « de convenance », 0.5 à celles des états non souverains ainsi qu'aux langues nationales *de facto* et 0.25 aux autres langues distinguées pour quelque raison que ce soit. Il faut ici noter que la population de l'état n'est pas prise en compte, l'incrément au score apporté par la Dominique (75000 habitants) à l'anglais est égal à celui apporté par la Chine (1,4 milliard d'habitants, soit 20000 fois plus peuplée) au mandarin.

b) En ce qui concerne les langues des états souverains nous retiendrons comme donnée brute le nombre d'états souverains dans lesquels la langue est officielle. Pour le malais par exemple nous en obtenons 4, Brunei, la Malaisie, l'Indonésie et Singapour.

c) Une langue ne pourra être retenue deux fois dans un même état souverain.

c.1) Si dans un même état souverain une langue est citée à deux ou trois niveaux nous ne la retiendrons qu'une fois, au niveau supérieur. Ce cas est très courant dans les états fédéraux. Le hindi officiel en Inde et dans une dizaine d'états de l'union est dans ce cas. Il ne sera retenu qu'une seule fois comme langue officielle de l'Union Indienne. De même pour des états non souverains, par exemple Curaçao étant formellement sous souveraineté néerlandaise, le néerlandais n'y sera pas retenu comme langue officielle.

c.2) Si une langue est officielle dans plusieurs états non souverains d'un même état souverain nous ne la retiendrons qu'une seule fois. Le Xhosa, officiel dans quatre états de l'Afrique du Sud, le Zoulou dans trois en sont des exemples, ils ne seront comptabilisés qu'une seule fois chacun, comme langue officielle d'états non souverains.

d) Si une langue est officielle dans un état souverain et une partie d'un autre état souverain elle est retenue pour les deux statuts. Le swati, officiel au Swaziland et dans la province du Mpumalanga en Afrique du Sud en est un exemple, il sera comptabilisé deux fois, langue officielle nationale au Swaziland et provinciale en Afrique du Sud.

### 2.A.5 Le nombre de traductions à partir de la langue

Nous utilisons ici les données de l'Index Translationum<sup>24</sup> que l'on trouve sur le site de l'UNESCO. L'index publie le nombre de traductions effectuées par langue depuis 1979. Les données peuvent être analysées par pays dans lequel la traduction a eu lieu, par année durant laquelle elle a eu lieu et par sujet. Les traductions sont classées en neuf catégories : généralités et bibliographie ; philosophie et psychologie ; religion et théologie ; droit, sciences sociales et éducation ; sciences exactes et naturelles ; sciences appliquées ; arts, jeux et sports ; littérature et enfin histoire, géographie et biographie. Le site n'utilise pas systématiquement les codes ISO 693-3 mais avec le nom de la langue indique en code qui s'en rapproche assez pour que l'utilisateur n'ait le plus souvent pas de problème dans l'identification de la langue concernée. Nous rencontrons cependant parfois des cas ambigus, comme l'attribution d'une traduction à une variété dialectale non décrite par Ethnologue. Par exemple

---

<sup>24</sup> <http://www.unesco.org/xtrans/bsstatlist.aspx>

## Le poids des langues dans le monde

L'index Translationium indique comme source de traduction "dialectes de l'Ijo du Sud-Est", dont le code pour l'index (IJS-DI) n'existe pas dans la norme ISO 693-3. Le code ISO [ijs] correspondant effectivement à l'ijo du Sud-Est nous affectons les traductions des dialectes à l'ijo du sud-est. Parfois Translationium indique une langue alors qu'il en existe deux variétés ou plus reconnues par Ethnologue. C'est le cas de cas de l'albanais, de l'azéri, du panjabi. Nous nous faisons alors le choix en fonction des critères suivants : le pays dans lequel les traductions sont faites donnent une indication, le rapport du nombre de locuteurs entre les variétés est important, une des variétés est langue officielle dans un pays et non les autres. Ainsi les données relatives à l'albanais sont affectées à l'albanais tosk [als], celles relatives à l'azéri à la variété septentrionale [azl], et celles du panjabi au panjabi occidental [pnb]. Il existe d'autres exemples de cette situation.

Un autre problème est posé par les langues qui pour des raisons politiques n'existent plus. Si l'hindi et l'ourdou ont divergé avant le début de la compilation translationium, il n'en est pas de même pour le serbocroate aujourd'hui éclaté entre serbe [srp], croate [hrv] et bosniaque [bos]. Nous avons attribué les données du serbocroate, qui n'existe donc plus, aux trois autres langues au prorata de leur nombre de traductions respectives compilées depuis la "création" des langues issues du serbocroate. Fort heureusement il n'existe pas encore de monténégrin.

L'arabe constitue un autre problème. L'index rapportait, au 22 octobre 2017, 12410 traductions à partir de l'arabe (ARA). En outre 3 traductions étaient reportées à partir de l'arabe marocain (ARY) et 5 à partir de l'arabe tchadien (SHU) et 1 à partir des dialectes de l'arabe (ARA-DI) sans autre précision. Dans la nomenclature ISO, le code [ara] est celui d'un *macrolangage* regroupant toutes les variétés. D'autre part [arb], [ary] et [shu] sont bien les codes de l'arabe standard, l'arabe marocain et l'arabe tchadien (arabe shuwa), et il n'existe évidemment pas de code pour les dialectes de l'arabe. Le tableau 3 regroupe ces données. Nous rencontrons un problème analogue avec la compilation d'article dans Wikipedia, plus de 540000 articles en arabe standard, 17000 en arabe égyptien et aucun dans les autres arabes dialectaux. Tout ceci montre que les différents *arabes parlés* ne sont pas des *langues écrites*.

Le problème réside probablement dans la collecte des données. Le site fait état d'un certain nombre de partenariats, bibliothèques nationales, instituts, universités et experts mais il semble que la collecte des données se fasse sur une base déclarative ce qui autoriserait retards, déclarations approximatives, incomplètes ou absentes. Nous avons respectivement affecté ce que Translationium décrit comme ARA, ARY et SHU aux langues codées [arb], [ary] et [shu].

	Arabe			Dialectes
	Standard	Marocain	Tchadien	
Code Translationium	ARA	ARY	SHU	ARA-DI
Code ISO 693-3	[arb]	[ary]	[shu]	?
Sans indication.	2	0	0	
Arts, Jeux, Sport	89	0	0	
Sciences exactes et naturelles	68	0	0	
Droit, Sciences sociales, Éducation	1072	0	0	1
Généralités, Bibliographie ...	28	0	0	

## Le poids des langues dans le monde

Histoire, Géographie, Biographie	693	0	0	
Littérature	4958	3	5	
Philosophie, Psychologie	319	0	0	
Religion, Théologie	4985	0	0	
Sciences appliquées	196	0	0	

TABLEAU 3. TRADUCTIONS À PARTIR DES ARABES STANDARD ET DIALECTAUX

### 2.A.6 Le nombre de traductions vers la langue

Nous utilisons ici les données de l'Index Translationum. On se reportera au paragraphe précédent pour plus de détails.

### 2.A.7 Les prix littéraires internationaux

Ce facteur a pour but de prendre en compte la reconnaissance de la culture véhiculée par une langue au travers des prix littéraires internationaux obtenus par les écrivains l'ayant utilisée.

Le premier prix venant à l'esprit est naturellement le plus prestigieux d'entre eux, le prix Nobel de littérature<sup>25</sup>. Cependant il est possible d'argumenter qu'il présente plusieurs biais. Le premier d'entre eux consiste à remarquer que la plupart des prix ont été attribués à des auteurs s'exprimant dans une langue originaire d'Europe occidentale. Environ 60% des prix ont été attribués à l'anglais, au français, à l'allemand ou à l'espagnol, le comité Nobel est "eurocentrique". Dans le même ordre d'idée il faut remarquer que la Suède à elle seule a récolté autant de prix que l'ensemble de l'Asie, huit prix suédois contre deux japonais et un seul chinois, bengali, turc, hébreu ou arabe. Mais il nous faut nuancer ce jugement. Ainsi le prix de Tagore permet de distinguer le bengali des autres langues importantes du sous-continent indien. De même des langues comme l'arabe, le chinois mandarin, le finlandais, l'hébreu, le hongrois, l'islandais, le serbe, le tchèque, le turc et le yiddish reçoivent une reconnaissance de la culture qu'elles véhiculent. Il faut aussi remarquer que l'espagnol et le portugais sont aujourd'hui plus des langues sud-américaines qu'occidentales et dans le cas de l'espagnol, l'apport de la culture sud-américaine est reconnu avec les prix attribués à Miguel Angel Asturias, Pablo Neruda, Gabriel García Márques, Octavio Paz et Mario Vargas Llosa.

Le deuxième type de controverse est de nature politique. L'académie suédoise est considérée comme pensant "à gauche", ce qui expliquerait que Jorge Luis Borges n'a pas été distingué à cause de son soutien aux dictatures argentine et chilienne. Jean Paul Sartre et Pablo Neruda, qui ne condamnaient pas les dictatures de gauche ont été distingués. L'académie a également été soupçonnée de favoriser l'Allemagne et de défavoriser la Russie. Tolstoï a été proposé seize fois, Merzkovsky huit fois, Berdyayev sept fois, ils n'ont jamais été récompensés et Ivan Bunin a été proposé dix-huit fois pour être finalement distingué en 1933.

La liste des auteurs reconnus comme majeurs qui n'ont jamais été distingués est longue : Marcel Proust, Ezra Pound, James Joyce, Vladimir Nabokov, Virginia Woolf, Jorge Luis Borges, Gertrude

<sup>25</sup> <http://www.nobelprize.org/>

## Le poids des langues dans le monde

---

Stein, August Strindberg, John Updike, Arthur Miller, Yannis Ritsos et bien d'autres. De nombreux auteurs "inconnus" des non spécialistes ont cependant été distingués : parmi les prix récents nous pouvons citer de Hertha Müller et Tomas Tranströmer.

On l'aura compris, le simple prix Nobel de littérature n'est pas suffisant pour atteindre le but défini au début de ce paragraphe. C'est la raison pour laquelle nous avons choisi de prendre en compte d'autres prix littéraires internationaux et avons choisis le prix Neustadt<sup>26</sup>, le prix Man Booker<sup>27</sup>, le prix Franz Kafka<sup>28</sup>, le prix Ovid<sup>29</sup>, le prix de Jerusalem<sup>30</sup>, l'American Award in literature<sup>31</sup> et le prix « Golden Wreath ». Nous avons également retenu le prix Prince des Asturies<sup>32</sup> mais uniquement à partir de 1999. Avant cette date les lauréats étaient tous de langue espagnole. Nous avons également retenu le prix Park Kyung-Ni décerné en Corée du Sud qui existe depuis 2011 et qui a une vocation internationale.

De nombreux autres prix littéraires existent mais n'ont pas pour vocation d'examiner des candidats originaires du monde entier<sup>33</sup>. Certains sont dédiés à la littérature asiatique, d'autres à la littérature arabe voire à une seule langue comme le prix nigérian Karaye consacré aux œuvres écrites en haoussa. Le prix Jnanpith<sup>34</sup> décerné en Inde a distingué des auteurs s'exprimant dans une des vingt-deux langues constitutionnelles (« scheduled languages ») du sous-continent indien. Nous avons bien entendu recherché des prix internationaux décernés dans d'autres pays que ceux appartenant au « monde occidental », il en existe peu, la plupart des prix récompensent des auteurs écrivant dans la langue du pays. Ces prix distinguent des individus plutôt que des langues, par construction leur domaine est limité et nous ne les retenons donc pas.

Les règles que nous appliquons sont les suivantes :

- a. Pour chacun de ces prix nous attribuons un point à la langue dans lequel le lauréat s'exprime.
- b. Si un prix est partagé, les deux langues se voient attribuer un point ou si les deux auteurs s'expriment dans la même langue celle-ci se voit attribuer deux points.
- c. Si un auteur a écrit en deux langues différentes et qu'il est récompensé pour l'ensemble de son œuvre les deux langues se voient attribuer un point. C'est le cas de Milan Kundera.

---

<sup>26</sup> <http://www.ou.edu/wlt/neustadt-prize.html>

<sup>27</sup> <http://www.themanbookerprize.com/prize/man-booker-international>

<sup>28</sup> <http://www.franzkafka-soc.cz/>

<sup>29</sup> [http://en.wikipedia.org/wiki/Ovid\\_Prize](http://en.wikipedia.org/wiki/Ovid_Prize)

<sup>30</sup> [http://www.jerusalembookfair.com/the\\_jerusalem\\_prize.html](http://www.jerusalembookfair.com/the_jerusalem_prize.html)

<sup>31</sup> [https://en.wikipedia.org/wiki/American\\_Award\\_in\\_Literature](https://en.wikipedia.org/wiki/American_Award_in_Literature)

<sup>32</sup> <http://www.fpa.es/premios/>

<sup>33</sup> [http://en.wikipedia.org/wiki/Man\\_Asian\\_Literary\\_Prize](http://en.wikipedia.org/wiki/Man_Asian_Literary_Prize)

<sup>34</sup> <http://jnanpith.net/index.html>

d. Si le même auteur reçoit plusieurs prix sa langue d'expression se voit attribuer autant de point que l'auteur a reçu de prix. C'est par exemple le cas de Amos Oz ou de Ismail Kadaré.

Même si pour les raisons discutées plus haut ce facteur ne reflète qu'imparfaitement ce que nous voulons quantifier, la reconnaissance internationale du niveau de culture d'une langue il introduit dans le baromètre un facteur important pour l'évaluation du poids des langues.

### 2.A.8 L'activité dans Wikipedia

Nous utilisons ici les données que l'on trouve sur le site des statistiques de Wikipedia<sup>35</sup>. Le nombre que nous retenons est la somme de tous les articles publiés dans Wikipédia depuis l'origine de l'encyclopédie jusqu'à la mise à jour la plus récente au moment où nous rassemblons les données. Signalons ici que Wikipedia n'utilise pas le code ISO 639-3 pour identifier sans ambiguïté les langues, ce qui pourrait théoriquement poser certaines difficultés mais n'a pas en l'occurrence constitué de problème majeur. Les ambiguïtés sont résolues de manière similaire à celle décrite au paragraphe 2.A.5.

### 2.A.9.L'enseignement au niveau universités

L'idée de ce facteur, introduit dans la troisième édition (2017) du baromètre des langues est de quantifier l'importance d'une langue par son enseignement au niveau universitaire. Il s'agit d'examiner les sites internet d'un échantillon des universités dans tous les pays du monde pour en extraire l'information sur les langues enseignées aux premiers niveaux de l'enseignement supérieur, le niveau doctorat (ou "post-graduate") étant exclu. Sont également pris en compte :

-Les organismes universitaires ou para-universitaires consacrés à l'enseignement de "langues rares" ou plutôt rarement enseignées. Il s'agit là de l'INALCO (Institut National des Langues et Civilisations Orientales, en France), du SOAS (School of Oriental and African Studies, au Royaume-Uni), du NARLC (National African Languages Resource Center, aux États-Unis), du CIIL (Central Institute of Indian Languages, en Inde), etc.,

-Des "foreign/modern languages centers" qui permettent aux étudiants de quelque niveau que ce soit de se familiariser avec une langue sans que celle-ci fasse formellement partie de leur cursus universitaire.

Il existe environ 20.000 établissements universitaires dans le monde, il n'est pas question de les examiner de manière exhaustive. Les règles que nous appliquons sont les suivantes :

#### 1. Nombre d'universités considérées

La première règle concerne le nombre d'universités choisies dans chaque pays :

---

<sup>35</sup> [http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias#Grand\\_Total](http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total)

# Le poids des langues dans le monde

---

1.a Nous sélectionnons au moins 10% du nombre total des universités dans le pays considéré. C'est-à-dire qu'au minimum une université sera sélectionnée sur un nombre total compris entre 1 et 10, qu'au minimum deux universités seront sélectionnées pour un nombre total compris entre 11 et 20, etc. Nous insistons sur le fait qu'il s'agit là d'un minimum. Ainsi en Inde nous avons considéré 32 universités sur un total possible de 150<sup>36</sup>.

1.b Pour les pays dans lesquels le nombre d'universités est supérieur à 100 nous examinons au minimum 10 universités et ne respectons pas toujours la règle 1.a des 10%. Par exemple il est possible de visiter les sites internet de près de 2000 universités aux États-Unis, nous en avons considéré 49. De même en Chine nous avons considéré 25 universités sur un total possible de 354. Nous nous expliquerons plus bas sur ce point.

## 2. Comment choisir les universités ?

Il s'agit ici de sélectionner les universités les plus représentatives dans le domaine de l'enseignement des langues étrangères. Plusieurs compilations sont disponibles, nous les exploitons.

2.a Les universités figurant dans le classement des meilleures universités pour l'enseignement des langues modernes<sup>37</sup>. Deux cents universités y figurent, nombre qui ne nous paraît pas suffisant.

2.b Les universités figurant dans un classement des meilleures universités au monde classées par pays<sup>38</sup>.

2.c Les universités figurant au classement des meilleures universités asiatiques dans le domaine "Arts and humanities"<sup>39</sup> Cent universités sont reprises dans ce classement.

2.d Un site listant "toutes" les universités dans "tous" les pays du monde<sup>40</sup>.

2.e Les sites internet d'établissements ou d'organismes consacrés à l'enseignement des langues sont également pris en compte. Les exemples les plus représentatifs sont l'INALCO<sup>41</sup>, le SOAS<sup>42</sup>, (School of Oriental and African Studies), le CIIL en Inde et le NALRC<sup>43</sup> (National African Languages Resource Center). Il peut y en avoir d'autres.

2.f Tout autre site universitaire accessible par d'autre moyens.

---

<sup>36</sup> <http://www.bulter.nl/universities/>

<sup>37</sup> <http://www.topuniversities.com/university-rankings/world-university-rankings/2011/subject-rankings/arts-humanities/modern-languages>

<sup>38</sup> <http://www.webometrics.info/>

<sup>39</sup> <http://www.topuniversities.com/university-rankings-articles/asian-university-rankings/top-universities-asia-arts-humanities-2014>

<sup>40</sup> <http://univ.cc/>

<sup>41</sup> <http://www.inalco.fr/>

<sup>42</sup> <http://www.soas.ac.uk/academic/>

<sup>43</sup> <http://www.nalrc.wisc.edu/>

## Le poids des langues dans le monde

---

Les règles définies aux paragraphes 1 et 2 ci-dessus sont appliquées avec souplesse et nous examinons autant d'universités qu'il est nécessaire pour arriver à une conviction raisonnable que l'information recueillie dans le pays considéré est représentative de la réalité. Le fait que nous examinons les universités par ordre décroissant de « qualité » si le site consulté propose un tel classement aide à obtenir cette conviction. L'apparition répétitive des mêmes langues dans les diverses universités considérées dans un même pays est également un élément convaincant.

Au final, nous essayons d'arriver à une situation dans laquelle les meilleures universités enseignant des langues vivantes dans un pays donné sont prises en compte et que tout ajout d'une ou plusieurs autres universités dans ce pays créerait une redondance.

### *3. Quelles sont les possibles sources d'approximations ?*

La méthode consistant à sélectionner un échantillon des universités et non pas la totalité introduit évidemment une source d'approximations dans l'étude. Plusieurs causes en sont possibles.

3.a. Le site de l'université est inexistant ou inaccessible. Ce cas se rencontre par exemple en Mongolie où nous n'avons pas été capables de trouver le site<sup>44</sup> "School of Foreign Languages and Cultures" de l'Université Nationale de Mongolie<sup>45</sup>. Il est difficile de savoir si l'inaccessibilité est permanente ou temporaire.

3.b. Le site est difficilement lisible car mal organisé. Par exemple l'"Universidad Nacional del Littoral" en Argentine donne la liste des départements, leur adresse internet mais ne décrit pas le contenu de l'enseignement dispensé<sup>46</sup>. Il faut alors explorer l'arborescence du site pour s'assurer que l'information n'est pas disponible par une autre voie. Une autre voie consisterait bien sûr à utiliser l'adresse mail indiquée pour demander l'information recherchée. Nous ne l'avons pas fait.

3.c. L'information est peu claire ou parcellaire. Par exemple, certaines universités africaines dans leurs départements de langues africaines n'indiquent pas clairement la liste des langues réellement enseignées de manière permanente et souvent n'indique rien de plus que "african languages".

3.d. Le site est dans une langue que nous ne comprenons pas. Ce problème se rencontre en Indonésie, et dans les pays très fermés comme la Corée du Nord et la Birmanie.

3.e. Seul le portail ou les premières pages du site de l'université sont accessibles dans une langue que nous comprenons et les pages sur lesquelles pourraient se trouver l'information recherchée ne sont pas traduites. La situation est analogue à celle du cas précédent, nous l'avons rencontrée par exemple en Hongrie.

3.f. Une ou plusieurs langues "rares" sont enseignées dans une université que nous n'avons pas considérée en appliquant les critères définis plus haut.

3.g. Le site de l'université est classé à risque ou malveillant par notre logiciel antivirus.

---

<sup>44</sup> <http://sflc.num.edu.mn/>

<sup>45</sup> <http://www.num.edu.mn/>

<sup>46</sup> <http://www.fhuc.unl.edu.ar/>

3.h. Les langues officielles au niveau national ou fédéral ne sont pas considérées. L'allemand en Autriche, l'anglais et l'afrikaans en Afrique du Sud, l'anglais et l'hindi dans l'Union Indienne, le malais le mandarin, le tamoul et l'anglais à Singapour en sont des exemples. Les langues ayant un statut hybride sont considérées même dans le pays où elles bénéficient de ce statut. Les langues constitutionnelles en Inde, les langues des provinces en Afrique du Sud sont donc prises en compte y compris dans ces pays. La conséquence en est que leur importance est surestimée.

3.i Enfin la nature même de la langue est sujet d'incertitude. En voici quelques exemples :

-Lorsqu'un site indique "arabe" sans plus de précision, nous décidons qu'il s'agit de l'arabe standard [arb].

-En revanche dans le cas du malais il existe dans Ethnologue une macrolangue [msa], un malais standard [zsm] sans locuteurs L1 et un malais parlé en Malaisie [zlm]. Nous avons choisi de retenir ce dernier. Ce choix nous semble refléter plus précisément l'importance du malais parlé en L1.

-Dans le cas du népalais nous avons fait le choix inverse de la macrolangue [nep] au dépend du népalais [npi] car les deux composantes de celle-là sont parlées au Népal et beaucoup moins en Inde et au Bhoutan

-Dans le cas de langues comme l'azéri ou le kurde pour lesquelles plusieurs variantes coexistent notre choix dépend du contexte. L'azéri du nord [azj] étant langue officielle en Azerbaïdjan est retenu au dépend de l'azéri du sud minoritaire en Iran, même si ce dernier a plus de locuteurs.

Lorsque de organismes comme le SOAS ou l'INALCO indique par exemple "berbère" nous retenons plusieurs variétés de berbère.

Les sources d'imprécision sont donc nombreuses, mais dans la plupart des cas elles concernent des langues "mineures" et ne changent donc que peu de choses dans l'analyse de la situation.

#### 4. Résultats

L'idée n'est pas de quantifier le nombre de fois où une langue est enseignée mais la proportion d'universités qui la proposent rapportée au nombre d'université qui *pourraient* la proposer. Pour être clair nous éliminons pour chaque langue les universités des pays dans lesquels la langue est officielle. L'anglais aux Etats-Unis, au Royaume Uni etc., le français en France, Belgique Côte d'Ivoire etc. Ce calcul est fait pour le monde entier

Le processus que nous venons de décrire nous a permis de compiler 327 différentes langues (caractérisées par leur code ISO 639-3) enseignées dans 1142 universités situées dans 198 pays. Le total des universités dans le monde étant de l'ordre de 20.000 nous estimons la marge d'incertitude à environ 3%. Rappelons que les universités retenues sont choisies, lorsque c'est possible en fonction de leurs réputations d'excellence dans le domaine de l'enseignement des langues étrangères.

### 2.A.10. Systèmes graphiques

Lorsqu'on compare les vingt premières langues de notre baromètre, tableau 4 ci-dessous, on remarque que dans nos trois précédentes éditions quatorze d'entre elles sont toujours présentes

## Le poids des langues dans le monde

(anglais, français, espagnol, allemand, néerlandais, russe, japonais, suédois, italien, mandarin, polonais, portugais, hongrois, danois) , à des places parfois différentes. Elles sont toujours présentes dans le classement 2022. Trois n'apparaissent que deux fois (catalan, finnois, norvégien) et que cinq n'apparaissent qu'une fois (arabe, hébreu, suisse allemand, grec et turc).

	2010	2012	2017	2022
1	Anglais	Anglais	Anglais	Anglais
2	Français	Espagnol	Français	Français
3	Espagnol	Français	Espagnol	Espagnol
4	Allemand	Allemand	Allemand	Allemand
5	Néerlandais	Russe **	Russe **	Russe **
6	Japonais **	Japonais **	Italien	Italien
7	Suédois	Néerlandais	Portugais	Suédois
8	Arabe **	Italien	Japonais **	Roumain
9	Italien	Portugais	Néerlandais	Portugais
10	Danois	Mandarin **	Suédois	Polonais
11	Finnois	Suédois	Mandarin **	Néerlandais
12	Russe **	Turc	Polonais	Catalan
13	Mandarin **	Norvégien	Tchèque	Tchèque
14	Hébreu **	Polonais	Croate	Croate
15	Polonais	Danois	Roumain	Mandarin **
16	Portugais	Finnois	Serbe	Hongrois
17	Hongrois	Hongrois	Hongrois	Indonésien
18	Suisse allemand	Roumain	Coréen **	Japonais **
19	Grec **	Catalan	Norvégien	Norvégien
20	Catalan	Tchèque	Danois	Finnois

Tableau 4 Les vingt premières langues dans les quatre éditions du baromètre

Certaines langues apparaissent donc : turc, norvégien, finnois, roumain et tchèque dans la version 2012, croate, serbe et coréen dans la version 2017. Ces modifications s'expliquent par le changement de certaines langues au regard de nos facteurs, par l'ajout de nouveaux facteurs (la véhicularité en 2012, l'enseignement des langues dans les universités en 2017) et, pour ce qui concerne l'arabe, le fait que nous prenions en compte en 2010 le seul arabe standard alors que nous avons ensuite introduit les différents arabes nationaux (arabe, égyptien, arabe algérien, etc.)

Mais il y avait dans ces trois classement une constante : la domination aux vingt premières places de langues utilisant l'alphabet latin. En 2010 six d'entre elles utilisaient un autre système de transcription graphique (japonais, arabe, russe, mandarin, hébreu, grec), trois en 2012 (russe, japonais, mandarin) et quatre en 2017 (russe, japonais, mandarin, coréen). Ces langues sont signalées par deux astérisques \*\*. Dans cette nouvelle version qui introduit le système graphique comme nouveau facteur le coréen a disparu de la tête du classement et si le russe demeure au cinquième rang le mandarin et surtout le japonais reculent sensiblement.

Cette particularité pose une question (concernant la corrélation entre la place d'une langue dans notre baromètre et son système d'écriture) et nous a menés à la réflexion suivante : le système graphique utilisé par une langue peut-il avoir une influence sur son expansion ? Est-il par exemple plus facile pour un anglophone d'apprendre à lire le français ou le turc que le chinois, pour un arabophone d'apprendre à lire l'ourdou ou le farsi que l'hindi, pour un russophone d'apprendre à lire le serbe ou le bulgare que l'ourdou ? Expliquons nous :

## Le poids des langues dans le monde

Le linguiste Nicolas Tournadre, après avoir rappelé à notre propos que nous avons « proposé de parler de « poids des langues » et, dans cette veine, on pourrait proposer la notion de « poids des écritures »<sup>47</sup>, a détaillé ce qui constituait à ses yeux cette difficulté :

*« La difficulté des systèmes d'écritures est relativement aisée à établir en se fondant sur le nombre de signes à mémoriser et sur la complexité interne de ces signes. Les écritures logographiques ont quelques milliers de signes, les syllabiques en comportent quelques centaines, tandis que les alphasyllabaires et les alphabets n'en ont généralement que quelques dizaines. Il est donc indiscutable que les systèmes logographiques sont plus difficiles à maîtriser que les syllabaires et que ces derniers sont plus complexes que les alphasyllabaires et les alphabets »<sup>48</sup>.*

Pour vérifier cette hypothèse que nous partageons avec lui, nous avons donc décidé d'ajouter un nouveau facteur dans cette quatrième version de notre baromètre.

Comme pour la version 2017 du baromètre, nous avons classé 634 langues. Parmi les 44 systèmes graphiques utilisés pour leur transcription, 23 ne le sont que pour une langue, 9 que pour 2 langues. Le tableau 5 ci-dessous indique les systèmes les plus utilisés parmi nos 634 langues :

Système graphique Alphabet	Occurrence	Système graphique Alphabet	Occurrence
Latin	379	Sinogrammes	29
Arabe	74	Gé'ez	13
Cyrillique	34	Bengali	10
Devanagari	30	Laotien	5

**Tableau 5. Compilation des systèmes graphiques**

La somme n'est pas égale à 634, nous avons limité ce tableau aux systèmes utilisés par au moins cinq langues. Certains systèmes sont spécifiques d'une langue et certaines langues ne sont tout simplement pas écrites.

On voit que l'alphabet latin domine largement, très loin devant les alphabets arabe ou cyrillique et les sinogrammes. Cette domination s'explique bien sûr par des raisons historiques. Par le fait d'abord pour les deux premiers (latin et arabe) que l'écriture alphabétique soit née au cours du premier millénaire avant JC dans le bassin méditerranéen et pour l'alphabet cyrillique qu'il ait été créé, au 9<sup>ème</sup> siècle après par des moines (Cyrille et Méthode) s'inspirant essentiellement de l'alphabet grec. Mais elle s'explique aussi par les expansions religieuses et impérialistes : les alphabets latin et arabe

<sup>47</sup> N. Tournadre, *Le prisme des langues*, Paris, L'Asiathèque, 2016, page 101,

<sup>48</sup> Op. cit. page 287

## Le poids des langues dans le monde

se sont diffusés à la fois avec les religions chrétienne ou musulmane et avec les impérialismes arabes ou européens.

Nous avons donc compilé pour les 634 langues que nous retenons depuis la version 2017 de notre baromètre le ou les systèmes graphiques utilisés. Une langue peut n'être pas écrite ou utiliser selon les pays dans lesquels elle est parlée utiliser un deux ou trois systèmes graphiques différents. Ceci peut être dû à des raisons politiques et/ou religieuses. Donnons l'exemple du turkmène, tableau 6.<sup>49</sup>

Systèmes graphiques utilisés par le turkmène	
Au Turkménistan	Cyrillique Latin officiel depuis 1991
En Iran et en Afghanistan	Alphabet arabe

**Tableau 6. Systèmes graphiques utilisés par le turkmène**

La compilation des systèmes utilisés nous a donné un fichier de 712 lignes le résumé étant donné dans le tableau 7.

Système graphique	Nombre d'occurrences	Système graphique	Nombre d'occurrences	Système graphique	Nombre d'occurrences
Latin	421	Khmer	3	Llao	1
Arabe	80	Nuosu bburma	3	Malayalam	1
Cyrillique	39	Géorgien	2	Meitei mayek	1
Devanagari	30	Gujarati	2	Mongol bitchig	1
Sinogrammes	30	Hébreu	2	Nüshu	1
Langue non écrite	17	Taï-le	2	Oriya	1
Ge'ez	12	Arménien	1	Pahawh	1
Bengali	10	Assamais	1	Singhalais	1
Batak	6	Balinais	1	Syllabaire cherokee	1
Laotien	5	Chakma	1	Syriaque	1
Thaï	5	Ge'ez	1	Tamoul	1
Tifinagh	5	Hangul	1	Télougou	1
Tibétain	4	Hmong	1	Thaana	1
Birman	3	Kana	1	Tigalari	1
Grec	3	Lati	1	Utkala lipi	1
Kannada	3	Lisu	1		

**Tableau 7. Compilation des systèmes graphiques et de leur utilisation**

<sup>49</sup> <https://www.axl.cefan.ulaval.ca/asia/turkmenistan-1General.htm>

## Le poids des langues dans le monde

---

Les scores de chacun des systèmes graphiques sont ensuite calculés comme suit :

Nous affectons la valeur 0 à l'absence de système graphique.

Le domaine des valeurs du nombre d'occurrences s'étendant sur plus de deux ordres de grandeur nous effectuons une transformation logarithmique et calculons le score d'un système en appliquant la formule :

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

Le symbole  $\text{Log1P}(\text{occurences})$  signifie que nous prenons le logarithme du nombre d'occurrences augmenté de 1. Ceci permet d'avoir la valeur finie zéro pour l'absence de système graphique (le logarithme de zéro n'est pas défini). Les scores varient de 0 (pas de système graphique) à 1 (alphabet latin).

Enfin pour chacune des langues nous prenons la moyenne des scores des systèmes graphiques utilisés. Les scores s'établissent encore entre 0 (langue non écrite) à 1 pour les langues n'utilisant que l'alphabet latin.

Le résultat de l'influence de ce nouveau facteur sur le classement des langues apparaît dans la figure 2 ci-dessous. Y sont représentées 634 langues, avec en ordonnée le rang 2017, en abscisse le rang 2021.2 Les langues en rouge au-dessous de la première diagonale sont moins bien classées en 2021 qu'elles ne l'étaient en 2017. Leur score système graphique est inférieur à 0,6. Les langues en vert sont celles qui ont un score système graphique égal à 1, c'est-à-dire qu'elles utilisent uniquement un alphabet latin. Les langues au-dessus de la première diagonale, mieux classées en 2021 qu'elles ne l'étaient en 2017 sont majoritairement vertes.

## Le poids des langues dans le monde

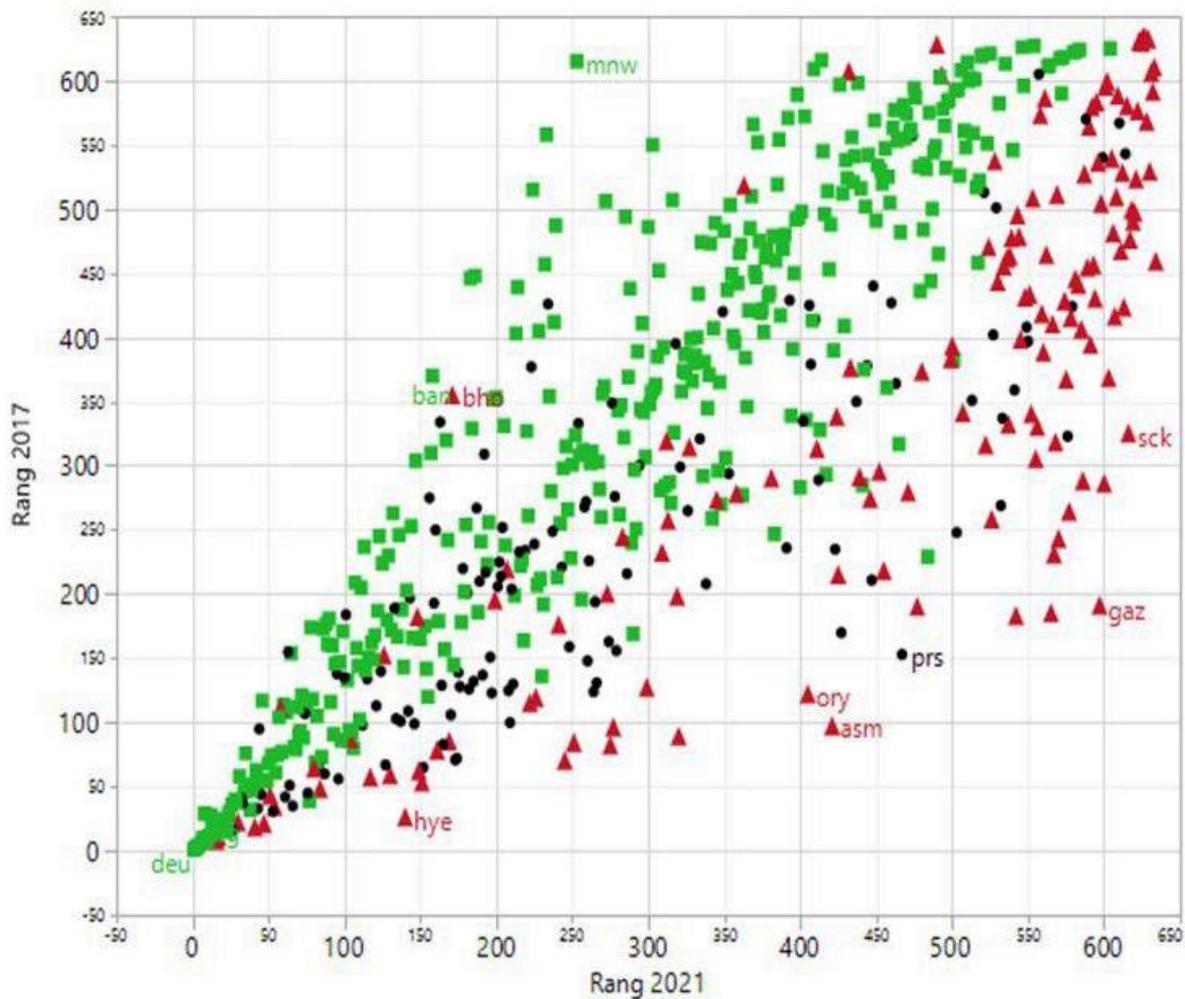


Figure 2. Comparaisons des classements 2017 et 2022

Si nous comparons maintenant le classement 2022 (avec facteur écriture) et le classement 2017 (sans écriture), toujours pour les vingt premières langues, le tableau 8 ci-dessous nous montre que les langues qui reculent ou disparaissent dans le classement (en gras) ont, à l'exception du portugais, un autre système graphique que l'alphabet latin, et que celles qui avancent ou apparaissent (en rouge) utilisent toutes l'alphabet latin.

	2017	2022
1	anglais	Anglais
2	Français	Français
3	Espagnol	Espagnol
4	Allemand	Allemand
5	Russe	Russe
6	Italien	Italien
7	<b>Portugais</b>	Néerlandais
8	<b>Japonais</b>	Suédois

## Le poids des langues dans le monde

9	Néerlandais	Roumain
10	Suédois	Portugais
11	Mandarin	Polonais
12	Polonais	Catalan
13	Tchèque	Tchèque
14	Croate	Croate
15	Roumain	Mandarin
16	Serbe	Indonésien
17	Hongrois	Hongrois
18	Coréen	Japonais
19	Norvégien	Norvégien
20	Danois	Finnois

**Tableau 8. Comparaison des classement 2017 et 2022**

Cette corrélation entre les changements dans le rang des langues et l'introduction du facteur graphique dans notre baromètre, apparaît donc comme une relation de causalité

## 2.B Les facteurs contextuels

Ce sont les facteurs qui ne sont pas spécifiques d'une langue particulière mais du ou des pays dans lesquels une langue est parlée, du contexte dans lequel elle vit.

### 2.B.1 L'index de développement humain

Nous utilisons ici les données que l'on trouve sur le site du Programme des Nations Unis pour le Développement (PNUD) qui publie annuellement un rapport sur l'état de développement des divers pays du monde<sup>50</sup>. Nous utilisons l'édition la plus récente de ce rapport, mise en ligne au printemps 2017. Les données relatives à l'IDH y figurent dans le tableau 1 des pages 212 et suivantes. L'IDH est un indice composite prenant en compte le produit national brut par individu, l'espérance de vie à la naissance et le niveau d'éducation. Il quantifie le niveau de développement d'un pays. Il est probable que dans de nombreux pays l'indice de développement humain ne soit pas le même pour toutes les régions, pour toutes les ethnies y vivant et donc pour toutes les langues que ces ethnies utilisent. Mais nous ne disposons pas de données plus précises et faisons l'hypothèse que l'indice est constant dans tout le pays, quelque puisse être l'hétérogénéité de celui-ci. Pour affecter une valeur à chaque langue, nous faisons une moyenne pondérée de l'indice dans chacun des pays dans lesquels la langue est parlée. Supposons par exemple que le somali soit parlé en Somalie (51% des locuteurs), en Éthiopie (30%), au Kenya (16%) et à Djibouti (3%). L'indice de développement humain du somali sera donc calculé comme suit :

$$IDH_{\text{somali}} = 0,51 * IDH_{\text{Somalie}} + 0,30 * IDH_{\text{Éthiopie}} + 0,16 * IDH_{\text{Kenya}} + 0,03 * IDH_{\text{Djibouti}}$$

<sup>50</sup> <http://hdr.undp.org/en/2016-report>

Les nombres utilisés ci-dessus ne sont pas rigoureusement exacts mais ils suffisent à expliquer la méthode utilisée. Le site de l'UNDP ne fournit de données que pour les pays affiliés à l'ONU et pour lesquels un indice a été effectivement calculé, ce qui exclut notamment les pays non membres de l'ONU et les pays en guerre. Dans ce cas nous affectons au pays non documenté un indice estimé que nous décidons par analogie avec les pays voisins et/ou comparables. Ainsi, pour la Somalie, pays en guerre depuis plusieurs années et dont les structures étatiques sont défailtantes nous avons affecté une valeur égale à celle du Niger, la plus faible valeur publiée par le PNUD soit 0,348

### 2.B.2 L'indice de fécondité

Nous utilisons ici la même source que précédemment, les données relatives à la fécondité figurent dans le tableau 8 page 238 et suivantes du rapport. Le taux global de fécondité est le nombre de naissance par femme. Il est probable que dans de nombreux pays le taux de fécondité ne soit pas le même dans toutes les régions du pays, pour toutes les ethnies y vivant et donc pour toutes les langues que ces ethnies utilisent. Malheureusement nous n'avons pas accès à des données plus précises et faisons l'hypothèse que le taux de fécondité est constant dans tout le pays. Pour affecter une valeur à chaque langue, nous faisons une moyenne pondérée de l'indice dans chacun des pays dans lesquels la langue est parlée. Par exemple, le somali est parlé en Somalie (65% des locuteurs), en Ethiopie (30%), au Kenya (3%) et à Djibouti (2%). Le taux de fécondité du somali sera donc calculé comme suit :

$$\text{Fécondité}_{\text{somali}} = 0,51 * \text{Fécondité}_{\text{Somalie}} + 0,30 * \text{Fécondité}_{\text{Ethiopie}} + 0,16 * \text{Fécondité}_{\text{Kenya}} + 0,03 * \text{Fécondité}_{\text{Djibouti}}$$

Comme dans le cas de l'IDH, le site de l'UNDP ne fournit de données que pour les pays affiliés à l'ONU et pour lesquels un indice a été effectivement calculé, ce qui ici aussi exclut les pays non membres de l'ONU et les pays en guerre. Dans les cas où le pays n'est pas référencé sur le site nous utilisons, la même méthode que précédemment et estimons la fécondité par analogie avec les pays voisins et/ou comparables.

D'autres sources possibles pour ces données sont «Fecndity Index by country<sup>51</sup>», «Index Mundi<sup>52</sup>» et quelques autres. Les données sont en général cohérentes.

### 2.B.3 La pénétration du réseau internet

Nous utilisons ici les données que l'on trouve sur le site Internet World Stats<sup>53</sup> qui maintient à jour le nombre d'utilisateurs d'internet pour tous les pays du monde et à partir des données démographiques calcule un taux de pénétration en pourcentage de la population, pourcentage que nous avons repris. Le programme des Nations Unis pour le développement publie également un taux d'utilisateurs d'internet<sup>54</sup>, mais, dans un souci de cohérence avec nos travaux antérieurs nous avons conservé les données du site Internet World Stats. Ici encore il est probable que dans de nombreux

---

<sup>51</sup> <https://worldpopulationreview.com/country-rankings/total-fertility-rate>

<sup>52</sup> <https://www.indexmundi.com/>

<sup>53</sup> <http://www.internetworldstats.com/stats.htm>

<sup>54</sup> <http://hdr.undp.org/en/countries/profiles/>

pays le taux de pénétration d'internet ne soit pas le même dans toutes les régions du pays, pour toutes les ethnies y vivant et donc pour toutes les langues que ces ethnies utilisent. Et ici encore nous ne pouvons que considérer par hypothèse que le taux est constant dans tout le pays. Pour affecter une valeur à chaque langue, nous faisons une moyenne pondérée du taux dans chacun des pays dans lesquels la langue est parlée. Par exemple, le somali est parlé en Somalie (65% des locuteurs), en Ethiopie (30%), au Kenya (3%) et à Djibouti (2%). Le taux de pénétration du somali sera donc calculé comme suit :

$$\text{Taux}_{\text{Somali}} = 0,51 * \text{Taux}_{\text{Somalie}} + 0,30 * \text{Taux}_{\text{Ethiopie}} + 0,16 * \text{Taux}_{\text{Kenya}} + 0,03 * \text{Taux}_{\text{Djibouti}}$$

Les données utilisées sont celles indiquées sur le site utilisé en février 2021.

## 3. Traitement des Données

### 3.A Normalisation des valeurs

Les différents facteurs utilisés, comme ceux que nous pourrions ajouter, ne nous donnent pas de valeurs numériques de même type. Une langue est officielle ou non dans un certain nombre de pays, nous obtenons alors un ensemble de valeurs discrètes comprises entre 0 et le nombre le plus élevé de pays dans lesquels une langue est officielle (24 pour le français). Le taux de fécondité nous donne une valeur de type continue entre 1.1 (Macao, Hong Kong) et 6.9 (Niger). Le nombre de locuteurs peut prendre toute valeur entre 0 (une langue morte) et 888.000.000 (chinois mandarin).

Pour donner à chacun des facteurs une importance égale, nous sommes passés des valeurs brutes obtenues comme décrit plus haut à des valeurs normées, en procédant à une transformation linéaire suivant la formule :

$$\text{Valeur Normée} = \frac{(\text{Valeur brute}) - (\text{Valeur Brute minimale})}{(\text{Valeur Brute maximale}) - (\text{Valeur Brute minimale})}$$

Cette transformation affecte la valeur normée 1 à la valeur brute maximale du facteur, la valeur normée 0 à la valeur brute minimale et des valeurs intermédiaires réparties de façon linéaire pour les autres valeurs. Le résultat est que tous les facteurs varient entre 0 et 1 ce qui permet de leur affecter une importance égale dans le classement.

### 3.B Utilisation des logarithmes

Pour certains facteurs le domaine de variation est restreint et s'étend sur deux ordres de grandeur ou moins. La pénétration d'internet est strictement comprise entre 0 et 100%, l'indice de fécondité entre 1,1 et 76.9, l'indice de développement humain est par construction compris entre 0 et 1. Par contre le nombre de locuteurs d'une langue s'étend pratiquement sur neuf ordres de grandeur (de 0 à 888 millions), le nombre d'articles dans Wikipedia sur plus de six, les flux de traduction sur plus de cinq et le nombre de prix littéraires sur pratiquement 2. L'étendue de ces domaines de variation rend difficile la distinction entre les valeurs les plus faibles. Pour contourner cette difficulté nous utiliserons les logarithmes des valeurs brutes, ce qui a pour effet de rapprocher entre elles les valeurs les plus hautes et d'étaler les valeurs les plus faibles. Les valeurs normées, comprises entre 0 et 1 sont ensuite calculée comme indiqué dans le paragraphe précédent.

La figure 3 ci-dessous permettent de bien comprendre l'intérêt d'une telle transformation logarithmique.

## Le poids des langues dans le monde

Le graphique dans la partie supérieure de la figure représente la distribution du nombre de locuteurs. Le mandarin (cmn) en comporte 920 millions, l'hindi 680, l'espagnol 440, l'anglais 360 et le portugais 220. Mais la partie la plus importante de ce graphique est la barre verticale verte sur la gauche, elle contient 6093 langues de moins de 10 millions de locuteurs. Autant dire que ce facteur ne différencie que très peu ou pas du tout entre 0 et 10 millions de locuteurs, ce dont on peut difficilement se satisfaire. En statistique une telle distribution des valeurs est désignée comme dissymétrique positive. Elle est caractérisée de manière chiffrée par le coefficient de dissymétrie qui, pour les amateurs de mathématiques, est le moment centré d'ordre 3 normalisé par le cube de l'écart type. Dans ce cas particulier ce coefficient est égal à 36 ce qui est bien entendu une valeur très élevée.

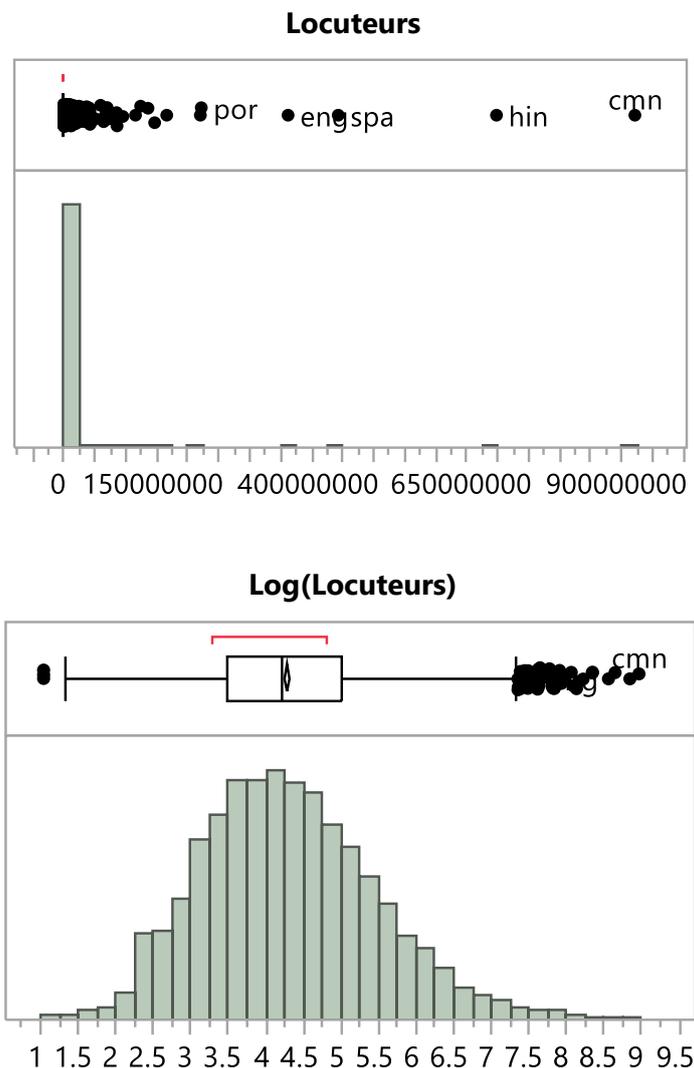


FIGURE 3. DISTRIBUTION DU NOMBRE DE LOCUTEURS. LINÉAIRE ET LOGARITHMIQUE

Le second graphique (partie inférieure) représente le facteur  $\text{Log}(\text{locuteurs})$  également. Il s'agit des mêmes langues ayant le même nombre de locuteurs. La distribution est maintenant pratiquement symétrique, et le coefficient de dissymétrie égal à 0,38. Le facteur devient alors beaucoup plus

## Le poids des langues dans le monde

discriminant, il fait une différence plus nette entre les langues ayant un nombre de locuteurs moyen et faible. Il faut cependant noter qu'il y a un prix à payer pour cette amélioration, il se situe en haut de l'échelle où le groupe de langues les plus parlées en L1 se différencie moins bien du mandarin que précédemment (groupe de points à droite).

Le tableau 9 permet de visualiser la question. Il représente en fonction d'un nombre de locuteurs variant de 1000 (par exemple le lacandon) à 920 millions (le mandarin), le logarithme de ce nombre et les valeurs normées calculées ainsi que défini plus haut dans le paragraphe 2.C.1 pour les valeurs brutes et leurs logarithmes. On y voit très bien l'effet recherché, l'étalement des valeurs basses et moyenne (en rouge), ainsi que le prix à payer, la compression des valeurs élevées (en bleu).

Langue	Locuteurs	Log (Locuteurs)	Norm (Loc)	Norm (Log(Loc))
Mandarin	920 M	20,641	<b>1,000</b>	<b>1,0000</b>
Espagnol	443 M	19,908	<b>0,481</b>	<b>0,960</b>
Javanais	116 M	18,566	<b>0,126</b>	<b>0,886</b>
Afrikaans	10 M	16,127	<b>0,011</b>	<b>0,753</b>
Kambaata	1 M	6,000	<b>0,001</b>	<b>0,626</b>
Alago	100000	5,000	<b>0,000</b>	<b>0,500</b>
Zapotec, Ozolotepec	10000	4,000	<b>0,000</b>	<b>0,373</b>
Lacandon	1000	3,000	<b>0,000</b>	<b>0,247</b>

TABLEAU 9. UTILISATION D'UNE TRANSFORMÉE LOGARITHMIQUE

Pour décider quels seront les facteurs qui subiront cette transformation logarithmique sur nous baserons sur le rapport entre la valeur la plus haute et la plus basse. Si ce rapport est égal ou supérieur à 100, valeur choisie arbitrairement, nous utiliserons la transformation logarithmique des données brutes. Cinq facteurs sur treize seront ainsi traités : le nombre de locuteurs, les deux flux de traductions, le nombre d'articles dans Wikipedia et l'enseignement au niveau universitaire.

### 3.C Indépendance statistique entre les données

Dans tous les problèmes multifactoriels, il faut prendre garde à ce que la multiplication des facteurs n'entraîne une redondance trop importante dans les données. L'approche statistique de cette question consiste à calculer ce que l'on appelle le "coefficient de corrélation linéaire de Pearson", du nom du mathématicien anglais qui l'a défini. L'expression mathématique du coefficient n'a pas d'intérêt ici, il suffit de savoir qu'il varie de -1 à +1. Une valeur de 0 démontre l'absence de corrélation, l'indépendance entre deux colonnes de valeurs. C'est bien sûr la situation idéale souhaitée. Une valeur de 1 indique une corrélation parfaite, les deux facteurs considérés sont totalement équivalents, la redondance est totale et la non prise en compte de l'un des deux facteurs ne fait perdre aucune information. Une valeur de -1 indique également une corrélation parfaite mais dans le sens négatif. On obtient généralement une valeur intermédiaire. Toute valeur inférieure à 0.5 (en valeur absolue) montre une indépendance satisfaisante des deux facteurs, toute valeur supérieure à 0.85 montre une redondance importante, les valeurs intermédiaires s'interprètent en fonction des circonstances.

# Le poids des langues dans le monde

Pour bien comprendre examinons les quatre graphiques de la figure 4, ils représentent en abscisse le facteur internet et en ordonnée de haut en bas et de gauche à droite les facteurs IDH, Wikipedia, entropie et fertilité. Les coefficients de corrélations sont respectivement 0.859, 0.460, 0.155 et -0.762. Il existe une corrélation positive forte entre IDH et internet, une faible corrélation entre Wikipedia et internet, pas de corrélation avec l'entropie et une corrélation négative modérée avec la fertilité. La conclusion est que les facteurs internet et IDH donne essentiellement la même information alors que internet et entropie sont totalement indépendants l'un de l'autre. Nous discutons plus bas de la manière de traiter ce problème en utilisant les coefficients atténuateurs.

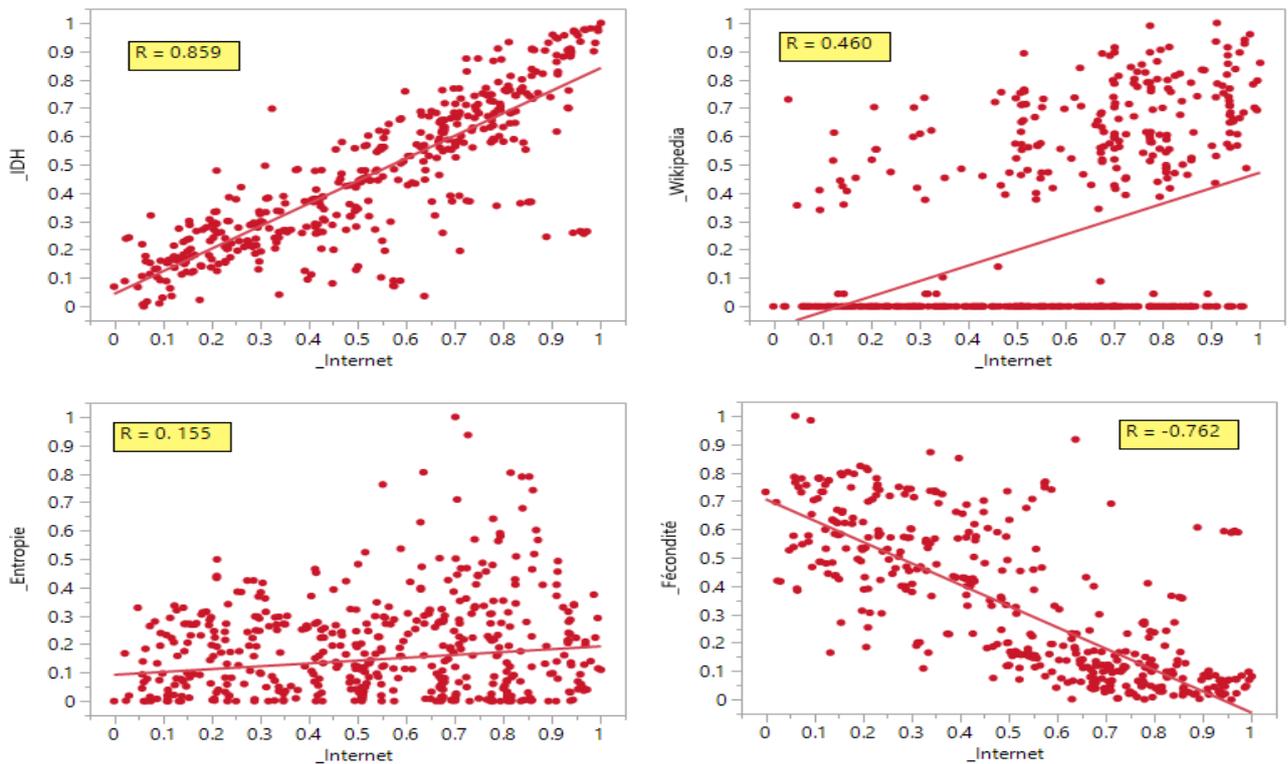


FIGURE 4 CORRÉLATION ENTRE QUELQUES FACTERS

Le tableau 10 indique toutes les corrélations entre les paramètres des langues figurant dans le baromètre.

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

TABLEAU 10. COEFFICIENTS DE CORRÉLATION, 634 LANGUES DU BAROMÈTRE

La conclusion est que nous disposons dans ensemble de facteurs raisonnables les trois quarts des coefficients de corrélation sont inférieure à 0,5

### 3.D Coefficients atténuateurs

En considérant les facteurs que nous avons choisis et les classements que nous avons effectués, l'utilisateur de notre travail peut considérer que tel ou tel facteur n'est pas pertinent ou ne l'intéresse pas et est trop redondant avec tel autre. La non prise en compte de tel ou tel facteur peut être jugée utile dans le cas où l'on étudie un problème particulier. Ainsi, si l'on doit décider dans quelles langues, les menus, tutoriels et programme d'aide d'un nouveau logiciel doivent être rédigés on retiendra plus particulièrement les facteurs locuteurs, accès internet et article dans Wikipedia, ce qui permettra d'optimiser le nombre de clients potentiels. Nous avons donc décidé d'utiliser un ensemble de coefficients « atténuateurs » que l'on utilisera comme multiplicateurs des valeurs normées des facteurs. Ces coefficients prennent une valeur comprise entre 0 et 1. La valeur 0 signifie que le facteur n'est pas pris en compte, la valeur 1 qu'il est jugé de première importance. Toute valeur intermédiaire est possible et est au choix de l'utilisateur du baromètre. Le score global que nous utiliserons pour classer les langues sera donc calculé en appliquant la formule :

$$\text{Score} = \sum_{i=1}^n f_i * w_i$$

dans laquelle le signe  $\sum$  indique que l'on fait la somme sur tous les facteurs de la valeur  $f_i$  du  $i^{\text{ème}}$  facteur multipliée par le coefficient atténuateur  $w_i$  choisi pour ce facteur. Ce score « global » peut varier de manière continue entre 0, tous les produits  $w_i * f_i$  sont nuls) à 12, nombre de facteurs utilisés, situation théorique dans laquelle tous les  $w_i$  et tous les  $f_i$  seraient alors égaux à 1.

37

## 4 Faut-il classer toutes les langues ?

Nous disposons donc d'un fichier contenant 6155 langues décrites par treize facteurs, ce qui nous permet de calculer un score et de les classer toutes les unes par rapport aux autres. Cela est-il raisonnable ?

Lorsque nous examinons leurs valeurs pour les facteurs que nous avons retenus, nous constatons que nombre d'entre elles ne sont parlées que dans un seul pays (leur entropie est donc nulle), ou n'ont aucune fonction véhiculaire, ou n'ont aucun statut officiel, ou n'ont donné lieu à aucune traduction répertoriée dans la base Translationium, ou n'ont reçu aucun prix littéraire, ou n'ont donné lieu à aucun article dans Wikipédia et ne sont enseignées dans aucune université. Plus important encore, de nombreuses langues cumulent plusieurs voire pratiquement toutes ces caractéristiques négatives ! Les comparer les unes aux autres n'a pas beaucoup de sens. Qui y a-t-il de commun entre le mandarin et le féroïen (900 millions et 58000 locuteurs), l'haoussa langue véhiculaire importante dans la bande sahélienne enseignées dans les écoles de langues africaines et orientales et une langue parlée par 2000 personnes dans un village du delta du Niger et inconnue 20 kilomètres plus loin. Quel sens y aurait-il à déclarer savamment que le papaminto, créole des Antilles néerlandaises est classée au 127<sup>ème</sup> rang et le hiri motu, pidgin de Papouasie Nouvelle-Guinée au 678<sup>ème</sup>. Il nous faut bien sûr faire un choix.

# Le poids des langues dans le monde

Mais choisir, c'est forcément éliminer, faire des mécontents et aussi s'exposer à faire des erreurs. Dans les deux éditions précédentes du baromètre nous nous étions fondés sur le nombre de locuteurs, 5 millions en 2010 et 500,000 en 2012 ce qui nous avait conduit à faire figurer 137 puis 563 langues dans le baromètre. Ces choix nous ont semblés judicieux à l'époque, notre vision a changé lorsque nous avons réalisé l'édition 2017 du baromètre et nous avons décidé de complexifier nos critères de sélection.

## 4.A Choix basé sur le nombre de locuteurs

Nous expliquons dans ce paragraphe le choix des 634 langues retenues en 2017. Pour des raisons de cohérence et de facilité de comparaison nous avons, dans cette présente édition, retenu strictement les mêmes 634 langues.

Considérons la figure 5 qui pour les 6141 langues (base de données 2017) relie le score total au logarithme du nombre de locuteurs. À droite de la ligne rouge verticale se situent toutes les langues ayant plus de 500.000 locuteurs. La ligne rouge horizontale se situe à la valeur 2,5 pour le score global. Cette valeur est purement arbitraire mais elle met en évidence le fait qu'un choix de 500.000 locuteurs « promeut » 434 langues (en bleu sur la figure) au détriment de 182 autres (en brun) qui ont un score supérieur dans un classement global de toutes les langues du monde. Nous avons souhaité mettre en cause ce choix basé uniquement sur le nombre de locuteurs.

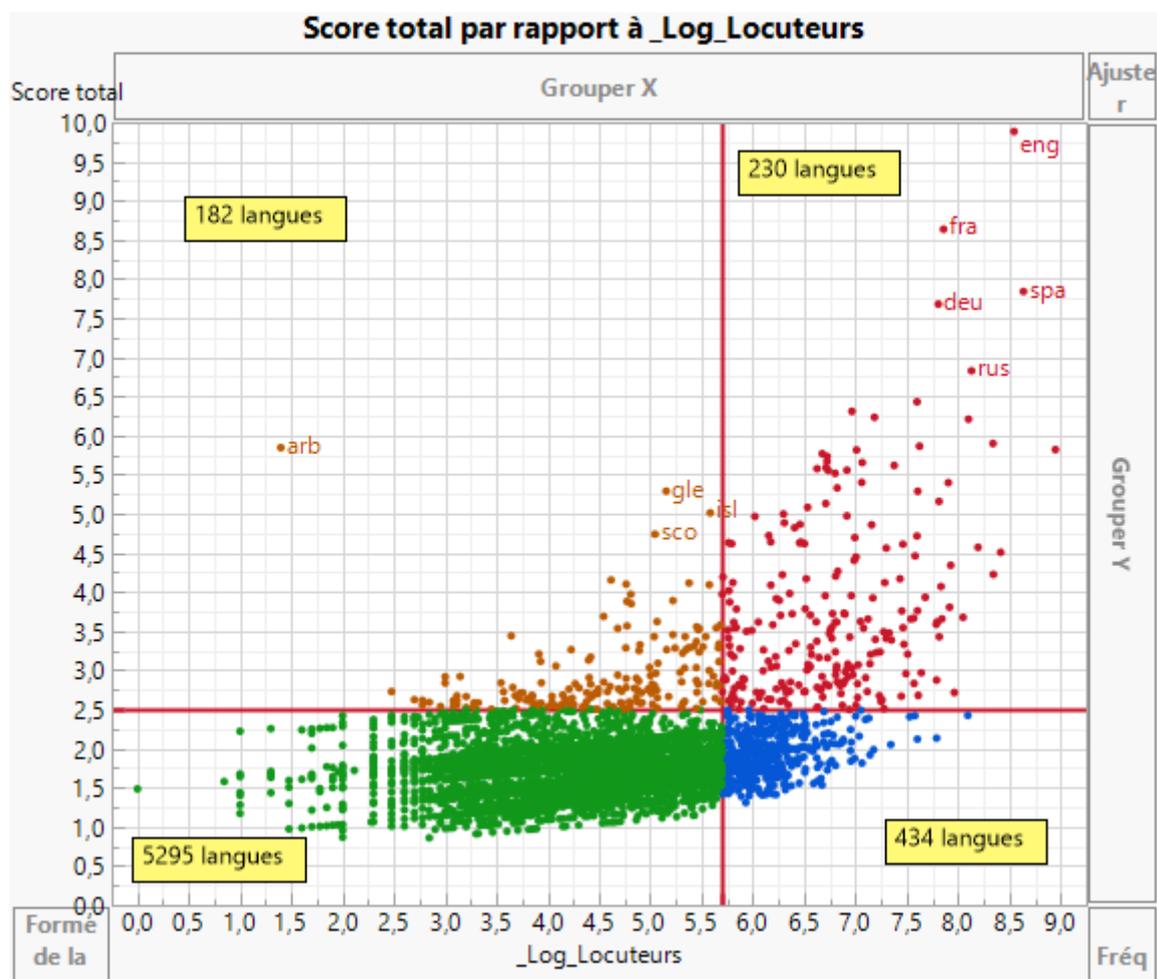


FIGURE 5. SÉLECTION DES LANGUES DU BAROMÈTRE BASÉE SUR LE NOMBRE DE LOCUTEURS

# Le poids des langues dans le monde

Pour diminuer cet inconvénient nous avons tracé une ligne non plus verticale mais oblique joignant les points de coordonnées {0,6} et {8,0}. Ceci élimina les langues « bleues » ayant les scores les plus faibles et donna une chance à des langues « brunes » ayant des scores élevés d'être réintégrées. Nous y ajoutons une autre condition : nous n'avions retenu que les langues ayant plus de 300.000 locuteurs. La figure 7 permet de visualiser le résultat de ces deux critères.

## 4.B Choix basé sur l'importance des facteurs conjoncturels

Il existait un autre problème que nous souhaitions traiter : l'influence sur le score total des paramètres *conjoncturels*, IDH, Internet et fécondité. Le score « du pays » (IDH + Internet + fécondité) ne dépend que du ou des pays dans lesquels la langue considérée est parlée. Faisons le rapport entre ce score du pays et le score total. La figure 6 et les tableaux 11 et 12 ci-dessous nous amènent aux réflexions suivantes :

Le rapport varie entre 0,20 et 0,91. La valeur de la médiane de cette distribution nous indiquait que pour une langue sur deux ce rapport est supérieur à 0,7. Pire encore il était supérieur à 0,5 pour 5781 langues soit 94% du total.

Le tableau 11 indique par exemple qu'un certain nombre de langues indigènes australiennes auraient été classées dans les 600 premières parmi 6141 et que cela est du non pas à la langue elle-même mais à l'Australie qui compte pour 88% du score, cela n'aurait eu pas de sens.

A l'inverse, le tableau 12 montre les vingt langues pour lesquelles le rapport était le plus faible. On y retrouvait dix des vingt langues les mieux classées par le baromètre, ce qui était satisfaisant, le classement des langues les plus « lourdes » ne dépendait que partiellement de critères géographiques, des paramètres que nous qualifions de contextuels.

Notre conclusion était qu'il apparaissait raisonnable de définir une limite supérieure pour ce rapport, nous avons choisi la valeur de 2/3. Ceci éliminait les langues qui « profitent » du pays dans lequel elles sont parlées : langues aborigènes d'Australie, langues indiennes du Canada, same de Norvège et bien d'autres.

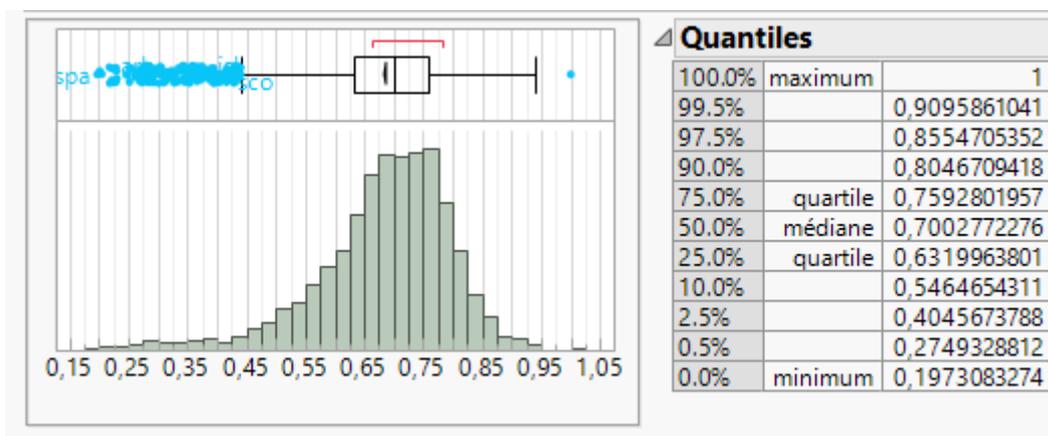


FIGURE 6. DISTRIBUTION DU RAPPORT (SCORE CONJONCTUREL)/(SCORE GLOBAL)

## Le poids des langues dans le monde

_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
nbg	[nbg]	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

TABLEAU 11. RAPPORT (SCORE CONJONCTUREL) / (SCORE TOTAL) ÉLEVÉ

_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

TABLEAU 12. RAPPORT (SCORE CONJONCTUREL) / (SCORE TOTAL) FAIBLE

### 4.C Choix final de 634 langues

La figure 7 visualisait donc notre choix. Les langues retenues figurent en rouge, les rejetées en bleu. Le logarithme du nombre de locuteurs est porté en abscisse. La ligne rouge verticale sélectionne les langues avec plus de trois-cent-mille locuteurs. La ligne oblique bleue permet de diminuer l'inconvénient traité au paragraphe 4.A. Nous y ajoutons la nécessité d'avoir un rapport (score conjoncturel) / (score total) supérieur à 0,6667. Les langues rejetées par ce dernier filtre sont celles figurant en bleu au-dessus et à droite des deux lignes bleue et rouge. On notera que l'arabe «standard», le gaélique et le scots et bien d'autres peuvent se juger sacrifiés.

## Le poids des langues dans le monde

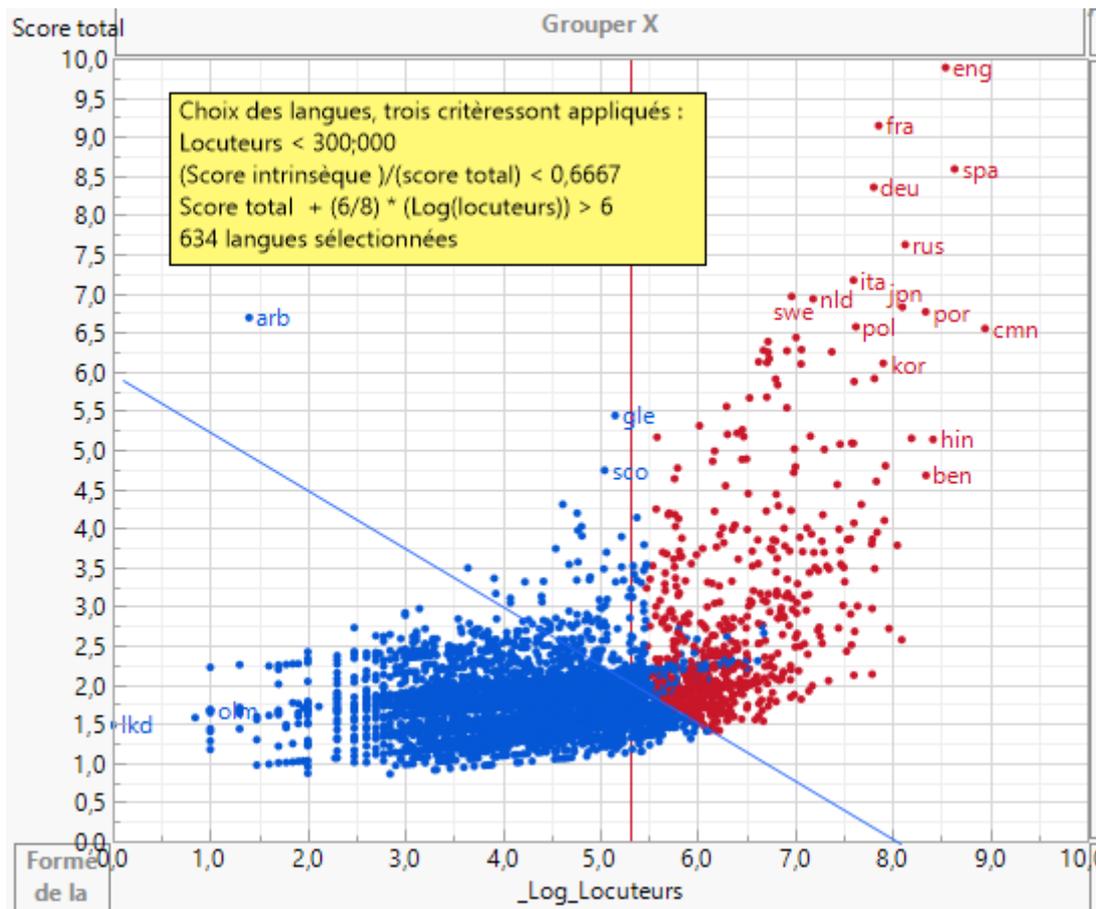


FIGURE 7 SÉLECTION DES LANGUES BASÉES SUR TROIS CRITÈRES

Nous restions alors avec 634 langues, ce sont celles qui ont été classées dans notre édition du baromètre 2017. Ce sont celles qui sont classées dans cette présente édition.

A partir de ce point nous redéfinissons les valeurs 0 et 1 de tous les facteurs qui correspondent donc désormais aux valeurs minimales et maximales des facteurs des langues retenues. Le classement devient donc un classement interne cohérent de ces langues et de ces langues seules.

Sur le diagramme ci-dessus l'arabe standard [arb] apparaît comme un point aberrant, isolé de tous les autres. Son score dans l'ensemble des 6141 langues est très élevé et le classe au 12<sup>ème</sup> rang. Cependant dans toutes les compilations des langues et de leurs locuteurs les arabes dialectaux sont retenus comme langue L1 des arabophones et il est impossible de connaître le nombre de locuteurs de l'arabe standard, en supposant qu'il y en ait. Le critère des 300.000 locuteurs minimum entraîne donc son élimination du classement final. Nous sommes évidemment conscients du fait que l'arabe est une grande langue au niveau mondial, officielle dans bien des organisations internationales et héritière d'une grande histoire et d'une grande culture, mais comment faire ? La situation dans les pays arabophones semble quelque peu schizophrène, on y parle une langue que l'on n'écrit pas et on y écrit une langue que l'on ne parle pas. Nous ne sommes pas responsables de cette situation, mais elle est la réalité.

## 5 Du bon usage du baromètre

Nous l'avons dit plus haut, notre baromètre se veut flexible et il est toujours possible, en utilisant les coefficients atténuateurs attachés aux facteurs, de moduler le score de manière telle qu'il réponde aux questions que se pose l'utilisateur. Nous souhaitons maintenant proposer quatre exemples de scores que nous qualifions de « scores standard ». Ils donnent quatre points de vue différents sur le classement des langues du monde.

Les cinquante premières langues pour chacun des quatre scores que nous allons décrire sont reportées dans le tableau 13.

### 5.A Score global

Le score prend en compte tous les facteurs que nous avons définis plus haut en affectant le coefficient 1 à tous les facteurs, c'est le score maximum possible que peut atteindre la langue considérée. Il donne une vue générale du classement des langues dans le monde. Il est à proprement parler la mesure du poids des langues au niveau mondial

A l'examen de la tête du classement (tableau 8) notre attention est attirée par la présence d'une forte proportion de langues européennes (cellules colorées en bleu) : dix, vingt et trente-quatre dans les dix, vingt-cinq et cinquante premières langues. En tête du classement nous trouvons principalement les langues des pays qui ont constitué un empire colonial, mais le japonais et le mandarin sont également bien placés. Tous ces langues allient un nombre de locuteurs élevé, une production culturelle abondante ou reconnue et elles sont parlées dans des pays dont la puissance économique est importante. L'importance des facteurs sociaux économiques est très apparente lorsque l'on considère les rangs du néerlandais, du suédois, du norvégien, du finnois et du danois qui se trouvent tous entre la 9<sup>ème</sup> et la 24<sup>ème</sup> place.

Ce score global est donc un point de vue mais il n'est bien sûr pas le seul et nous voudrions maintenant introduire d'autres manières de juger du poids des langues.

### 5.B Score intrinsèque

Nous l'avons vu, des milliers de langues ont des valeurs nulles sur plusieurs voire la majorité des paramètres intrinsèques, ceux qui ne dépendent pas des pays dans lesquels ces langues sont parlées. Ceci signifie que si elles n'ont pas un nombre de locuteurs très élevé l'essentiel de leur score total est dû aux facteurs contextuels qui se rapportent aux pays dans lesquels la langue est parlée. Ce phénomène peut se reproduire à un degré plus ou moins important pour la plupart des langues, c'est pourquoi il apparaît intéressant de considérer un classement prenant uniquement en compte les dix facteurs qui ne se rapportent qu'à la langue et non plus à son aire de distribution. Nous l'appelons le score intrinsèque.

Lorsque l'on se reporte au tableau 8 des classements on observe une stabilité des toutes premières langues classées (cellules vertes). Le groupe des langues nordiques citées plus haut recule en bloc (cellules grises), elles ont perdu l'avantage que leur conférait leurs pays très développés. On observe aussi l'avancée des langues du sous-continent indien ainsi que des langues asiatiques qui progressent de un ou plusieurs rangs.

Ce classement peut être considéré comme celui des langues d'avenir voire de l'avenir. Les langues asiatiques et africaines peuvent être considérées comme « pénalisées », lorsque nous considérons le score global, par le fait d'appartenir à des pays moins avancés que les pays « occidentaux ». Lorsque les pays dans lesquels elles sont parlées auront rattrapé tout ou partie de leurs « retards » économique et/ou culturel, elles progresseront dans le classement global.

### 5.C Score démographique

Dans les pays en voie de développement, le niveau économique et le niveau d'éducation pris en compte par les facteurs contextuels ne sont pas les seuls points dont on attend qu'ils s'améliorent dans un avenir à moyen ou long terme. Il est probable que le progrès économique entrainera le progrès culturel pris en compte par les facteurs flux de traduction, Wikipedia et prix littéraires. Pour anticiper cette évolution, il apparaît intéressant de définir un classement ne prenant en compte que les éventuels points forts actuels des langues de ces pays en développement, à savoir le nombre de locuteurs et la véhicularité. Nous qualifions ce score de démographique.

Le point important ici est la présence majoritaire des langues du sous-continent indien ainsi que des langues asiatiques, on en trouve respectivement 8, 19 et 40 dans les 10, 25 et 50 premières langues classées. En corollaire, la présence européenne est très largement diminuée.

### 5.D Score prestige

De manière similaire les langues des pays développés ont des points forts qui les distinguent des langues des pays en voie de développement. Le statut officiel de langues comme l'anglais ou le français provient de l'importance des anciens empires coloniaux et en a fait la langue des élites dans un grand nombre de pays. Le niveau élevé d'éducation des pays développés a pour corollaire la reconnaissance de cette culture par l'attribution de prix littéraires ainsi que le développement des flux de traduction. Cet aspect du poids des langues est mis en évidence par le score prestige qui est la somme de ces trois facteurs auquel s'ajoute le facteur universités.

Ce classement est en quelque sorte complémentaire des deux précédents et on observe le retour des langues européennes, huit, dix-neuf et trente-trois dans les dix, vingt-cinq et cinquante premières respectivement.

On notera également que l'indonésien (bahasa indonesia) qui dans les trois classements précédents se situait aux 17<sup>ème</sup>, 12<sup>ème</sup> puis 4<sup>ème</sup> rang se retrouve maintenant en 36<sup>ème</sup> position.

Pour nous résumer nous pouvons dire que les scores global et prestige couronnent les langues "établies" alors que les scores intrinsèque et démographique donnent une vision de ce que pourrait être le panorama des langues du monde dans l'avenir.

## Le poids des langues dans le monde

Rang	Classement suivant différents scores			
	Total	Intrinsèque	Démographique	Prestige
1	anglais	anglais	anglais	anglais
2	français	français	ourdou	français
3	espagnol	espagnol	français	espagnol
4	allemand	allemand	indonésien	allemand
5	russe	russe	yué	russe
6	italien	italien	tagalog	italien
7	suédois	portugais	javanais	japonais
8	roumain	roumain	hindi	mandarin
9	portugais	mandarin	thaï	portugais
10	polonais	polonais	dyula	polonais
11	néerlandais	suédois	tok pisin	suédois
12	catalan	indonésien	russe	tchèque
13	tchèque	catalan	swahili	néerlandais
14	croate	swahili	créole anglais du Cameroun	coréen
15	mandarin	croate	mandarin	danois
16	hongrois	tchèque	oromo, ouest central	norvégien
17	indonésien	néerlandais	lingala	hongrois
18	japonais	hongrois	igbo	hébreu
19	norvégien	farsi	bamanankan	grec
20	finnois	turc	amharique	serbe
21	turc	japonais	mòré	turc
22	danois	ourdou	roumain	finnois
23	farsi	finnois	azéri du Nord	croate
24	swahili	serbe	allemand	roumain
25	slovaque	coréen	espagnol	farsi
26	serbe	hindi	zoulou	catalan
27	tagalog	norvégien	haoussa	bulgare
28	estonien	danois	sotho du Nord (sepedi)	hindi
29	lithuanien	slovaque	oromo borana	slovène
30	coréen	vietnamien	sango	slovaque
31	slovène	tagalog	farsi	estonien
32	ukrainien	ukrainien	afrikaans	ukrainien
33	vietnamien	bulgare	thaï du Nord Est	lithuanien
34	islandais	afrikaans	xhosa	islandais
35	grec	estonien	krio	bengali
36	ourdou	lithuanien	hiligaynon	indonésien
37	galicien	grec	tswana	ourdou
38	basque	slovène	nyanja	arménien

## Le poids des langues dans le monde

---

39	hindi	azéri du nord	oromo de l'Est	macédonien
40	bulgare	kazakh	efik	basque
41	hébreu	bosniaque	créole anglais du Nigéria	vietnamien
42	kazakh	bengali	dari	ouzbèque du Nord
43	azéri du Nord	ouzbèque du Nord	sotho du sud (sesotho)	tamoul
44	bosniaque	haoussa	arménien	galicien
45	gallois	thaï	wolof	thaï
46	letton	malais	assamais	gallois
47	haoussa	galicien	vietnamien	biélorusse
48	malais	basque	catalan	tibétain, central
49	afrikaans	arménien	asturien	bosniaque
50	asturien	albanais, tosk	malais des moluques	géorgien

**TABLEAU 13. TÊTE DU CLASSEMENT POUR 4 SCORES STANDARD**

### 5.D Scores personnalisés

Il vous appartient, utilisateur du baromètre, de bâtir votre propre score en jouant sur les curseurs des coefficients atténuateurs. Chacun des paramètres peut se voir attribuer un coefficient variant de manière continue entre 0 et 1 suivant votre vision des langues ou le problème quei vous est posé.

A vous de jouer !