Baromètre des langues dans le monde

Ce travail a été réalisé par MM. Alain Calvet et Louis-Jean Calvet en 2017 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/ ou écrivez à Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Sommaire

1. Introduction

En avril 2010, nous mettions en ligne sur le site de l'Union Latine un « baromètre des langues du monde »[1], que le lecteur consultera avec profit afin de comprendre la méthode que nous utilisions 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 avions 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 nouvelle édition du baromètre a été mise en ligne en 2012[2]. Elle permettait le classement d'un nombre de langues plus important et utilisait un facteur de plus pour effectuer ce classement.

Nous mettons aujourd'hui en ligne une troisième édition de notre baromètre. Nous utilisons maintenant douze 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 *leur 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 plus de 500.000 pages en « arabe » sans autre spécification et 17.000 en arabe «égyptien». De manière similaire si nous nous intêressons au nombe 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 mettre un peu d'ordre dans ce désordre, nous utilisons la

[1] http://www.observatoireplurilinguisme.eu/en-us/cross-cutting-themes/reference-texts-en-us-6/6620-barometre-calvet-des-langues-du-monde-source-portalingua

[2] http://wikilf.culture.fr/barometre2012/

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 sesotho, sepedi et tswana) doivent ils être considérés comme une seule langue ou trois langues différentes. Le complexe des langues peul doit 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'interintellingibilité. 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.

Nous nous en sommes donc strictement tenus à la classification dite ISO 693-3[3] qui bien qu'imparfaite présente l'avantage d'être cohérente et d'attacher à chaque langue un code univoque de trois lettres. Il est ainsi possible de résoudre les 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"[4] qui en fait une seule langue avec environ vingt-cinq millions de locuteurs.

Il nous faut donc décider quelles langues nous examinerons et pour cela trouver une compilation fiable des langues du monde. A notre connaissance il en existe au moins trois accessibles sur le web, Ethnologe[5], Joshua[6] et People Groups[7]. Le problème est que ces trois sites mélangent une volonté scientifique de compilation des langues et de leurs locuteurs et un aspect religieux plus ou moins présent. Dans nos deux premières éditions nous avions 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 16553 enregistrements. Joshua cite les sources de ces données, elles sont Ethnologue 19ème édition dans

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

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

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

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

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

16445 cas, World Fact Book 2015[8] dans 89, UNESCO 2010 dans 19. Dans 33 cas la source n'est pas indiquée. Finalement nous avons donc obtenu une base de 6244 langues différentes prenant en compte un peu moins de 7,4 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 6141 langues regroupant 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 Sene (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[9] 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 6141 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 »[10] 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.

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

[8] https://www.cia.gov/library/publications/the-world-factbook/

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

[10] Dictionnaire des langues, E. Bonvini et alii, Quadrige P.U.F. 2011.

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 etpeuvent être séparés en deux catégories

En premier lieu nous décrirons ceux qui se rapportent à une langue et à elle seule, nous les appelerons *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 connait avec précision le nombre d'habitants dans le monde et certains géographes considérent 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 fonction qui permet de quantifier le « désordre ». Elle a été utilisée à l'origine en thermodynamique, puis a trouvé des applications en théorie de l'information et plus récemment en linguistique[11]. Nous l'utilisons ici pour différencier une langue parlée dans un seul pays

[11] http://unesdoc.unesco.org/images/0014/001421/142186f.pdf

d'une langue parlée dans plusieurs pays. 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 :

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

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 Σ 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	0.705	0.199	1.166	2.536	3.219	0.123
Locuteurs L1	135 M	126M	350M	433M	?	888M

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

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 très 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 nomreux 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 bien supérieure. L'arabe standard n'a pas un nombre important de locuteurs en langue première mais est considérée comme une langue présente dans tous les pays arabo-musulmans, son entropie est 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 situations 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 éfléchit et s'exprime le plus naturellement. Nous utiliserons donc ici la notion de langue première, ou 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 enfin apprendre sur le tas différentes langues que l'on n'utilisent 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 L2 pour toutes les situations dans lesquelles une langue autre que la L1 est utilisée dans la vie sociale.

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.

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, c'est-à-dire un fait qui ne résulte pas d'une décision gouvernementale *in vitro* mais de pratiques sociales *in vivo*.

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 parlent 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 Etats 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 avions 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 20ème édition indique pour un nombre important de langues, nous en avons relevé 154, le nombre de locuteurs en L1 et en L2.

- Les études par pays du site Laval sont 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[12,13] 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 fat cependant être prudent et s'assurer que le terme secondary speakers correspondbien à des locuteurs de L2.

- Les sites gouvernements se rapportant aux résultats de référendums sont également utiles dans les cas où les langues maternelles et secondes sont documentées[14].

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[15] et plus de six-cents millions selon une autre[16]. Le français est parlé en seconde langue par environ cinquante millions[17], ou cent cinquante trois millions[18]. 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,

---

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

[13] http://nalrc.indiana.edu/

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

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

[16] http://en.wikipedia.org/wiki/World_language et références citées

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

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

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.

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	http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
		17 M		http://www.lmp.ucla.edu/Profile.aspx?LangID=7&menu=004
		21 M	21 M	http://www.tlfq.ulaval.ca/axl/afrique/ethiopie.htm
		17 M	?	http://www.nalrc.indiana.edu/brochures/amharic.pdf
		32 M?		http://www.plc.sas.upenn.edu/languages/amharic.html
Anglais	[eng]	341 M	167 M	http://www.nationsonline.org/oneworld/most_spoken_languages.htm
		371 M	611 M	https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers
		340 M	170 M	http://www.vistawide.com/languages/top_30_languages.htm
		372 M	612 M	Ethnologue 20ème édition
Tamoul	[tam]	68 M	8 M	http://www.ethnologue.org/show_language.asp?code=tam
		67M	8 M	http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
		68 M	9 M	http://www.vistawide.com/languages/top_30_languages.htm
		68 M	8 M	http://en.wikipedia.org/wiki/World_language
		66 M	M's	http://www.lmp.ucla.edu/Profile.aspx?LangID=99&menu=004
		52M	M's	http://www.plc.sas.upenn.edu/languages/tamil.html

TABLEAU 2. DIFFERENTES 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 souligne en même temps les graves lacunes dans la connaissance des situations sociolinguistiques. En particulier les données, lorsqu'elles existent, ne sont pas actualisées de manière régulière. Il est dans ce domaine souhaitable 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[19].

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.

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 Etats Unis d'Amérique etc..

---

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

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 Etats 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 papiamento 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 le communauté autonome de Galice sont dans cette situation. Ces cas sont comptabilisés.

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 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 Vénézuela (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, araona, baure, bésiro, canichana, cavineño, cayubaba, chácobo, chimán, ese ejja, guaraní, guarasu'we, guarayu, itonama, leco, machajuyai-kallawaya, machineri, maropa, mojeño-trinitario, mojeño-ignaciano, moré, mosetén, movima, pacawara, puquina, quechua, sirionó, tacana, tapiete, toromona, uru-chipaya, weenhayek, yaminawa, yuki, yuracaré 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 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, gwich'in, inuvialuktun, inuktitut et inuinnaqtun[20]. Sous l'index "Official Languages" ce site Internet propose l'annonce ci-dessous :

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

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 1er
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.*

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[21]:

*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 manikakan et trois songaï, quant au madenkan, nous ne le trouvons nulle part. La conséquence de tout ceci est que, comme dans le cas du nombre de locuteurs, il nous faudra admettre un certain degré d'approximation dans nos données.

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 indépendamment trois facteurs correspondants aux langues des trois niveaux décrits et les combinerons ensuite en affectant un coefficient 1 aux langues officielles des états souverains, 0.5 à celles des états non souverains et 0.25 aux autres langues distinguées pour quelque raison que ce soit.

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, Bruneï, la Malaisie, l'Indonésie et Singapour.

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

c1) 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.

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

c2) 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 elle est retenue pour les deux facteurs. Le swati, officiel au Swaziland et dans la province du Mpumalanga en Afrique du Sud en est un exemple, il sera comptabilisé une fois pour chaque facteur.

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

Nous utilisons ici les données de l'Index Translationium[22] 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 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 le hindi et le 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 serbo-croate. 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

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

de l'arabe. Le tableau 3 regroupe ces données. Nous rencontrons un problème analogue avec la compilation d'articledans 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			
	Standard	Marocain	Tchadien	Dialectes
Code Translationium Code ISO 693-3	ARA [arb]	ARY [ary]	SHU [shu]	ARA-DI ?
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	
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 A PARTIR DES ARABES STANDARD ET DIALETCAUX

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

Nous utilisons ici les données de l'Index Translationium. 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[23]. 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'occidentaleset 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 condamnait 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 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[24], le prix Man Booker[25], le prix Franz Kafka[26], le prix Ovid[27],le prix de Jerusalem[28], l'American Award in literature[29] et le prix « Golden

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

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

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

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

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

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

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

Wreath ». Nous avons également retenu le prix Prince des Asturies[30] à partir de 1999. Avant cette date les lauréats étaient tous de langue espagnole.

De nombreux autres prix littéraires existent mais n'ont pas pour vocation d'examiner des candidats originaires du monde entier[31]. 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. Nous ne tenons pas compte des prix dont le domaine est trop limité.

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 attibuer 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.

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[32]. 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 ou nous rassemblons les données. Signalons ici que Wikipedia n'utilise pas le code ISO 693-3 pour identifier sans ambigüité 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 pour la première fois dans le 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 :

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

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

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

-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 ans African Studies, au Royaume-Uni), du NARLC (National African Languages Resource Center, aux États-Unis), du CIIL (Central Institue 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 :

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[33].

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 au 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[34]. 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[35].

---

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

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

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

2.c Les universités figurant au classement des meilleures universités asiatiques dans le domaine "Arts and humanities"[36] Cent universités sont reprises dans ce classement.

2.d Un site listant "toutes" les universités dans "tous" les pays du monde[37] .

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[38] , le SOAS[39], (School of Oriental ans African Studies), le CIIL en Inde et le NALRC[40] (National African Languages Resource Center). Il peut y en avoir d'autres.

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

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[41] "School of Foreign Languages and Cultures" de l'Université Nationale de Mongolie[42]. 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[43]. Il faut alors explorer l'arborescence du site pour s'assurer que

---

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

[37] http://univ.cc/

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

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

[40] http://www.nalrc.indiana.edu/

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

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

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

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.

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 le 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 Azebaïdjan 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 291 différentes langues (caractérisées par leur code ISO 639-3) enseignées dans 966 universités situées dans 172 pays. Le total des unviversités dans le monde étant de l'ordre de 20.000 nous estimons la marge d'incertitude à environ 3%.

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[44]. 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 Ethiopie (30%), au Kenya (16%) et à Djibouti (3%). L'indice de développement humain du somali sera donc calculé comme suit :

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

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 défaillantes nous avons affecté une valeur égale à celle du Niger, la plus faible valeur publiée par le PNUD soit 0,348

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

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 :

$$Fécondité_{somali} = 0,51 * Fécondité_{Somalie} + 0,30 * Fécondité_{Ethiopie} + 0,16 * Fécondité_{Kenya} + 0,03 * IFécondité_{Djibouti}$$

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.

2.B.3 La pénétration du réseau internet

Nous utilisons ici les données que l'on trouve sur le site Internet World Stats[45] 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[46], 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 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 :

$$Taux_{somali} = 0,51 * Taux_{Somalie} + 0,30 * Taux_{Ethiopie} + 0,16 * Taux_{Kenya} + 0,03 * Taux_{Djibouti}$$

Les données utilisées sont celles indiquées sur le site utilisé en février 2017.

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

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

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 7,6 (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 Utililisation 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 7,6, 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 1 ci-dessous permettent de bien comprendre l'intêret d'une telle transformation logarithmique.

Le graphique y dans lapartie supérieure de la figure représente la distribution du nombre de locuteurs normé entre 0 et 1. Le mandarin (cmn) est à la valeur 1, suivi d'assez loin par l'espagnol puis l'anglais (spa et eng) avec respectivement 433 et 351 millions de locuteurs. Mais la partie la plus importante de ce graphique est la barre verticale verte sur la gauche, elle contient 6090 langues (ligne horizontale bleue) dont le nombre de locuteurs va de 0 à 22 millions, la référence de valeur 1 étant le mandarin avec 888 millons de locuteurs leur score va de 0 à 0,025. Autant dire que ce facteur ne différencie pas entre 1000 et 20 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 38,96 ce qui est une valeur très élevée.

Le second graphique (partie inférieure) représente le facteur Log(locuteurs) également normé entre 0 et 1. Il s'agit des mêmes langues ayant le même nombre de locuteurs. La ligne rouge verticale se situe à la valeur 0,82 et à sa gauche se trouvent les 6090 langues qui se situaient entre 0 et 0,025 sur le graphique supérieur La distribution est maintenant pratiquement symétrique, et le coefficient de dissymétrie égal à 0,335. Le facteur devient alors beaucoup plus 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 symbolisées par des points rouges se différencient moins bien du mandarin que précédemment.

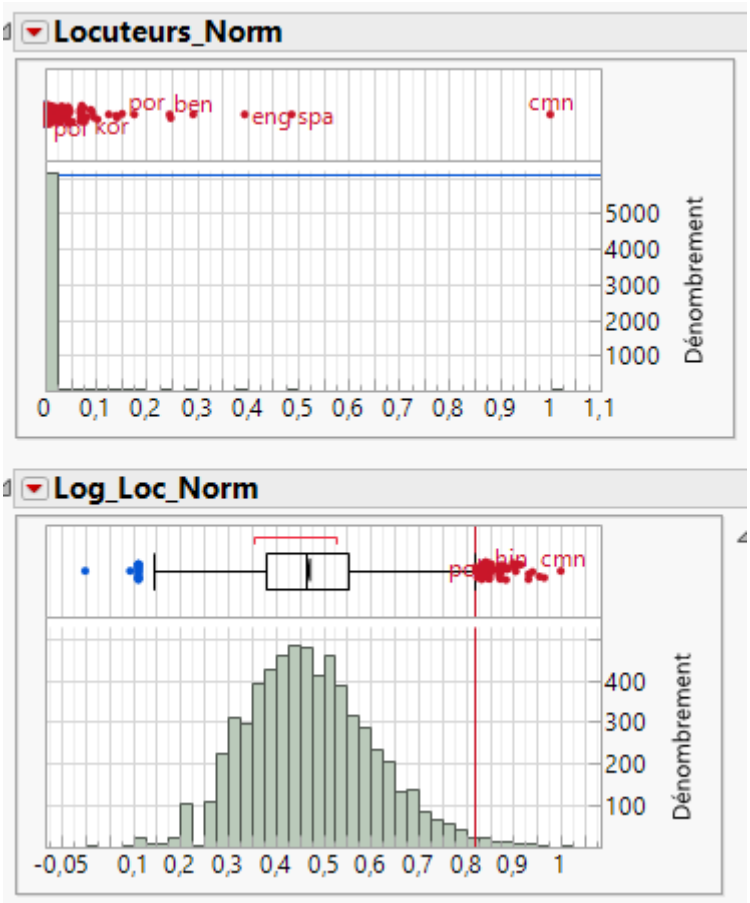


FIGURE 1 DISTRIBUTION DU NOMBRE DE LOCUTEURS. LINEAIRE ET LOGARITHMIQUE

Le tableau 4 permet de visualiser la question. Il représente en fonction d'un nombre de locuteurs variant de 1000 (par exemple le bas saxon) à 888 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, ainsi que le prix à payer, la compression des valeurs élevées.

Langue	Locuteurs	Log (Locuteurs)	Norm (Loc)	Norm (Log(Loc))
Mandarin	888 M	8,948	**1,000**	**1,0000**
Espagnol	433 M	8,637	**0,487**	**0,965**
Javanais	112 M	8,048	**0,126**	**0.899**
Arabe hijazi	10,12 M	7,004	**0,011**	**0,783**
Morysien	1,079 M	6,033	**0,001**	**0,674**
Cachoube	100000	5,000	**0,000**	**0,559**
Zapotèque ozolope	10000	4,000	**0,000**	**0,447**
Bas saxon	1000	3,000	**0,000**	**0,340**

TABLEAU 4. UTILISATION D'UNE TRANSFORMÉE LOGARITHMIQUE

Pour décider quels seront les facteurs qui subiront cette transformation logarithmique sur nous baserons sur le coefficient de dissymétrie : toute valeur supérieure à 10, valeur choisie arbitrairement, justifiera le traitement logarithmiqe des données. Sept facteurs sur douze seront ainsi traités : le nombre de locuteurs, les deux flux de traductions, le nombre d'articles dans Wikipedia, les prix littéraires, l'enseignement au niveau universitaire et le statut des langues.

3.C Indépendance statistique entre les données

Dans tous les problèmes multifactoriels, il faut prendre garde à ce que la multiplication des facteurs n'entraine 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 a mis 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 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.

Pour bien comprendre examinons les quatre graphiques de la figure 2, 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 0.919, 0.507, 0.186 et -0.779. 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 conclusin est que les faceurs internet et IDHdonne 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.

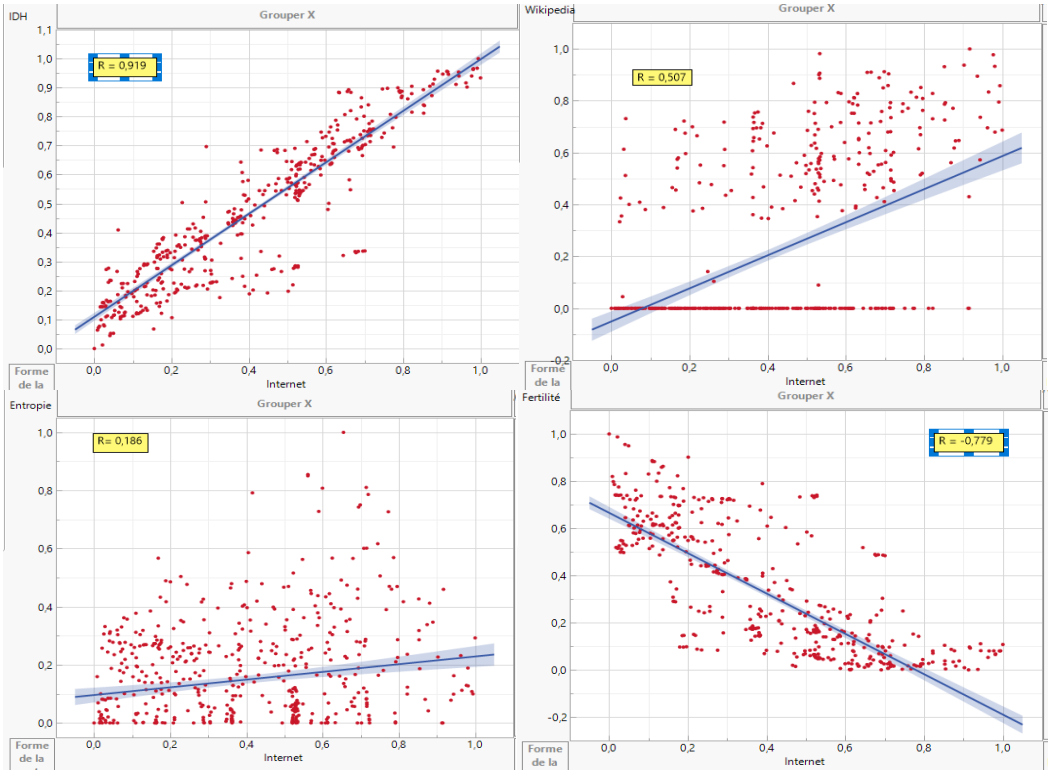


FIGURE 2 CORRÉLATION ENTRE QUELQUES FACTERS

Le tableau 5 indique toutes les corrélations entre les paramètres des langues figurant dans le baromètre.

Corrélations

	Locuteurs	Entropie	IDH	Fertilité	Internet	Cible	Source	Wikipedia	Statut	Universités	Véhicularité	Prix
Locuteurs	1,0000	0,1118	0,1468	-0,1777	0,1566	0,4680	0,4864	0,5312	0,5134	0,6070	0,2455	0,3915
Entropie	0,1118	1,0000	0,1688	-0,0741	0,1861	0,3120	0,3188	0,2132	0,3808	0,2553	0,0550	0,2238
IDH	0,1468	0,1688	1,0000	-0,8700	0,9194	0,4679	0,4925	0,5323	0,3174	0,3368	0,0011	0,3295
Fertilité	-0,1777	-0,0741	-0,8700	1,0000	-0,7790	-0,3759	-0,3818	-0,4380	-0,2422	-0,2475	-0,0365	-0,2098
Internet	0,1566	0,1861	0,9194	-0,7790	1,0000	0,4540	0,4863	0,5069	0,3180	0,3427	0,0384	0,3247
Cible	0,4680	0,3120	0,4679	-0,3759	0,4540	1,0000	0,9402	0,7540	0,7826	0,7877	0,2098	0,6171
Source	0,4864	0,3188	0,4925	-0,3818	0,4863	0,9402	1,0000	0,7746	0,8048	0,8036	0,2056	0,6546
Wikipedia	0,5312	0,2132	0,5323	-0,4380	0,5069	0,7540	0,7746	1,0000	0,6486	0,6406	0,2291	0,4508
Statut	0,5134	0,3808	0,3174	-0,2422	0,3180	0,7826	0,8048	0,6486	1,0000	0,8457	0,3193	0,7443
Universités	0,6070	0,2553	0,3368	-0,2475	0,3427	0,7877	0,8036	0,6406	0,8457	1,0000	0,2790	0,8369
Véhicularité	0,2455	0,0550	0,0011	-0,0365	0,0384	0,2098	0,2056	0,2291	0,3193	0,2790	1,0000	0,0979
Prix	0,3915	0,2238	0,3295	-0,2098	0,3247	0,6171	0,6546	0,4508	0,7443	0,8369	0,0979	1,0000

TABLEAU 5. COEFFICIENTS DE CORRELATIONS , 634 LANGUES DU BAROMETRE

La conclusion est que nous disposons dans ensemblede facteurs raisonable les deux tiers des coefficients de corrélation sont inférieur 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 multiplicateur 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 :

$$\text{Score} = \sum_{i=1}^{n} fi * wi$$

dans laquelle le signe ∑ indique que l'on fait la somme sur tous les facteurs de la valeur f_i du $i^{ème}$ 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, situatio théorique dans laquelle tous les w_i et tous les f_i seraient alors égaux à 1.

4 Faut il classer toutes les langues ?

Nous disposons donc d'un fichier contenant 6141 langues décrites par douze 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 enseignéesdans 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 féroïenn (900 millions et 58000locuteurs), le 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 papamiento, créole des antilles néerlandaises est classée au 127ème rang et le hiri motu, pidgin de Papouasie Nouvelle-Guinée au 678ème. Il nous faut bien sûr faire un choix.

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 aujourd'hui changé et nous avons décidé de complexifier nos critères de sélection.

4.A Choix basé sur le nombe de locuteurs

Pour nous expliquer, considérons la figure 3 qui pour les 6141 langues relie le score total au logarithme du nombre de locuteurs. A 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 toutesles langues du monde. Nous avons souhaité mettre en cause ce choix basé uniquement sur le nombre de locuteurs.

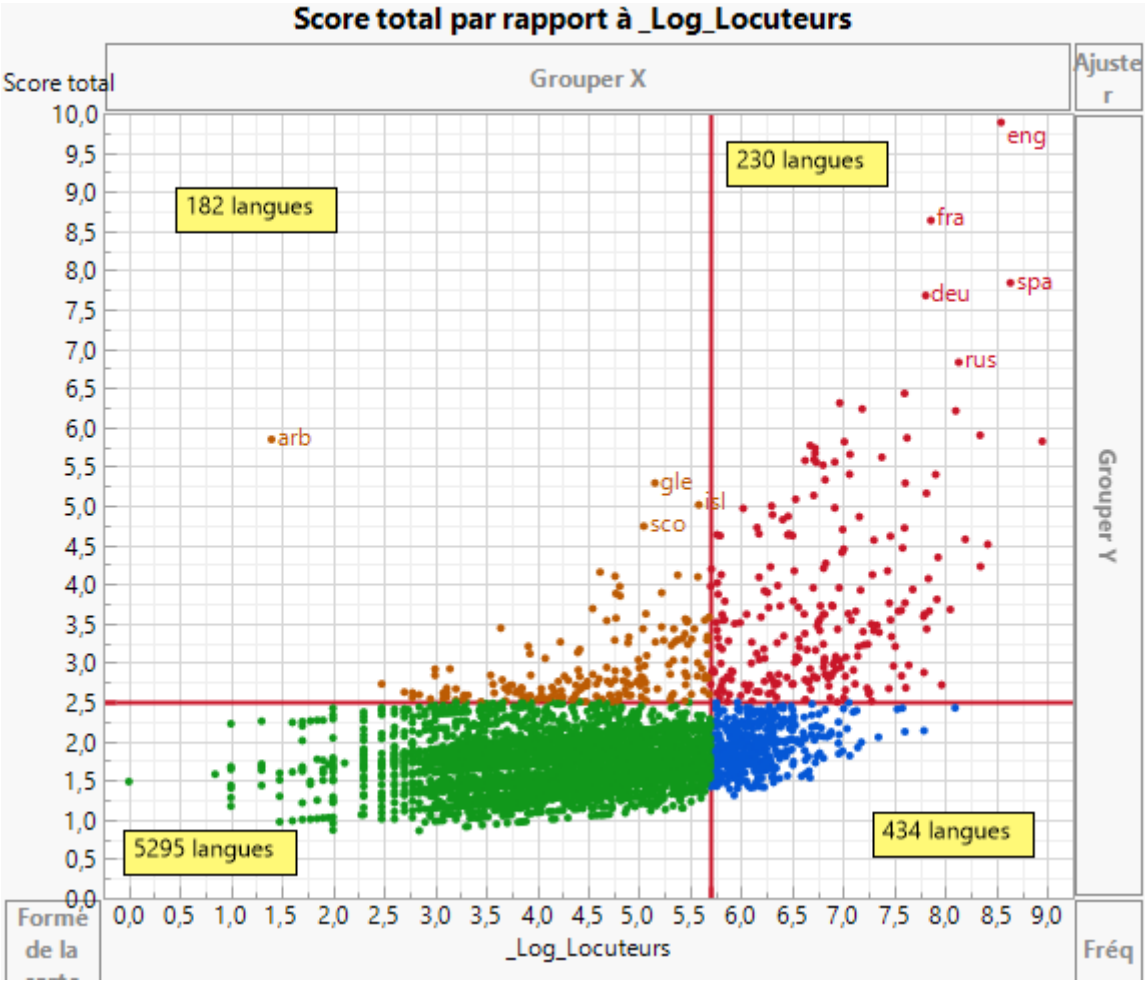


FIGURE 3. SELECTION DES LANGUES DU BAROMETRE BASEE SUR LE NOMBRE DE LOCUTEURS

Pour diminuer cet inconvénient nous tracerons une ligne non plus verticale mais oblique joignant les points de coordonnées {0,6} et {8,0}. Ceci éliminera les langues « bleues » ayant les scores les plus faibles et donnera une chance à des langues « brunes » ayant des scores élevés d'être réintégrées. Nous y ajouterons une autre condition : nous ne retiedrons que les langues ayant plus de 300.000 locuteurs. La figure 5 permet de visualiser le résultat de ces deux critères.

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

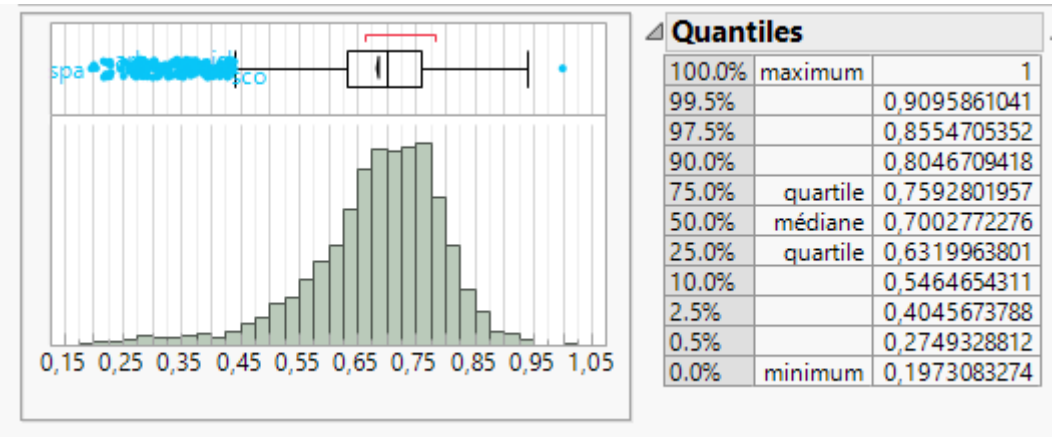
Il existe un autre problème que nous souhaitons aborder ici : l'influence sur le score total des paramètres «conjoncturels», IDH, Internet et fécondité. Le score «du pays» (IDH + Internet + fécondité) ne dépend que du ou des pays dans lesquels la langue considérée est parlée. Faisons le rapport entre ce score du pays et le score total. La figure 4 et les tableaux 6 et 7 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 indique que pour une langue sur deux ce rapport est supérieur à 0,7. Pire encore il est supérieur à 0,5 pour 5781 langues soit 94% du total.

Le tableau 6 indique par exemple qu'un certain nombre de langues indigènes australiennes seraient 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 pas de sens.

A l'inverse, le tableau 7 montre les vingt langues pour lesquelles le rapport est le plus faible. On y retrouve dix des vingt langues les mieux classées par le baromètre, ce qui est satisfaisant, le classement des langues les plus « lourdes » ne dépend que partiellement de critères géographiques.

Notre conclusion est qu'il apparait raisonnable de définir une limite supérieure pour ce rapport, nous avons choisi la valeur de 2/3. Ceci éliminera 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.



Quantiles

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

FIGURE 4. DISTRIBUTIO DU RAPPORT (SCORE CONJONCTUREL)/(SCORE GLOBAL)

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

TABLEAU 6. RAPPORT (SCORE CONJONCTUREL)/(SCORE TOTAL) ELEVE

_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 7. RAPPORT (SCORE CONJONCTUREL)/(SCORE TOTAL) FAIBLE

4.C Choix final de 634 langues

Le diagramme XXx visualise 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 paraghe 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.

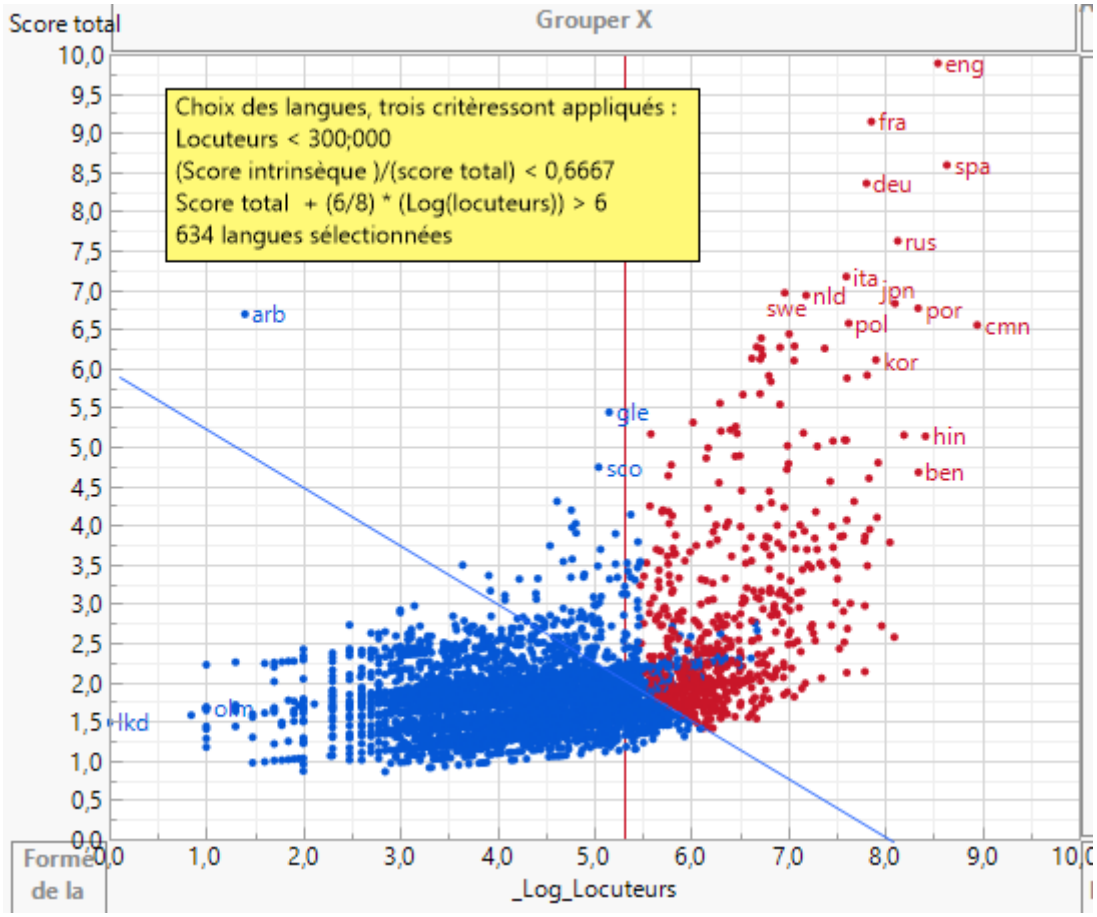


FIGURE 4 SELECTION DES LANGUES BASEES SUR TROIS CRITERES

Il nous reste alors 637 langues, qui seront classées dans notre édition du baromètre 2017.

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 ainsi un classement interne cohérent de ces langues et de ces langues seules.

Sur le diagrame ci-dessus l'arabe standard [arb] apparait comme un point aberrant, isolé de tous les autres. Son score dans l'ensemble des 6141 langues est très élevé et le classe au 12ème 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 connaitre le nombre de locuteurs L1 de l'arabe standard, en supposant qu'il y en ait. Le critère des 300.000 locuteurs minimum entraine donc *ipso facto* 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 il est impossible de déroger à une règle (ici celle du seuil des 300.000 locuteurs) au bénéfice d'une seule langue. La situation des pays arabophones pourrait en effet être définie comme une *schizoglosie* : on y parle une langue que l'on n'écrit pas et on y écrit une langue que l'on ne parle pas. Et il est impossible d'aller contre les faits pour introduire dans un classement qui repose sur le traitement statistique de données objectives.

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 8.

Dans ce tableau nous avons codé par un fond bleu les langues européennes, vert les langues africaines, ocre les langues asiatiques, mauve les langues de l'ex-URSS et bronze les langues du moyen-orient et du monde arabo-musulman. Ceci permet de visualiser les grandes tendances des classements.

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 (tableau8) notre attention est attirée par la présence d'une forte proportion de langues européennes : seize dans les vingt premières et plus de la moitié des cinquante premières langues. En tête du classement nous trouvons les langues des pays qui avait constitué un empire colonial, mais le japonais, le mandarin et le turc 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 9ème et la 24ème 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 elles 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 lequels 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 apparait intéressant de considérer un classement prenant uniquement en compte les neuf 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. Ensuite le portugais prend la septième place qui était occupé par le néerlandais. Le groupe des langues nordiques citées plus haut recule en bloc. Une observation intéressante est 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. Mandarin, turc, farsi, indonésien, hindi ou swahili sont en partie « 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 sans doute 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 part 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 apparait 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.

La présence de langues (sept) européenne est maintenant largement diminuée. Ce sont les langues de l'Inde, de l'Asie et surtout de l'Afrique (24 langues) qui figurent dans ce classement.

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, sept, quatorze et trente dans les dix, vingt et cinquante premières respectivement.

On notera également que l'indonésien (bahasa indonesia) qui dans les trois classements précédents se situait aux 28ème, 22ème puis 7ème rang se retrouve maintenant en 37ème 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.

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 qui vous est posé.

A vous de jouer !

Rang	Scores			
	Total	Intrinsèque	Démographique	Prestige
1	anglais	anglais	anglais	anglais
2	français	espagnol	ourdou	français
3	espagnol	français	français	allemand
4	allemand	allemand	yué	espagnol
5	russe	russe	tagalog	russe
6	italien	italien	javanais	italien
7	portugais	mandarin	indonésien	japonais
8	japonais	portugais	thaï	mandarin
9	néerlandais	japonais	dyula	portugais
10	suédois	polonais	russe	polonais
11	mandarin	roumain	mandarin	suédois
12	polonais	néerlandais	tok pisin	tchèque
13	tchèque	suédois	hindi	hongrois
14	croate	coréen	swahili	hébreu
15	roumain	serbe	oromo, ouest central	danois
16	serbe	croate	igbo	coréen
17	hongrois	hongrois	lingala	norvégien
18	coréen	tchèque	créole anglais Cameroun	serbe
19	norvégien	turc	amharique	turc
20	danois	grec	bamanankan	grec
21	grec	arménien	roumain	croate
22	hébreu	indonésien	azéri du Nord	finnois
23	catalan	farsi	allemand	néerlandais
24	finnois	swahili	espagnol	roumain
25	turc	catalan	mòoré	slovène
26	arménien	hindi	haoussa	bulgare
27	farsi	ourdou	zoulou	farsi
28	indonésien	hébreu	néerlandais	catalan
29	swahili	bulgare	sotho du Nord (sepedi)	hindi
30	slovaque	danois	farsi	albanais, tosk
31	bulgare	norvégien	afrikaans	slovaque
32	slovène	finnois	sango	ukrainien
33	ourdou	ukrainien	oromo borana	estonien
34	hindi	bengali	thaï du Nord Est	bengali
35	kazakh	slovaque	xhosa	macédonien
36	tagalog	thaï	arménien	lithuanien
37	ukrainien	kazakh	hiligaynon	indonésien
38	lithuanien	tagalog	nyanja	ourdou
39	albanais, tosk	vietnamien	sindhi	letton
40	estonien	slovène	krio	bosniaque
41	bosniaque	tamoul	tswana	arménien
42	macédonien	albanais, tosk	oromo de l'Est	islandais
43	thaï	malais	sotho du sud (sesotho)	basque
44	malais	macédonien	wolof	galicien
45	azéri du Nord	bosniaque	efik	tamoul
46	vietnamien	azéri du Nord	créole anglais du Nigéria	vietnamien
47	islandais	afrikaans	assamais	ouzbèque du Nord
48	bengali	lithuanien	vietnamien	thaï
49	letton	ouzbèque du Nord	dari	biélorusse
50	galicien	estonien	catalan	gallois

TABLEAU 8. TETE DU CLASSEMENT POUR 4 SCORES STANDARD