maaya

réseau mondial pour la diversité linguistique

# Étude sur la présence de la langue française dans le cyberespace

**Etude réalisée par Daniel Pimienta**
**pour le compte de MAAYA.**

**Rapport final #2 - mai 2017**
**auteur: Daniel Pimienta**

# RESUMÉ EXÉCUTIF

Cette étude sur la présence la langue française dans le cyberespace fait suite à deux études similaires réalisées en 2012 et 2013. Les principes méthodologiques sont les mêmes mais ils ont été améliorés et appliqués sur un nombre d'indicateurs à la fois beaucoup plus important et plus diversement équilibré, ce qui valide un traitement statistique des données collectées et la création d'indicateurs. La méthode consiste à rechercher des sources numériques qui fournissent une série de micro-indicateurs à propos des langues dans l'Internet et dont le traitement permettra d'obtenir, directement ou indirectement, et pour chaque langue, des indicateurs exprimés en pourcentage mondial. Un nombre limité des sources rencontrées s'applique directement aux langues; un nombre beaucoup plus important s'applique aux pays. Les indicateurs-pays sont convertis en indicateurs-langue par une méthode originale de pondération des valeurs d'indicateurs-pays avec la répartition des langues (L1) dans chaque pays. Les résultats obtenus pour les L1 sont ensuite recalculés pour la somme L1 plus langue seconde (L2) à partir d'hypothèses globales sur les L2 exprimées en termes de taux d'accroissement des valeurs L1 pour obtenir les valeurs de L1+L2. Les résultats permettent de quantifier de manière raisonnable la présence de la langue française dans l'Internet en comparaison avec l'ensemble des langues, pour un ensemble de 6 indicateurs (internautes, trafic, usages, contenus, indexes et interfaces). A partir de ces 6 indicateurs sont construits 3 macro-indicateurs qui renseignent la puissance, la capacité et le gradient des langues dans l'Internet. Les résultats en terme de puissance montrent une langue française (L1+L2) en très solide quatrième position (après l'anglais, le chinois et l'espagnol) et avec une marge d'avance importante sur les langues qui la suivent (russe, allemand et portugais). La quote-part moyenne mondiale du français dans l'Internet est estimée à 7,5% (L1+L2). Les indicateurs établis prennent en compte les personnes connectées (5,6%), le trafic (7,8%), les usages (7,4%), les contenus - surtout en relation avec Wikimédia- (9,3%), des indexes évaluant les pays dans la société de l'information (7,4%) et les langues d'interface ou de traduction (7,3%). La marge d'avance moyenne sur la langue en cinquième position est proche de 3%. L'ensemble des méthodes, sources et traitements réalisés est analysé en profondeur pour en comprendre les limitations et les biais possibles; cette analyse permet d'en déduire et d'en comprendre les implications sur la qualité des résultats produits. Certaines sources sont marquées par un très fort biais en faveur de l'anglais, au détriment, en premier lieu, des langues non occidentales; l'analyse plus fine des résultats permet toutefois de déterminer que la correction de ces biais importants, si elle était possible, pourrait rabaisser la quote-part de l'anglais qui ressort des calculs (de 37% à 32%) au bénéfice des langues largement sous-estimées comme le chinois, l'hindi, le bengali, l'ourdou et l'arabe, sans que cela ne change la quatrième place du français. De même, diverses simulations réalisées sur des scénarios différents avec les valeurs de L2, ou encore dans une première tentative d'intégration du concept plus large de langue d'usage de l'Internet (Li), montrent la solidité du français dans cette quatrième place. L'ensemble des expérimentations conduites avec le dispositif produit par cette étude démontre sa capacité pour jouer un rôle d'outil de modélisation et simulation, voire même de prospective.

# Table des matières

## LISTE DES TABLES ET FIGURES

## TABLES



## FIGURES

## 1. Avertissement méthodologique

**À propos de la gestion des données concernant les langues première et seconde (L1+L2):**
Quand une quantité (ou un pourcentage mondial) est exprimée pour quantifier les locuteurs d'une langue première ou seconde qui sont connectés à l'Internet, il faut comprendre que si l'on fait la somme pour toute les langues de cette variable, elle sera logiquement supérieure à la population mondiale des internautes (et la somme des pourcentages par rapport à la population mondiale sera supérieur à 100%). En effet, la même personne si elle est bilingue apparaitra 2 fois dans cette comptabilité (3 fois si elle est trilingue, 4 fois si elle parle quatre langues, etc.). Il en est de même quand il est dit que 230 millions de personnes (données de Ethnologue[1]) ont le français comme première ou deuxième langue: si l'on comptabilisait toutes les personnes par langue de cette manière on trouverait un total mondial bien supérieur au 7,3 milliard d'êtres humains et un pourcentage supérieur à 100%. Avec les données prises comme hypothèses (les données L2 de Ethnologue et une estimation d'un accroissement de 18% pour le reste des langues), cette valeur mondiale serait de 9,1 milliards et le "facteur de multilinguisme" atteindrait 25% (donc le pourcentage de référence serait de 125%). Ce nombre de 1,25 peut se comprendre intuitivement comme le nombre moyen de langues parlées par personne.

Ainsi, quand on parle de la population mondiale L1+L2, on parle d'une population de locuteurs de l'ensemble des langues, qui se mesure avec un nombre dont l'excédent par rapport à la population mondiale représente le comptage des personnes monolingues plus bilingues[2] (avec les données actuelles, un peu plus de 1,8 milliards, soit 25% de la population mondiale).

Dans le scénario qui est adopté, basé sur les données de Ethnologue et avec un supplément moyen de 18% accordé par prudence (et ajusté pour créer l'arrondi) aux langues absentes de la liste des 140 langues traitées, les quotes-parts (ce ne sont plus des pourcentages au sens strict) de données seront calculées sur la base d'une somme qui serait égale à 125% de la population mondiale.

Cette notion s'applique tout aussi logiquement à l'ensemble des concepts mesurés: internautes, trafic, usages, contenus, interfaces et indexes. Quand il est dit que la quote-part du français (L1+L2) dans l'Internet est de 7,5%, on peut comprendre que le pourcentage de la part du trafic francophone (courriel, chat, navigation sur le web, participation dans les réseaux sociaux, contenus) peut être estimée en moyenne à 7,5% en ayant clairement conscience que le multilinguisme va s'appliquer à tous ces éléments (les sites web consultés peuvent être en plusieurs langues, de même pour les flux de courriels ou de chat et les échanges dans les réseaux sociaux) et donc si on additionnait les mêmes pourcentages pour l'ensemble des langues du monde le total serait supérieur à 100%. Dans le cas de la Toile, il est probable que dans la réalité le degré de multilinguisme d'un site web soit supérieur à celui de ses visiteurs, cependant l'approximation qui consiste à les considérer équivalents est tout à fait acceptable. Pour les personnes qui serait heurtées par un pourcentage avec des totaux supérieurs à 100%, ce qui semble une contradiction sémantique avec la structure du mot pourcentage, il est préférable en effet de parler de *quote-part* et il faut faire l'analogie avec des statistiques portant sur le "pourcentage" d'abonnements à des services mobile qui dans certains pays dépassent les 100% puisque certaines personnes cumulent ces abonnements (idem pour les téléphones mobiles). Dans les exemples exposés on utilise l'expression "par habitant" au lieu de pourcentage. Dans ce rapport l'abus

[1] Un portail de données démo-linguistiques. https://www.ethnologue.com/
[2] Dans cette logique une personne qui possède 3 langues est comptabilisée 3 fois, une fois pour la langue maternelle et une fois pour chacune des langues secondes.

de langage que représente l'utilisation du mot *pourcentage* sera accepté et le mot *quote-part* (ou d'autre fois le mot *taux*) sera également utilisé pour exprimer ces chiffres portant sur l'ensemble de locuteurs-langues, population dont le total est supérieur à la population humaine.

Pour permettre des comparaisons plus faciles avec des données du même genre qui seraient exprimées en pourcentage de la population mondiale, chaque fois qu'il sera question de données L1+L2 en pourcentage une valeur par rapport à ce total (de 125% avec les données démo-linguistiques actuelles) sera indiquée et en même temps (parfois dans une note de bas de page) le même pourcentage normalisé a 100% sera indiqué afin de permettre des comparaisons avec des données exprimées de cette manière par ailleurs.

## 2. Introduction

Ce projet, réalisé entre octobre 2015 et mars 2017, et qui fait l'objet du protocole d'accord du 19/10/2015 (référence DNFDL F/ITF/OBS/AW/ph/20151015-004), propose de mesurer la présence de la langue française dans l'Internet, de constituer un système informatique permettant de gérer de manière autonome de futures mesures, et de préparer une page web compatible avec le site de l'OIF, avec les résultats des différentes mesures, analyses et synthèses.

Le travail se base sur la collecte d'informations quantitatives concernant l'usage des langues dans de nombreuses applications et espaces de l'Internet. La compilation et l'organisation de ces données permettent la mesure de micro-indicateurs de la présence des langues et la mise en perspective des résultats par la construction d'une série d'indicateurs qui mesurent la quote-part correspondante à la langue française dans l'Internet, en termes d'internautes, de trafic généré, d'usage de services, de contenus, d'interfaces à des logiciels et langues de traduction et en fonction d'une série d'indexes qui évaluent des critères sociétaux. Une synthèse de l'ensemble des indicateurs de la présence du français dans l'Internet est produite en forme d'un macro-indicateur qui englobe l'ensemble des paramètres pour l'ensemble des langues et qui permet ainsi d'estimer sa place mondiale et sa quote-part dans l'Internet par rapport à l'ensemble des langues. Les regroupements des micro-indicateurs par thématique permet de plus de différencier le potentiel du français selon ces thématiques et de déterminer ainsi les atouts du français dans l'Internet et ses faiblesses relatives[3].

La complexité mise en jeu par la méthode, résultant de la grande quantité de données à manipuler et de la transformation de certaines données par pays en données par langue, appelle un soin particulier dans la rédaction de la partie méthodologique et dans la description des limites et biais possibles associés avec chacun des paramètres.

Il est important de comprendre que l'orientation principale de l'étude est celle de **mesurer la puissance mondiale des langues dans l'Internet** et en conséquence les indicateurs seront, dans la mesure du possible, exprimés **en pourcentages mondiaux**. Cette approche tient évidement compte à la fois de la taille de la langue en termes de population et de sa **capacité** propre dans l'Internet (en relation avec ses locuteurs) ou de son **gradient** dans l'Internet (en relation avec ses locuteurs qui sont connectés) et elle met logiquement en avant les langues qui ont un bassin important de locuteurs, dans la proportion de leur présence dans l'Internet.

[3] Cela sera réalisé seulement dans le cas où il y a suffisamment de micro-indicateurs dans cette thématique pour que la statistique prenne sens.

Si l'on choisissait une autre métrique, comme par exemple des pourcentages par langue ramenés au total de locuteurs de cette langue, on obtiendrait avec les même données, des résultats équivalents mais mesurant cette fois la **capacité des langues dans l'Internet** (indépendamment de leur taille). Avec ce critère, des langues à forte capacité dans l'Internet, même si la taille de leur population ne leur permet pas d'atteindre une puissance planétaire[4], apparaitraient au dessus des langues avec des populations qui s'expriment en centaines de millions, dont le français, lequel détient cependant, en terme de capacité, un score très haut de 3[5], qui le place à la dixième place parmi les 140 langues exposées[6] (l'ensemble des langues de plus de 5 millions de locuteurs).

Les résultats obtenus montrent en synthèse que le français **maintient et renforce sa position de quatrième langue de l'Internet**, tous critères confondus, avec une quote-part mondiale de l'ordre de **7,5%**[7] pour une population d'internautes de **5,6%**[8]. La **grande solidité de cette quatrième place**, avec une **avance forte sur les langues suivantes,** a de plus été évaluée et confirmée par diverses simulations sur les paramètres du modèle établi.

Les domaines très forts du français (en terme de trafic vers des sites) sont les **cours en ligne ouverts et massifs, la recherche scientifique, les réseaux sociaux pour la musique et les réseaux sociaux professionnels.** Les domaines forts sont les **l'informatique et les jeux.** Il faut aussi noter une bonne présence du français dans **la galaxie Wikipédia.**

En synthèse, le français est une des langues phares de l'Internet avec des atouts dans les **contenus plutôt professionnels**, en particulier **à caractère scientifique et ouverts**.

## 3. Contexte et antécédents

Cette étude consiste en une troisième étape dans une série de mesures réalisées avec des méthodologies comparables, pour obtenir un classement du français dans l'Internet. La première a été réalisée en 2012, dans le cadre d'une étude soutenue par l'OIF et la seconde, réalisée en 2013, a nourri le chapitre Internet du rapport 2014 "Le Français dans le Monde", publié chez Nathan.

Pendant la période 1998-2007, les deux responsables de cette étude collaboraient, à partir de leurs institutions respectives, l'Association Réseaux & Développement (FUNREDES) et l'Union Latine, pour des méthodes de mesure de la places des langues dans l'Internet susceptibles de fournir des indicateurs reproductibles et fiables[9]; à la même période d'autres initiatives[10] existaient avec les mêmes objectifs. A partir de 2007, l'évolution de la taille du Web et des politiques des moteurs de

[4] C'est le cas par exemple, dans l'ordre, des langues suivantes hébreu, suédois, finnois, néerlandais, hongrois, allemand, anglais, tchèque, italien et danois qui surpassent le français dans cette métrique.
[5] Une valeur de 1 représente une capacité normale (proportionnelle à la présence dans le monde non virtuelle); une valeur inférieure à 1 serait une valeur faible, et la capacité est d'autant plus forte que la note est supérieure à l'unité. A titre de comparaison l'anglais a une capacité mesurée de 3,6 et l'arabe de 0,9.
[6] L'onglet SYNTHESES du fichier de Excel de référence présente les classements de toutes les langues par capacité ainsi que par pourcentage de personnes connectées et par gradient.
[7] De 6,2% si ramené à une base de 100%.
[8] De 4,5% si ramené à 100%.
[9] Voir *Douze années de mesure de la diversité linguistique sur l'Internet: bilan et perspectives*, D. Pimienta, D. Prado et Á. Blanco, UNESCO, 2009. Accessible à: http://unesdoc.unesco.org/images/0018/001870/187016f.pdf
[10] En particulier l'ambitieux Language Observatory Project (LOP) dirigé par le japonais Yoshiki Mikami de l'université de Nagaoka.

recherche a rendu obsolètes ces méthodes et a créé un vide dans la production d'indicateurs des langues dans l'Internet[11]. L'OIF a alors contribué, avec l'Union Latine et l'UNESCO, à un effort proposé par MAAYA, entre 2010 et 2012, pour lancer des travaux de recherche ambitieux, avec l'objectif de combler ce vide, et susceptibles d'être financés par le Programme Cadre de l'Union Européenne (projet DILINET[12]). Cependant cet effort n'a pu aboutir malgré l'obstination et la qualité des équipes de recherche impliquées dans un consortium regroupant des acteurs prestigieux de la recherche européenne.

C'est pour combler ce vide d'une manière plus pragmatique et plus économique, quoique moins ambitieuse, que cette méthode nouvelle, basée sur l'observation du comportement des langues dans une grande variété d'espaces et d'applications de l'Internet, a été proposée par Daniel Prado en 2012 et a ouvert une nouvelle collaboration avec Daniel Pimienta[13], sous le chapeau institutionnel de MAAYA.

Si le cadre méthodologique est commun entre les trois mesures historiques, il existe une complexification (par l'augmentation du nombre des éléments mesurés) et des perfectionnements méthodologiques à chaque étape. Cette troisième étude permet d'atteindre un niveau de perfectionnement et de généralisation notable de la méthode qui justifie la mise au point d'indicateurs capables de mesurer la quote-part du français dans les principaux éléments fonctionnels de l'Internet et la mise à disposition d'un outil de simulation.

Le cadre méthodologique commun entre les trois études consiste à utiliser le plus grand nombre possible de sources numériques disponibles pour quantifier la place des langues dans l'Internet, soit directement, lorsque les statistiques accessibles concernent la langue[14], ce qui est malheureusement peu fréquent, soit indirectement, et de manière plus fréquente, en transformant des données par pays en données par langue[15].

Cette transformation de données pays en données langue fait de cette méthode une approche originale qui n'a pas de précédent identifié jusqu'à aujourd'hui et qui lui confère la capacité de traiter la question linguistique dans l'Internet, dans un contexte où les indicateurs de langue sont devenus, au mieux, très peu fiables mais surtout et le plus souvent inexistants. Cette approche est soutenue par des hypothèses qui demandent à être évaluées et un certain nombre de précautions doivent être prises pour en assurer la cohérence et la fiabilité. La discussion sur les limites et les contrôles qui ont été réalisés pour garantir la fiabilité d'une méthode qui fait appel à une certaine complexité, tant dans les calculs réalisés que dans la compréhension des concepts qui en résultent, occupera en conséquence une part importante de ce rapport.

Cette dernière étude est un aboutissement de la méthode grâce aux bénéfices des leçons apprises lors des études passées et à l'importance de ce nouvel effort consenti par l'OIF.

[11] Il semblerait que le cataclysme vécu par le Japon en mars 2011 a signé un arrêt du projet LOP à la même période.
[12] http://dilinet.org
[13] Et avec Álvaro Blanco de FUNREDES.
[14] Par exemple, le nombre d'articles de Wikipédia, ou le nombre de livres chez Amazon.
[15] Par exemple, la répartition en pourcentage du trafic généré par pays du moteur de recherche Google permet de déduire le trafic par langue vers Google.

# 4. Méthodologie

## 4.1Indicateurs des langues dans l'Internet

Les types de pratiques ou usages de l'Internet se rapportent à des *applications* (par exemple le moteur de recherche Google ou le réseau social Facebook) ou des *espaces*, au sens large du terme (par exemple les téléphones intelligents ou le gouvernement électronique). Quand des sources quantitatives fiables pourront être identifiées, des *micro-indicateurs* de la place des langues dans l'Internet seront définis en relation avec l'application ou l'espace en question. La catégorisation des espaces et applications permet, en regroupant les micro-indicateurs correspondant, et par de simples calculs de moyenne, de créer *des indicateurs* représentatifs de la présence des langues dans l'Internet. Six grandes catégories d'indicateurs ont été ainsi identifiés qui mesurent la quote-part de la langue selon des éléments caractéristiques de l'Internet:

- **Internautes** (ou personnes connectées): qui se rapporte aux locuteurs de chaque langue ayant accès à l'Internet. Un seul micro-indicateur (offert par l'IUT) permet d'informer cette catégorie et servira comme source fondamentale pour le reste des travaux.

- **Usages**: qui se rapporte aux abonnements de ces locuteurs à des applications ou aux moyens de connexion à l'Internet. Onze micro-indicateurs permettent de construire cet indicateur.

- **Trafic**: qui se rapporte au trafic généré par les internautes vers des applications spécifiques qui correspondent à des *thématiques* établies. Trois cent seize micro-indicateurs permettent de construire cet indicateur (le nombre important permettra d'obtenir des statistiques par thématique).

- **Indexes**: qui se rapporte à des classements permettant d'attribuer une note comparative dans les différents éléments constitutifs de la société de l'information. Cinq micro-indicateurs permettent de construire cet indicateur pour le moment, mais cela pourrait être étendu.

- **Contenus**: qui se rapporte aux contenus dans l'Internet pour chaque langue et qui, pour le moment, regroupe principalement des données de la galaxie Wikimédia. Treize micro-indicateurs permettent de construire cet indicateur.

- **Interfaces** et **langues de traduction**: qui se rapporte à la présence des langues dans les interfaces à des applications ou comme langue de traduction. Vingt trois micro-indicateurs permettent de construire cet indicateur.

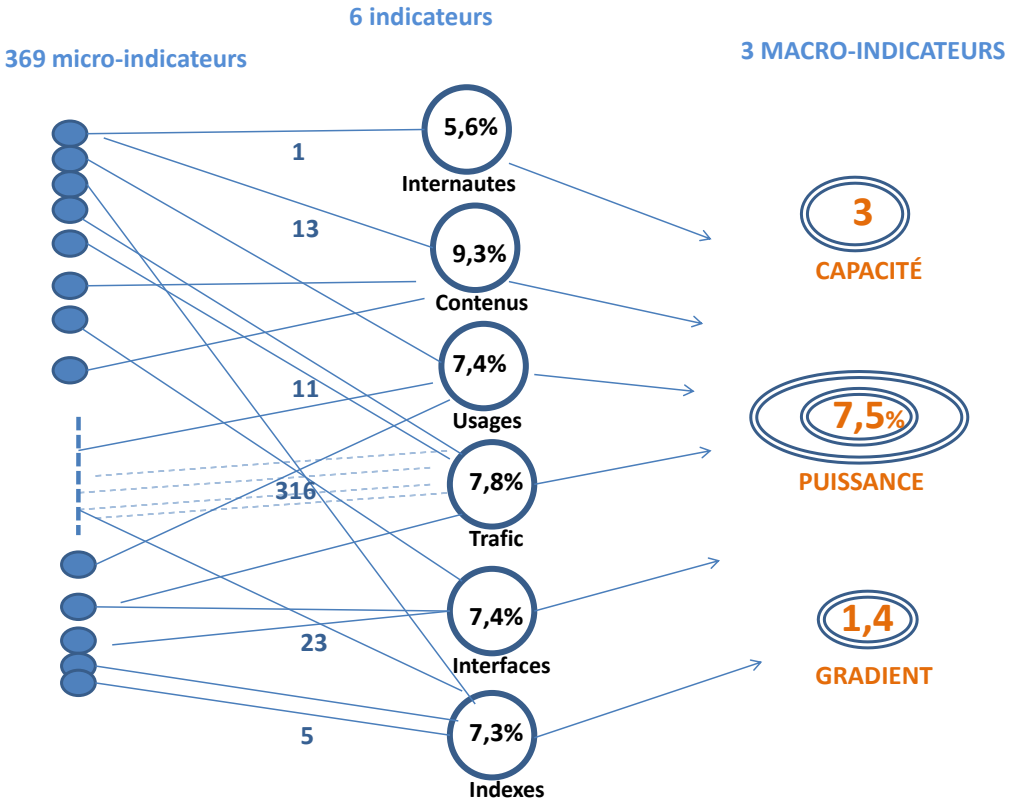
Finalement trois macro-indicateurs de la présence des langues dans l'Internet expriment la synthèse de l'ensemble:

- un indicateur de la **puissance** de la langue dans l'Internet, objet principal de cette étude, lequel mesure la quote-part de la langue dans l'Internet, moyenne des 6 indicateurs précédents;

- un indicateur de la **capacité** de la langue dans l'Internet lequel, en mesurant le rapport entre la puissance et le pourcentage de la population mondiale dans cette langue, indique la capacité propre à cette langue dans l'Internet[16];

- un indicateur du **gradient** de la langue dans l'Internet lequel, en mesurant le rapport entre la puissance et le pourcentage de personnes connectées à l'Internet, indique l'impact ans l'Internet des locuteurs de cette langue connectés à l'Internet.

FIGURE 1 : **Représentation graphique des indicateurs et résultats**




L'ensemble des micro-indicateurs est présenté en *Annexe 1*. Les indicateurs sont détaillés dans un tableau du chapitre *4.5.3 Traitements pour obtenir les résultats*.

## 4.2 Micro-indicateurs par langue ou par pays

Un ensemble de 369 micro-indicateurs constitue l'ensemble des données mesurées pour cette édition. Certains (36) concernent directement les langues dans l'Internet, les autres (333) concernent les pays .

### 4.2.1 Par langue

Les micro-indicateurs par langue concernent les **contenus** (13) et les **interfaces** (23) et sont au nombre de 36 dans cette édition.

Les **contenus** sont exprimés en termes d'unités par langue (par exemple, nombre d'articles Wikipédia par langue) ou en pourcentage mondial (estimations de contenus de W3Techs). Dans le premier cas, la valeur est transformée en pourcentage mondial en divisant par le nombre total. L'indicateur **contenu**

[16] Une capacité supérieure à 1 indique une forte capacité.

est calculé comme la moyenne réduite à 20%[17] des 13 micro-indicateurs relatifs aux contenus (un relatif aux nombre de livres par langue dans Amazon, W3Tech et 11 en relation avec Wikimédia).

Les **interfaces et langues de traduction** sont simplement exprimés par un nombre binaire exprimant l'absence ou la présence de la langue dans l'application. Il y a 23 applications renseignées. L'indicateur **interfaces** indique dans un premier temps le pourcentage de présence de la langue dans l'ensemble des applications mesurées et ensuite cette valeur est transformée en pourcentage mondial par pondération par rapport au pourcentage d'internautes dans chaque langue.

### 4.2.1 Par pays

Plus de 90% des micro-indicateurs de cette édition de l'étude proviennent de sources qui renseignent par pays, les sources numériques pouvant être rapportées de plusieurs manières:

- ✓ en **quantités** (par exemple, le nombre d'internautes par pays);
- ✓ en **pourcentages nationaux** (par exemple, le pourcentage de consultation à Facebook par pays, ou le pourcentage du trafic Internet réalisé par mobile dans chaque pays) ;
- ✓ en **pourcentages mondiaux** (par exemple la répartition du trafic mondial vers Facebook.com par pays);
- ✓ en **notation sur un barème fixé** (c'est le cas des index qui apportent une notation à des pays selon des critères définis comme par exemple, WebIndex[18] qui fournit des index par pays à propos de la société de l'information prenant des valeurs de 0 à 1 ou de 0 à 100, selon le cas);

Pour pouvoir réalisé un travail homogène, les micro-indicateurs qui seront construits à partir de ces sources seront finalement toujours exprimés en termes de **pourcentage mondial** pour obtenir, après la transformation de données par pays en données par langues, un résultat également en termes de pourcentage mondial.

La transformation de données par pays en données par langue se réalise par **pondération des valeurs par rapport au nombre de locuteurs de chaque langue dans chaque pays**.

Dans certains cas la même opération de pondération à partir des pourcentages de locuteurs de langue dans chaque pays permet de ramener les résultats en terme de pourcentage mondial en une seule opération

Les sources peuvent exprimer des valeurs pour le présent (2016) ou plus souvent pour une date plus ancienne, voire même parfois pour des projections pour le futur.

La transformation de ces différents types de données numériques (quantité, pourcentage de différents types, indicateurs normalisés) en des données faisant sens pour les langues, après pondération avec les données démo-linguistiques par pays, demande une grande attention au type de données et surtout à la signification de ces données.

[17] La moyenne réduite est la moyenne du vecteur des valeurs après élimination des valeurs extrêmes les plus hautes et les plus basses. Dans ce cas la fenêtre d'exclusion a été fixée à 20% ce qui veut dire que les 10% des notes les plus hautes et les plus basses ne participent pas de la moyenne, ce qui protège raisonnablement des biais. Cette méthode permet en particulier de réduire, pour le macro-indicateur trafic, l'influence du choix d'applications favorisant certaines langues (voir 7.4.7 *Limitations/biais liés au degré de localité des sources*).
[18] http://thewebindex.org/

D'autre part, les sources de données par pays disponibles ne couvrent que rarement l'ensemble des pays de la planète[19] et très souvent, seulement une minorité. Ainsi cette transformation de données par pays en données par langues à partir d'un nombre limité de pays exige, pour éviter l'écueil de négliger les langues des pays non renseignés, des techniques d'extrapolation pour complémenter, de manière raisonnable à partir des données existantes, les données pour les pays non renseignées par la source (voir *le chapitre 4.5.1 Extrapolations*).

## 4.3 Sources d'information

Pour réunir cet ensemble de données, des dizaines de sources d'information différentes ont été répertoriées (statistiques, sondages, évaluations, indexes, bases de données, etc.) puis évaluées ensuite en fonction de leur pertinence pour l'étude, de la confiance que l'on pouvait leur accorder et également de leur adéquation avec la méthode. Certaines sources offrant des résultats pour un nombre trop limité de pays n'ont pas été retenues et d'autres pouvant exprimer des pourcentages par pays qui perdent leur sens quand ils sont ramenés aux langues ont été écartées.

Il est aisé de trouver les statistiques par pays concernant l'accès à l'Internet et aux nouvelles technologies (grâce notamment à l'UIT, aux Nations Unis, à la Banque mondiale et à certains autres sites provenant d'ONG et d'organismes publiques ou parapubliques). Par contre, il est bien plus difficile de trouver des sources offrant des résultats par langue, en dehors de l'univers *Wikimédia* et de l'univers du *libre* qui sont marqués par un authentique souci de transparence tant pour les données que pour les méthodologies qui les ont produites.

Les acteurs économiques propriétaires des réseaux sociaux, des moteurs de recherche, des applications mobiles, etc. font en général un secret de leurs statistiques. Les seules qui sont capables d'offrir des données sur les applications (avec une énorme concentration sur les réseaux sociaux et sur tout ce qui a un rapport avec le marketing dans le monde numérique) sont des compagnies de marketing de l'Internet qui demandent souvent de fortes sommes pour divulguer leur données. Ces entreprises ont souvent des relations privilégiées avec les grandes applications dont elles peuvent être des amplificateurs d'audience ou ont parfois les moyens financiers pour développer des méthodes puissantes quoique souvent approximatives (comme c'est le cas d'Alexa[20], maintenant une succursale de Amazon, ou de W3Techs[21]).

Pour trouver des sources, il y a deux options à combiner:
- soit ramasser quelques miettes gratuites, le plus souvent des filtrages d'information partielle réalisés par une des ces compagnies comme un moyen d'appâter les clients pour payer le gros des informations;
- soit payer pour obtenir les données.

La première approche a été appliquée systématiquement en déployant des moyens de recherche intense; cependant il faut savoir que les méthodologies ne sont en général pas transparentes et que ces

[19]Les seules sources disponibles pour tous les pays de la planète concernent le nombre d'internautes, le nombre de lignes de téléphonie fixe, les comptes de haut-débit fixe, les comptes haut-débit mobile, le nombre de téléchargements OpenOffice, le nombre d'utilisateurs Facebook selon Owloo et l'usage de serveurs par site Web.
[20] Alexa (http://alexa.com) est une entreprise qui fournit des statistiques et des classements à propos des sites Internet, sur la base d'estimation de trafic.
[21] Un portail de statistiques technologiques. https://w3techs.com/

données sont rarement pérennes. Cette étude prévoyait un budget limité pour des sources payantes. Alexa et Statista[22] ont été finalement sélectionnées comme sources payantes en raison de leur meilleur rapport qualité/prix[23] et ont nourri les statistiques souhaitées pour la plupart des applications ou espaces. Cela a permis de réduire, par souci d'homogénéité, le nombre final de sources retenues à une quinzaine. Quant à Alexa, ce choix permet de couvrir très large[24] mais il faut savoir que ce sont des données de trafic par pays qui présentent des biais (voir le chapitre *7.4.6 Biais potentiels de Alexa.com et W3Techs*).

La liste complète des sources retenues est détaillée en *Annexe II. Sources retenues.*

## 4.4 Données démo-linguistiques

### 4.4.1 Le cas des langues maternelles (L1)

Pour les langues maternelles, l'étude s'appuie sur un apport considérable de Daniel Prado qui a repris les données de la source du projet Joshua[25] qui offre la répartition des plus de 7500 langues du monde par pays et l'a, d'une part, corrigé quand nécessaire à partir de diverses sources, et d'autre part, adapté à la division par pays choisie pour cette étude.

L'onglet LOC1 du fichier Excel de travail répond donc à cette définition, pour l'ensemble des langues retenues et l'ensemble des pays sélectionnés:

$LOC_1(i,j)$ = Le nombre de locuteurs L1 pour la langue i dans le pays j.

La source portant sur 7500 langues, il restait à faire un choix sur la liste des langues qui allaient être sélectionnées pour cette étude portant sur la présence des langues dans l'Internet. Aujourd'hui, l'estimation du nombre de langues qui ont une présence dans l'Internet est de l'ordre de 500. Une possibilité pouvait être de s'approcher autant que possible de cette liste; une autre possibilité était de sélectionner les langues pour lesquelles Wikipédia offre des statistiques. Après plusieurs essais le choix s'est finalement porté sur la **liste des 140 langues ayant plus de 5 millions de locuteurs**. les données concernant les autres langues ont été cumulées par pays dans une catégorie appelée "reste des langues".

La raison de ce choix est que l'hypothèse simplificatrice sur laquelle repose la méthode de pondération, qui conduit à transformer des données par pays en données par langue, ne permet pas d'apporter une confiance suffisante dans les résultats pour les langues dont le nombre de locuteurs est faible. L'objectif premier de l'étude étant d'évaluer la puissance du français dans l'Internet en comparaison avec les autres langues les plus puissantes, ce choix permet d'y répondre avec une grande marge de sécurité tout en évitant de produire des résultats (comme par exemple pour les langues de France - voir 7.2.5 *Le cas des langues de France*) qui ne présenteraient pas un critère de confiance suffisant.

[22] Statista (http://statista.com) est un généraliste des statistiques qui possède des équipes par thème disposées à aider les clients.
[23] Un prix mensuel de dix dollars des États-Unis pour utiliser pleinement les statistiques de trafic de Alexa et un abonnement annuel de 2,100 dollars des États-Unis (un prix spécial négocié) pour l'utilisation des services de Statista.
[24] Plus de 80% de nos micro-indicateurs sont mesurés a l'aide de ce service.
[25] Un portail de données démo-linguistiques. https://joshuaproject.net/

### 4.4.2 Le cas des langues secondes (L2)

La méthode idéale pour traiter le cas des L2 serait évidemment de faire de même que pour les L1 et de produire une matrice telle que:

$LOC_2(i,j)$ = Le nombre de locuteurs L1+L2 de la langue i dans le pays j.

Il suffirait ensuite d'appliquer sans aucun changement tout le reste de la méthode pour obtenir les résultats attendus avec une précision très appréciable. Malheureusement, ce travail n'existe pas et l'existence de fortes divergences entre les experts sur les valeurs en question rendent la réalisation d'une telle matrice très difficile en l'état actuel des connaissances sur le sujet. Il est donc proposé une autre approche.

Le principe de la méthode choisie est extrêmement simple et consiste, pour chaque langue, à obtenir un nombre qui représente l'accroissement à appliquer aux quantités L1 pour obtenir la valeur L1+L2. Ce nombre sera égal à 1 pour la langue i si L2(i) = 0. L'auteur de ce rapport a utilisé de manière homogène pour sa méthode d'introduction des L2 les propositions de Ethnologue tout en conservant intactes les données établies par ailleurs pour L1 (en d'autres termes si Ethnologue fournit un total L1+L2 il est appliqué sur le L1 existant). Ainsi les langues et valeurs concernées sont indiquées dans le tableau suivant, par ordre alphabétique des codes ISO.

TABLE 1 : **Valeurs retenues pour les L2**

| Code ISO | Langue | L2 | L1 | L1+L2/L1 |
|---|---|---|---|---|
| Awa | Awadhi | **45 400** | 11 821 647 | 1,0038 |
| Ben | Bengali | **19 202 880** | 248 094 781 | 1,0774 |
| Bho | Bhodjpouri | **160 000** | 48 656 324 | 1,0033 |
| Ces | Tchèque | **2 540 000** | 10 121 000 | 1,251 |
| Deu | Allemand | **52 689 000** | 69 419 000 | 1,759 |
| Ell | Grec moderne | **57 000** | 11 455 800 | 1,005 |
| Eng | Anglais | **611 563 010** | 344 179 157 | 2,7769 |
| Fas | Persan | **18 434 114** | 43 495 306 | 1,4238 |
| **Fra** | **Français** | **153 485 770** | **76 660 700** | **3,0021** |
| Hau | Haoussa | **19 500 000** | 45 494 090 | 1,4286 |
| Hin | Hindi | **121 230 000** | 374 692 040 | 1,3235 |
| Ita | Italien | **3 085 000** | 41 769 700 | 1,0739 |
| Jpn | Japonais | **11 500** | 125 376 800 | 1,0001 |
| Kan | Kannada | **9 000 000** | 43 729 877 | 1,2058 |
| Khm | Khmer | **1 000 000** | 16 030 000 | 1,0624 |
| Mar | Marathi | **3 000 000** | 82 230 980 | 1,0365 |
| Msa | Malais | **175 200 000** | 53 798 640 | 4,2566 |
| mya | Birman | **10 000 000** | 31 704 180 | 1,3154 |
| Orm | Oromo | **170 000** | 33 129 700 | 1,0051 |
| Pol | Polonais | **454 000** | 40 479 300 | 1,0112 |
| Por | Portugais | **11 180 000** | 217 192 370 | 1,0515 |

| | | | | |
|---|---|---|---|---|
| Rus | Russe | **113 273 820** | 130 013 300 | 1,8712 |
| Sin | Singhalais | **2 000 000** | 11 838 680 | 1,1689 |
| Som | Somali | **95 600** | 19 801 000 | 1,0048 |
| Spa | Espagnol | **91 308 400** | 434 569 400 | 1,2101 |
| Tam | Tamoul | **8 000 000** | 74 551 992 | 1,1073 |
| Tel | Télougou | **5 000 000** | 87 584 764 | 1,0571 |
| Tha | Thaï | **40 000 000** | 28 549 500 | 2,4011 |
| Tur | Turc | **380 300** | 64 175 300 | 1,0059 |
| Urd | Ourdou | **94 045 800** | 99 137 906 | 1,9486 |
| Yor | Yoruba | **2 000 000** | 38 104 300 | 1,0525 |
| Zha | Zhuang | **20 000** | 16 305 000 | 1,0012 |
| Zul | Zoulou | **15 700 000** | 12 159 200 | 2,2912 |
| **Zzz** | **RESTE** | **236 200 000** | **891 034 646** | **1,2651** |
| | **TOTAL** | **1 824 031 594** | **7 296 012 569** | **1,2500** |

Notes:
- Le malais inclut l'indonésien
- Les valeurs L1 utilisées pour le chinois étant supérieures aux valeurs L1+L2 indiquées par Ethnologue L2 a été laissé à 0 pour le chinois.
- Par prudence une valeur de 1,265 a été inscrite pour le reste des langues: l'effet, marginal sur les calculs, est d'éviter de trop forcer les notes des 140 langues sélectionnées vers le haut. La valeur a été finalement choisie pour permettre un arrondi à 25% de l'effet du multilinguisme.

Le taux mondial d'augmentation résultant (1,25 avec les hypothèses choisies) est le résultat de l'opération de pondération:

$$Tg = \sum_{j=1}^{j=P} L1(j) \times T(j)$$

où P est le nombre total de langues, L1(j) le nombre de personnes de langue maternelle j et T(j) le taux d'accroissement pour la langue j.

Note: Il est remarquable que le français est, après le malais, la langue qui a le taux le plus haut parmi l'ensemble des langues.

Pour passer les résultats en L1 en résultats en L1+L2 le taux d'accroissement est utilisé et les résultats en L1 sont ensuite pondérés par rapport à la distribution des taux d'accroissements T.

$$R_{L1+L2}(i) = Tg \times R_{L1}(i) \; / \; \sum_{j=1}^{j=L} R_{L1}(j) \times T(j)$$

La division opère la pondération pour obtenir des résultats cadrés sur 100% et le produit par Tg est seulement là pour normaliser les résultats sur un pourcentage égal à Tg.

Cette méthode est évidement moins précise qu'une solution qui aurait pu différencier les taux de connectivité à l; Internet pour chaque pays; ses limitations et biais possibles, ainsi qu'une variante peu complexe de la méthode pour la rendre plus précise, sont discutés au chapitre *7.2.2 Le cas des L2*.

Il est important de noter que l'OIF peut proposer le cas échéant son propre schéma pour les L2 et tant la mise à jour des entrées que la refonte des calculs et des résultats sont instantanés.

### 4.5 Traitements réalisés

Les indicateurs exprimés directement en termes de langue (contenus et interfaces) ne présentant aucune difficulté particulière, l'attention va être portée sur les indicateurs exprimés par pays.

Le principe du calcul de base est celui du produit matriciel entre, d'un côté, une matrice LOC ayant en abscisse (lignes) les différentes langues sélectionnées et en ordonnée (colonnes) les différents pays considérés (la dernière colonne cumulant les valeurs pour l'ensemble des pays non considérés, la dernière ligne cumulant les valeurs pour le reste des langues non traitées) et, de l'autre côté, le vecteur RP comprenant l'ensemble de valeurs présentées par les sources pour chacun des pays considérés pour l'application ou l'espace étudié.

LOC(i,j) = Le nombre de locuteurs de la langue i dans le pays j.

$RP_n(j)$ = La valeur mesurée pour le micro-indicateur n dans le pays j.

$RL_n(i) = \sum_{j=1}^{j=P} LOC(i,j) \times RP_n(j)$ où P est le nombre total de pays.

Le produit matriciel $RL_n$ = LOC +.x $RP_n$, en notation APL[26], ou =SommeProduit(LOC;$RP_n$) en notation Excel, est une opération de pondération qui produit un nouveau vecteur cette fois-ci de la taille du nombre de langues pour lequel $RL_n(i)$ = la valeur du micro-indicateur n déduite pour la langue i à partir des données démo-linguistiques (LOC) et des valeurs mesurées par pays pour le micro-indicateur n ($RP_n$). Le cœur de la méthode est en fait une pondération des valeurs mesurées pour ce micro-indicateur dans chaque pays par rapport à la présence de la langue j dans chacun des pays[27]. La validité de cette pondération et ses biais potentiels sont discutés dans le chapitre 7.*1 Limitations propres à la méthode*.

Pour bien comprendre:

- les totaux verticaux de LOC représentent le nombre total d'habitants pour le pays j;
- les totaux horizontaux de LOC représentent le nombre total de locuteurs de la langue i dans le monde;
- le total du vecteur $RP_n$ représente le total mondial pour le micro-indicateur dans le monde (qui peut être une quantité mondiale dans le cas d'une quantité par pays[28]);
- les totaux de $LP_n$ sont les mêmes que ceux de $RP_n$ mais cette fois-ci la répartition est faite directement par langue et non plus par pays.

[26] APL, "A Programming Language", qui est à la fois un formalisme mathématique et son implémentation sous forme de langage de programmation, a été conçu par Kenneth. Iverson. Pour plus de détails voir https://fr.Wikipédia.org/wiki/APL_(langage)

[27] Le principe de pondération peut se comprendre intuitivement en prenant l'analogie courante du système de notation d'un étudiant en examen avec des coefficients différents pour chaque matière: la *moyenne pondérée*, qui déterminera le succès final, s'obtient en faisant la somme des notes dans chaque matière multipliées par le coefficient correspondant à cette matière et en divisant le total par la somme des coefficients.

[28] Ou une valeur de 100% s'il s'agit d'une répartition complète en pourcentages par pays (par exemple le pourcentage de ligne fixe téléphoniques par pays par rapport au pourcentage mondial).

L'inconvénient majeur de cette méthode de calcul est qu'elle ne permet pas d'obtenir des résultats crédibles si tous les pays ne sont pas renseignés car le calcul matriciel transforme cette absence en valeur nulle ce qui pénalise de manière inacceptable les résultats pour les langues fortement présentes dans les pays pour lesquels la source n'indique aucune information (a fortiori, dans le cas de cette étude, si les pays francophones majeurs sont absents).

### 4.5.1 Extrapolations

La manière de surmonter cet obstacle est de remplir de la meilleure manière possible les valeurs absentes des pays non renseignés. C'est faisable de manière raisonnable, dans la plupart des cas, en extrapolant les informations manquantes à partir des informations existantes et d'autres informations connues par ailleurs. Il ne s'agit pas de rechercher une précision absolue mais plutôt une méthode simple dont les écarts par rapport à la réalité seront réels mais de peu d'impact sur les résultats des traitements statistiques.

Deux méthodes différentes ont été adoptées pour résoudre l'ensemble des cas:

**a) L'extrapolation au prorata des pourcentages de personnes connectées par pays.**
Cette méthode s'appliquera quand il est raisonnable de considérer que la quantification des valeurs du micro-indicateur considéré est naturellement proportionnelle au pourcentage mondial que représente le pays en termes de personnes connectées (c'est la cas par exemple des trafic vers des sites web). Selon que les données de la source s'expriment en pourcentage ou en quantités il faudra extrapoler (ou pas) en premier lieu le total mondial de la quantité en question. Le reste extrapolé des quantités non renseignées ou le pourcentage restant, selon le cas, sera répartit au prorata des poids respectifs des pays restants en termes d'internautes.

**b) Méthode des quartiles**
Dans cette méthode, il est attribué aux pays non renseignés un des quartiles des données des pays renseignés en fonction du pourcentage de personnes connectées. Après plusieurs essais, il est apparu opportun de déterminer ainsi l'attribution des quartiles:

- Si le pays a moins de 15% de sa population connectée : note la plus basse
- Si le pays a plus de 15% mais moins de 35% de sa population connectée : premier quartile
- Si le pays a plus de 35% mais moins de 65% de sa population connectée : médiane
- Si le pays a plus de 65% mais moins de 85% de sa population connectée : troisième quartile
- Si le pays a plus de 85% de sa population connectée : note la plus haute.

En règle générale les micro-indicateurs pour lesquels aucune méthode d'extrapolation n'apparait de manière évidente sont les mêmes pour lesquels le sens de la transformation pays vers langue n'apparait pas non plus clairement et ils seront donc écartés.

La variété de significations possibles pour les valeurs de $RP_n$ montre la nécessité de distinguer les différents types de données quantitatives qui vont être manipulées et dans la pratique d'effectuer les techniques d'extrapolations de manière conforme à la nature du type.

### 4.5.2 Différents types de données

L'expérience a conduit à distinguer les types de données suivant:

**NB: Quantité par pays.** Le cas le plus simple comme par exemple dans le cas de l'espace "lignes téléphoniques fixes". Dans ce cas le calcul fera passer du nombre de lignes téléphoniques fixes par pays au nombres de lignes téléphoniques par langue (qu'il faut comprendre comme nombre de locuteurs de chaque langue ayant une ligne téléphonique fixe). Lorsque l'extrapolation s'avère nécessaire elle sera faite sur le principe du prorata.

**PN: Pourcentage dans le pays.** Par exemple, pourcentage de personnes ayant un téléphone mobile, par pays. Dans ce cas le calcul fera passer du pourcentage de personnes ayant un mobile par pays au pourcentage de locuteurs de chaque langue ayant un mobile. Parfois les pourcentages sont indiqués dans les sources comme un nombre de 0 à 1, autrefois de 0 à 100; les calculs doivent en tenir compte. Lorsque l'extrapolation s'avère nécessaire elle sera faite avec la méthode des quartiles.

**IX: Index.** Il s'agit d'une notation normalisée attribuée à chaque pays pour un critère donné. Par exemple la qualité des services de gouvernement électronique par pays, notée de 0 à 100 avec 100 comme la meilleure note (parfois noté de 0 à 1). Dans ce cas, le calcul aboutit à une valeur qui représenterait une sorte de pondération des services ramenée à la langue, en fonction des notes respectives des différents pays qui ont des locuteurs de cette langue. Ainsi si un groupe linguistique se répartissait à part égale entre 2 pays, avec une note maximum de 100 pour l'un d'entre eux et une note très faible de 50 pour l'autre la langue obtiendrait comme résultat une note moyenne de 75 qui qualifierait la qualité moyenne des services de gouvernement électronique pour les locuteurs de cette langue dans le monde. Lorsque l'extrapolation s'avère nécessaire elle sera faite avec la méthode des quartiles.

Dans ce cas précis, quoique le calcul fasse sens, le résultat obtenu peut surprendre et s'avérer le symptôme de la mauvaise prise en compte des langues dans ces classements internationaux. Ainsi, la France ayant la note maximale de 100 pour l'index des services en ligne de gouvernement électronique[29], mais le nombre de locuteurs du français se trouvant dans des pays beaucoup moins notés, la note du français tombera à 94 (L1) après pondération, ce qui est logique et cohérent. Par contre, les calculs réalisés pour un échantillon de langues incluant les langues locales de France indiquent pour des langues locales comme le corse ou les créoles de la Guadeloupe et la Martinique, une meilleure note que celle du français! La pondération d'une langue qui est parlée en très grande proportion en France explique ce résultat qui peut sembler paradoxal quand on sait que les services de e.gouvernement ne sont pas encore offerts en créole ou en corse (ce qui serait alors une bonne raison pour justifier une meilleure note). Cependant, si on prend en compte que la note de 100 n'est pas une note absolue mais simplement la meilleure note (ou, si on considère l'aspect linguistique, la moins mauvaise note!) et que plus de 90% des locuteurs de certaines langues locales qui sont résidents dans le pays qui a la meilleur note[30], la France, ce qui est en cause ce n'est pas la méthode mais l'index. D'une certaine manière il est bon que cette méthode permette de souligner les carences de ces indexes. En tout état de cause, la méthode de pondération ne permet pas de toute façon d'obtenir des résultats utilisables pour les langues de France (voir *7.2.5 Le cas des langues de France*) et en règle général les hypothèses simplificatrices de la méthode de pondération pour passer de données pays en données langue interdisent d'appliquer la méthode avec confiance pour des langues dont les locuteurs ne se comptent pas en millions (voir *7.1 Limitations propres à la méthode*).

[29] Selon les Nations Unis (https://publicadministration.un.org/egovkb/en-us/reports/un-e-government-survey-2016), donnée rapportée par WebIndex pour 2013 (http://thewebindex.org/wp-content/uploads/2012/05/Web_Index-Data_2014.xls)

[30] Dans le cas du créole de Guadeloupe présent dans l'échantillon de langues utilisé cette valeur est de 100%.

**PM: Pourcentages mondiaux.** Les données expriment, pour chaque pays, le pourcentage mondial concernant le micro-indicateur. Ainsi un micro-indicateur offre le pourcentage par pays du marché mondial du commerce électronique. Le passage à la langue va exprimer le même pourcentage ramené aux langues (si on accepte l'hypothèse très simplificatrice dans cas que le marché national se répartit de manière homogène entre les langues). Lorsque l'extrapolation s'avère nécessaire elle sera faite par la méthode des quartiles. Le cas le plus intéressant est la construction, à partir de l'utilisation des données de l'UIT concernant les personnes connectées à l'Internet, d'une indication relativement fiable du pourcentage de personnes connectées par rapport au total de locuteurs de chaque langue. Ce résultat est un bénéfice collatéral très important de la méthode quand on sait que les seules données disponibles depuis plusieurs années sont celles de InternetWorldStat[31] qui se réduisent aux seules 10 langues considérées comme les plus présentes. Voir *5.3.1 Comparaison des résultats avec ceux de InternetWorldStats*.

**PN: Pourcentage national.** Certaines sources indiquent des pourcentages nationaux comme par exemple le nombre d'abonnements mobiles pour 100 habitants (ce n'est pas vraiment un pourcentage car sa valeur est supérieure à 100 dans certains pays) et qui se traduit, après transformation, en pourcentage de locuteurs de chaque langue ayant un abonnement mobile. Certains cas de pourcentages nationaux s'avèrent réfractaires à la transformation en pourcentages par langue (voir *7.4.4 La question du sens de la transformation pays > langues*).

**PMI : Pourcentage mondial relatif aux internautes.** C'est le type de donnée le plus utilisé dans cette étude, grâce à la versatilité du service Alexa.com, qui permet de mesurer pour certains pays le pourcentage du trafic mondial vers un site web donné. Lorsque l'extrapolation s'avère nécessaire elle sera faite sur le principe du prorata.

Ce cas, étant donné son poids dans les résultats et le fait que le nombre de pays renseignés est généralement faible est celui qui a bénéficié le plus de la technique d'extrapolation. Il n'en reste pas moins que la méthode de Alexa présente des biais important qu'il faut analyser (voir *7.4.6 Biais potentiels de Alexa.com et W3Techs*).

### 4.5.3 Traitements pour obtenir les indicateurs

En dehors de **l'indicateur de l'UIT qui offre les pourcentages de personnes connectées à l'Internet, par pays, et qui est une pièce fondamentale de cette étude**, peut être aussi d'autres indicateurs concernant les technologies, et, à la limite, de certains indicateurs statistiques de Wikimédia[32], aucun indicateur pris individuellement ne reflète une vérité valide pour l'ensemble des langues. En effet, l'idéal d'un degré de mondialisation absolu, qui permettrait de considérer que les comportements des usages sont les mêmes face à cet indicateur, indépendamment du pays ou de la langue, sera difficilement vérifié. Seul un traitement statistique sur un ensemble important de micro-indicateurs sélectionnés avec des critères le plus prudemment divers et mondialisés possible, permet, en l'absence d'une mesure fiable des contenus par langue[33], de quantifier la présence des langues dans l'Internet et cela à condition de comprendre les déformations potentielles dans les résultats que peuvent apporter les instruments de mesures et les sources. Le grand nombre peut permettre aussi d'appréhender, par différentiation, les thématiques où une langue donnée montre plus ou moins de

[31] http://www.internetworldstats.com/stats7.htm

[32] Mais il ne faut pas perdre de vue que même si Wikipédia représente l'application la plus mondialisée de l'Internet, elle n'est pas très présente dans certain pays importants de l'Internet comme la Chine ou l'Inde et qu'elle ne permet de renseigner qu'au maximum 300 langues.

[33] L'étude établit que W3Techs ne peut pas être considéré comme une mesure fiable, très loin de là.

présence; dans l''etat actuel de l'étude, seul l'indicateur de trafic qui comprend 316 micro-indicateurs permet de se risquer à l'analyse thématique.

L'ensemble des indicateurs de la présence des langues dans l'Internet qui ont été créées sont présentées dans le tableau suivant. Tous sont exprimés en termes de quote-part mondiale sur la base de la population totale de locuteurs (population mondiale multipliée par le facteur multilinguisme, dans la version actuelle 25% plus grande que la population mondiale).

**TABLE 2: Description des indicateurs**

| MACRO-INDICATEUR | DÉFINITION | TÉCHNIQUE | FIABILITÉ |
|---|---|---|---|
| **A: INTERNAUTES** | Mono indicateur dérivé du % mondial de personnes connectées par pays de l'UIT. | Transformation pays -> langues sans extrapolation | Très forte<br>Biais très marginaux |
| **B: USAGES** | Regroupe 11 micro-indicateurs<br>- lignes fixes téléphoniques<br>- marché e.commerce<br>- téléchargement OpenOffice<br>- utilisateurs réseaux sociaux + projection 2021<br>- utilisateurs de Netflix + projection 2020<br>- utilisateurs Facebook, Twitter, YouTube, Linkedin | Transformation pays-> langue avec extrapolation Moyenne des micro-indicateurs | Forte.<br>Biais faibles.<br>Mais le nombre de micro-indicateurs aurait besoin d'être étendu pour donner plus de sens à la moyenne. |
| **C: TRAFIC** | Mesure du trafic vers une sélection de 316 sites web. | Transformation pays-> langue avec extrapolation Moyenne réduite à 20% des micro-indicateurs | Relativement bonne<br>Mais fort biais anglophone confirmé par des comparaisons entre trafic et nombre d'abonnés par pays. |
| **D: SOCIETE DE L'INFORMATION** | Regroupe 5 indexes de WebIndex qui qualifient selon les critères suivants:<br>- E.gouvernement<br>- Accès universel<br>- e.participation<br>- général<br>- infrastructure | Transformation pays-> langues avec extrapolation par la méthode des quartiles.<br>Puis transformation en pourcentage mondiale par pondération avec données UIT<br>Moyenne des micro-indicateurs | Bonne (il s'agit de données subjectives quantifiées par un organisme compétent).<br>Cette catégorie mérite d'être étendue. |
| **E:CONTENUS (Wikipédia et livres)** | Regroupe 13 micro-indicateurs<br>- Nombre de livres chez Amazon<br>- Mesures contenus par langue de W3Techs<br>- 11 indicateurs langue de Wikimédia: articles, utilisateurs ou éditeurs pour Wikipédia, wikilivre, wikiquote, wikisource, wikiversité, wiktionnnaire. | Reprise directe des données par langue.<br>Moyenne réduite à 20% des micro-indicateurs | Très forte pour Wikimédia et Amazon.<br>Mais très faible présence de certaines langues asiatiques majeures.<br>Le nombre de micro-indicateurs aurait besoin d'être étendu pour donner plus de sens à la moyenne. |
| **F:INTERFACES (et langues de traduction)** | Regroupe 23 micro-indicateurs binaires (présence ou non) | % de présence sur l'ensemble des 23 micro-indicateurs.<br>% mondial par pondération avec données UIT. | Parfaite. |

TABLE 3: **Description des macro-indicateurs**

| ***MACRO-INDICATEURS*** | **DÉFINITION** | **TÉCHNIQUE** |
|---|---|---|
| ***PUISSANCE*** | Mesure la quote-part de la langue dans l'Internet comme un % mondial | Moyenne des 6 indicateurs |
| ***CAPACITÉ*** | Mesure la capacité d'une langue dans l'Internet indépendamment du nombre de locuteurs | Rapport entre puissance et % mondial de locuteurs |
| ***GRADIENT*** | Mesure l'impact des locuteurs connectés indépendamment de leur nombre | Rapport entre puissance et % de locuteurs connectées |

Les 3 macro-indicateurs sont liés par les équations suivantes:

**Puissance** de la langue dans l'Internet = **Capacité** de la langue dans l'Internet x Population mondiale

**Puissance** de la langue dans l'Internet = **Gradient** de la langue dans l'Internet x Population connectée

La puissance et la population mondiale sont exprimées en termes de quote-part mondiale et la capacité et le gradient sont des quantités sans dimension normées à la valeur 1[34].

Des calculs générés automatiquement dans le fichier Excel de travail[35] permettent d'obtenir une valeur de quote-part associée à chaque langue pour chaque indicateur. La valeur moyenne entres les six indicateurs est considérée comme la valeur la mieux représentative de la puissance d'une langue dans l'Internet, probablement la meilleure approximation possible à la quote-part de contenus associés à cette langue en l'absence d'une mesure fiable. L'onglet SYNTHESE de la feuille de calcul se met à jour automatiquement après chaque changement et offre l'ensemble des résultats quantitatifs de cette étude ce qui permet de réaliser des simulations très facilement.

Il faut noter que:

- ✓ A- La donnée de l'UIT à propos des **internautes** à l'Internet est traitée, seule, comme un indicateur essentiel et "pèse" donc très fort dans les résultats, directement, dans la moyenne établie, qui lui donne un poids aussi important que les plus de 300 micro-indicateurs de trafic fournies par Alexa, et, indirectement, dans les techniques d'extrapolation et dans les techniques de pondération qui affectent transversalement l'ensemble des calculs. Tout l'édifice des traitements repose sur cette donnée considérée comme très fiable.

- ✓ C- Pour les données de **trafic** obtenues par Alexa.com, la force réside dans la quantité de sites utilisés quoique les résultats comporte un fort écart-type causé par la mesure de sites très marqués sur une population linguistique donnée (voir chapitre *7.4.7 Limitations/biais liés au degré de localité des sources*). La technique de moyenne réduite permet d'écarter les valeurs extrêmes et donne plus de confiance dans le résultat. D'autre part, ces données présentent un

[34] Une note proche de 1 est normale; une note très supérieure à 1 exprime un bon résultat ; une note très inférieure à 1 exprime un mauvais résultat.
[35] Voir dans les onglets RES1 et RES2 les colonnes marquées en rouge qui séparent les données par pays des données par langue (colonnes MC à MG).

fort biais en faveur de l'anglais et au détriment surtout des langues asiatiques (mais aussi peut être du portugais) qui est discuté dans le chapitre *7.4.6 Biais potentiels de Alexa.com et W3Techs.*

- ✓ D- Les **indexes** ouvrent une perspective très intéressante sur la mesure d'éléments de contexte comme le sont les indexes-pays créés par des organismes internationaux ou des ONGs. Pour le moment seul WebIndex[36], un travail de la Fondation Word Wide Web qui maintient une série d'indexes qui s'appuient sur les travaux sérieux réalisés par les services des Nations Unies, la Banque Mondiale et d'autres institutions. Les indexes retenus dans cette étude incluent une mesure par pays des services de gouvernement électronique, de la participation citoyenne à travers l'Internet, du degré d'accomplissement de l'accès universel et des infrastructures pour la société de l'information, ainsi qu'un index macro qui intègre l'ensemble des facteurs. Il est importante de savoir que WebIndex propose d'autres indicateurs au delà des cinq qui ont été utilisés[37] et que d'autres organismes offrent déjà ou vont offrir dans le futur des indexes qu'il serait très utiles d'intégrer dans une prochaine étape[38]. Le français apparait en bonne position dans les indexes analysés, confortant sa place de langue importante de l'Internet et de la société de l'information et améliorant encore sa place de quatrième langue de l'Internet.

- ✓ E: Cet indicateur est qualifié comme **contenus,** il est cependant très centré sur la galaxie Wikimédia laquelle offre l'énorme avantage d'offrir des statistiques détaillées et crédibles pour la présence des langues dans ses différentes activités. La complémentation par W3Techs (peu crédible) et les livres chez Amazon, élargit à peine la perspective et cet indicateur. qui s'adresse au secret le mieux gardé dans l'Internet, la répartition des contenus par langue, mériterait d'être élargit... à condition de trouver des sources crédibles, ce qui est devenu une gageure. Il faut noter cependant un fort biais du au fait que les grandes langues asiatiques sont très faiblement représentées dans Wikipédia[39]. Il est important de comprendre également que par sa nature (indicateur direct de langues) cet indicateur va pénaliser totalement (avec des notes de zéro) les langues en dehors de l'univers Wikipédia, lequel, quoiqu'étant celui qui se préoccupe du plus grand nombre de langues existantes (près de 300) ne traite que 300/7500 = 0,4% de l'univers linguistique complet.

[36] http://thewebindex.org/

[37] WebIndex propose également un index général où la France est 11ième, un index "contenus de relevance" pour lequel la France est 3ième, un index "libertés et ouverture" pour lequel la France est 18ième, un index "empowerment" pour lequel la France est 7ième, un index "accès et prix" pour lequel la France est 16ième, un index "éducation et conscience" (awareness) pour lequel la France est 25ième, un index "gratuit et ouvert", pour lequel la France est 18ième, un index "contenus et utilisation" pour lequel la France est 3ième (le Canada et la Belgique sont respectivement 10 et 12ième), in index économique pour lequel la France est 8ième, un index politique, pour lequel la France est 7ième, un index "social et environnement", pour lequel la France est 7ième et la Belgique et le Canada, respectivement 9ième et 11ième.

[38] The Economist vient de publier un rapport avec des macro-indicateurs pour mesurer le "degré d'inclusion" de l'Internet (https://theinclusiveinternet.eiu.com). l'ONG Open Knowledge publie un index pour mesurer les données ouvertes (http://index.okfn.org/). L'UNESCO a l'intention de construire des indicateurs pour mesurer le degré d"universalité de l'Internet dans les pays (http://www.unesco.org/new/en/media-services/single-view/news/unesco_call_for_proposals_defining_internet_universality_in/).

[39] A titre d'exemple, Wikpedia en chinois aurait quelques 700,000 articles alors qu'il existe deux options encyclopédiques chinoises (Hudong et Baidu Baike) qui représentent ensemble près de 12 millions d'articles (source: https://fr.Wikipédia.org/wiki/Wikipédia_en_chinois).

- ✓ F: La pénalisation des langues non ou faiblement connectées (mais également des langues bien connectées mais non prise en compte dans les **traductions** et les **interfaces**[40]) sera encore plus forte que pour E. Cet indicateur va faire reculer encore plus dans le classement les langues les moins bien notées mais il reflète une réalité que l'on ne peut ignorer.

### 4.5.4 Traitements pour obtenir des résultats différentiés par thème

Le seul indicateur possédant un nombre suffisant éléments pour permettre une analyse plus fine par thème est celui du trafic avec ses 316 sources. Pour chacun des site web il est associé un thème et les moyennes réduites à 20% sont calculées pour chaque thème en montrant la différence positive ou négative para rapport à la note de moyenne générale. Les thèmes retenus sont les suivants:
*blog, courrier électronique, informatique, jeux, messagerie électronique, MOOCS, moteurs de recherche, outils informatique, partage de fichiers, questions/réponses, recherche scientifique, répertoire, réseaux sociaux images, réseaux sociaux musique, réseaux sociaux professionnels, réseaux sociaux rencontres, réseaux sociaux textes, téléphonie IP.*

Il faut toute fois prendre ces résultats avec précaution, surtout dans le cas où le nombre de sites correspondant à un thème donné est faible. Les résultats sont présentés dans l'onglet DIFF.

### 4.5.5 Structure du fichier Excel de référence

TABLE 4 **Description du fichier de travail Excel**

| NOM | SIGNIFIE | CONTENU | Abscisse | Ordonnée | Traitements |
|---|---|---|---|---|---|
| UIT | | données UIT adaptées | | Code ISO pays | Adaptation pays |
| LOC1 | Locuteurs L1 | | Code ISO pays | Code ISO langue | Internautes par langue (col. GM et GN) |
| LOC2 | Locuteurs L2 | | Indicateurs | Code ISO langue | L1+L2/L1 + données par langue |
| PL | Pourcentage Langues | Pourcentage de locuteurs d'une langue par pays | Code ISO pays | Code ISO langue | |
| LI | Langue Internautes | Nombre d'internautes par langue selon les pays | Code ISO pays | Code ISO langue | |
| MIL | Micro-indicateurs par langue | Données saisies | Micro-indicateurs | Code ISO langue | |
| MIP | Micro-indicateurs par pays | Données saisies | Micro-indicateurs | Code ISO langue | |
| Ma | Masque Absence | 1 si pays renseigné | Micro-indicateurs | Code ISO langue | Sert à l'extrapolation |
| Mp | Masque Présence | 1 si pays non renseigné | Micro-indicateurs | Code ISO langue | Sert à l'extrapolation |
| EX | Extrapolation | Formules extrapolation | Micro-indicateurs | Code ISO langue | |
| RES1 | Résultats par langue L1 | Pondération pays -> langues | Indicateurs | Code ISO langue | Indicateurs L1 colonnes centrales |
| RES2 | Résultats par langue L1+L2 | Pondération pays -> langues | Indicateurs | Code ISO langue | Indicateurs L1+L2 colonnes centrales |
| SYNTHESES | Ensemble des résultats | | Code ISO langue | Macro-indicateurs | |
| DIFF | Différentiation thématique | | | | |
| Matrice | Définition micro-indicateurs | Thématiques | | | |

[40] Pour cet indicateur, par exemple, le créole de Guadeloupe a une note de zéro.

### 4.5.6 Organisation des calculs dans le fichier Excel

Les traitements pour obtenir les indicateurs (pondérations et moyennes) sont réalisés dans les colonnes centrales (marquées en rouge) des onglets RES1 et RES2 puis reportés dans l'onglet SYNTHESES. L'onglet LOC2 qui gère les quantités L2 a été complémenté par la copie d'une série de données (calculées par ailleurs dans les onglets RES1 et RES2) dans un souci de lisibilité de données essentielles. Cela concerne principalement les personnes connectées à l'Internet et les pourcentages qui en dérivent et en dessous l'analyse comparative avec les données de InternetWorldStats.

L'onglet SYNTHESES a été introduit dans la feuille de calcul pour permettre d'intégrer les traitements qui sont présentés en séquence pour chacune des 140 langues, avec reprise dessous d'un classement des 15 les plus présentes. Tous les éléments à l'exception de capacité, gradient et productivité[41] portent sur **des pourcentages mondiaux par langue**:

- Calcul du classement par rapport à la puissance
- Report du pourcentage d'internautes L1
- Report du pourcentage d'internautes L1+L2
- Report des résultats de l'indicateur trafic L1
- Report des résultats de l'indicateur trafic L1+L2
- Report de l'indicateur usages L1
- Report de l'indicateur usages L1+L2
- Report de l'indicateur contenus L1
- Report de l'indicateur contenus L1+L2
- Report de l'indicateur interfaces L1+L2
- Report de l'indicateur indexes L1+L2
- Calcul du macro-indicateur puissance, moyenne de tous les indicateurs L1
- Calcul du macro-indicateur puissance , moyenne de tous les indicateurs L1+L2[42]
- Calcul du macro-indicateur capacité, quotient du macro-indicateur puissance L1+L2 par le pourcentage de locuteurs de la langue L1+L2[43]
- Report du pourcentage mondial de locuteurs
- Calcul du macro-indicateur puissance L1+L2 ramené à 100%.
- Calcul de macro-indicateur gradient, quotient du macro-indicateur puissance L1+L2 par le pourcentage d'internautes de la langue L1+L2
- Report des données de contenus de W3TECHS
- Calcul de la productivité de production de contenus selon W3TECHS, quotient de pourcentage mondial de contenus par pourcentage mondial d'internautes

Ces mêmes résultats détaillés pour chacune des 140 langues considérées sont ensuite regroupés en dessous pour les 15 langues avec les meilleures valeurs de moyenne générale L1+L2.

[41] La productivité pour les contenus est le quotient entre le pourcentage mondial de contenus et le pourcentage mondial d'internautes.
[42] Cet élément est considéré comme le produit principal de l'étude.
[43] Une valeur de 1 montre une capacité normale, c'est à dire proportionnelle à la présence de la langue dans le monde. Une valeur inférieure à 1 montre une faiblesse; la valeur est d'autant supérieur à 1 que la langue a une forte capacité dans l'Internet.

Dans le même tableau, il est proposé un résultat purement spéculatif qui serait, au vue des résultats, de l'analyse des différents biais observés et de l'expérience passée sur ce genre d'étude, la meilleure estimation de l'auteur de ce rapport de ce que pourrait être la réalité (exprimée en termes de contenus), si l'on pouvait éliminer les biais. L'analyse des ratios précédents pour les langues mentionnées et le reste des langues sont les facteurs clef de cette spéculation.

Plus bas dans la feuille, les facteurs *capacité*, *gradients* et *pourcentages de locuteurs connectés* sont mis en perspective pour l'ensemble des langues puis triés.

Dans l'onglet LOC2 il est rajouté 3 colonnes qui permettent de comparer les résultats de cette étude avec :

- les valeurs proposées par InternetWorldStats pour les 10 langues les plus présentes
- les valeurs proposées par W3Techs pour le pourcentage de contenus par langue
- le quotient des valeurs de W3Techs par la moyenne générale produite pour L1 pour L2.

Ces données vont permettre de soutenir les discussions sur les biais des sources dans le chapitre *7. Limites méthodologiques, analyse des biais et contrôles réalisés*.

### 4.5.7 Les différents types de pondérations utilisés

La méthode utilisée repose sur la technique de pondération; il est utile d'identifier et récapituler les différents types de pondération utilisés dans les traitements et de présenter les hypothèses simplificatrices qui sous-tendent la validité des résultats obtenus par la pondération (les discussions sur les implications et les biais possibles de ces hypothèses sont énoncées au chapitre 7.1 *Limitations propres à la méthode*).

La pondération principale est celle qui permet de **transformer les données pays en données langue**: les données d'un micro-indicateur exprimé en terme de pays sont pondérées par rapport à la répartition des langues dans les pays ce qui permet d'obtenir la répartition des données du micro-indicateur par langue. L'hypothèse simplificatrice qui soutient la validité est que le critère exprimé par le micro-indicateur s'exprime de la même manière pour toutes les langues dans chaque pays.

La pondération qui permet de **transformer les résultats en termes de L1 en résultats en termes de L1+L2** est réalisée à partir des taux d'accroissements qui permettent de passer du nombre de locuteurs L1 au nombre de locuteurs L1+L2[44] pour chaque langue. L'hypothèse simplificatrice qui soutient cette méthode est que les pourcentages de connexion à l'Internet sont les mêmes pour les L2 que pour les L1 indépendamment des pays de résidence des L2.

La pondération qui permet **de transformer des résultat en termes de pourcentage par rapport à un critère donné (interfaces, indexes) à un pourcentage mondial** est faite par rapport aux pourcentages de personnes connectées par langue. Le principe qui soutient cette méthode est que les pourcentage mondiaux obtenus sont une déviation des pourcentages mondiaux d'internautes par langue en fonction le la répartition des pourcentages dans ce critère par langue. Une approximation plus intuitive du sens de cette pondération est fournie dans le *chapitre 7.4.8 A propos du principe de pondération.*

[44] Une méthode plus fine qui est discutée en 7.2.2 *Le cas des L2* ne pondère pas directement à partir des taux d'accroissement mais à partir de la même valeur pondérée par la méthode des quartiles par la répartition des locuteurs L2 dans les pays en fonction de leur pourcentage de connexion à l'Internet.

TABLE 5 Les 3 types de pondérations utilisés.

| | Démo-linguistique | Langues secondes | Internautes par langue |
|---|---|---|---|
| TYPE | Pays ---> langue | L1 ---> L1+L2 | % critère --> % mondial |
| APPLICATION | Données par pays | Résultats L1 | Donnée en % par critère |
| RÉSULTAT | Données par langue | Résultats L1+L2 | Donnée en % mondial |
| DONNÉES DE PONDÉRATION | Matrice LOC | Vecteur L1+L2/L1 par langue | Données IUT |
| CHAMP D'APPLICATION | Toutes les sources par pays | Tous les micro-indicateurs pays | Micro-indicateurs index et interfaces. |
| HYPOTHÈSE SIMPLIFICATRICE | Indépendance par rapport aux langues dans les pays | Indépendance des taux de connexion Internet par rapport aux pays | Modulation des taux de connexion à l'Internet en fonction du critère |

# 5. Résultats

## 5.1 Synthèse des résultats

Le nombre **d'internautes** locuteurs du français est estimé à une quote-part de **5,6%.** Le **trafic Internet** généré par ces internautes vers les contenus existants, place le français en **3ième place avec une quote-part de 7,8%** du trafic mondial. Pour l'indicateur "**usages**" qui mesure l'accès aux infrastructures et à des applications majeures[45], le français prend la **4ième place avec 7,4%** de quote-part. Pour l'indicateur **contenus** (Wikimédia, W3Techs et le nombre de livres proposés par Amazon), le français occupe la **2ième place avec une quote-part de 9,3%**. Si l'on prend en compte le macro-indicateur qui évalue les avancées des pays dans la société de l'information à partir **d'indexes** ramenés en pourcentages mondiaux par langue alors le français garde la **4ième place avec 7,3%**. Finalement si l'on mesure l'indicateur qui quantifie les **langues d'interface et de traduction** le français est en **quatrième position de nouveau avec 7,4%.** Finalement si l'on combine **l'ensemble des facteurs pour déterminer la puissance d'une langue dans l'Internet, le français occupe la 4ième place avec 7,5%**, derrière respectivement *l'anglais, le chinois et l'espagnol* et avec un écart supérieur à 3% par rapport aux suivants[46]. Si l'on s'intéresse à des facteurs qui sont indépendants du nombre de locuteurs, alors le français se maintient à la onzième place en terme de **pourcentage de personnes** connectées (81%), en termes de **capacité (3)**, et en terme de **gradient (1,4)**.

TABLE 6 Classement du français dans l'Internet

| CRITERES L1+L2 | CLASSEMENT MONDIAL | QUOTE-PART MONDIALE | QUOTE-PART /100% |
|---|---|---|---|
| **Internautes** | 4ième | 5,6% | 4,4% |
| **Trafic** | 3ième | 7,8% | 8,2% |
| **Usages** | 4ième | 7,4% | 6,6% |
| **Contenus** | 2ième | 9,3% | 7,1% |
| **Indexes** | 4ième | 7,3% | 5,8% |
| **Interfaces** | 4ième | 7,4% | 6,0% |
| **PUISSANCE** | **4ième** | **7,5%** | **6%** |
| **% Loc. connectés** | **10ième** | **81%** | |
| **CAPACITÉ** | **10ième** | **3** | |
| **GRADIENT** | **12ième** | **1,4** | |

[45] Cela inclut par exemple les lignes téléphoniques fixes, les personnes ayant un compte dans certains réseaux sociaux ou le nombre de téléchargement de OpenOffice.
[46] Russe, allemand, portugais et japonais avec respectivement 4,7%, 4,6%, 3,8%, 3,5%.

En plus de déterminer une place et une puissance pour le français dans l'Internet, l'étude permet, dans certain cas, grâce à la quantité de mesures mis en œuvre, de différencier sa position selon différentes thématiques établies, et ainsi d'apprécier ses forces et ses faiblesses. La table suivante présente la différence à la moyenne de l'indicateur trafic para rapport à la nature thématique des sites visités. Seuls sont mentionnés les thèmes qui ont été calculés avec un nombre de sites supérieur à 14, pour éviter de tirer des conclusions à partir d'un échantillon statistique insuffisant (sans ce critère les MOOCs apparaitraient en tête de très loin et les réseaux sociaux professionnels seraient en bonne place).

TABLE 7 **Différenciation par thème de la place du français dans l'Internet**

| THEME | Nombre de sites | Moyenne réduite 5% | ECART A LA MOYENNE | QUALIFICATION |
|---|---|---|---|---|
| Recherche | 17 | 17,8% | 173% | +++ |
| Moteurs de recherche | 28 | 14,0% | 115% | +++ |
| Jeux | 17 | 13,7% | 110% | +++ |
| Outils informatique | 40 | 8,5% | 30% | ++ |
| Réseaux sociaux rencontres | 15 | 6,7% | 3% | |
| Réseaux sociaux images | 39 | 6,2% | -5% | |
| Courrier électronique | 14 | 6,2% | -5% | |
| Blog | 14 | 5,8% | -11% | - |
| Réseaux sociaux textes | 42 | 5,5% | -16% | - |
| **Moyenne réduite 20%** | **42** | **5,5%** | | |

Ainsi il apparait que le français est particulièrement présent dans l'Internet dans ce qui a un rapport avec la **recherche scientifique**, et à un degré moindre **les jeux et les moteurs de recherche.** Les bonnes notes dans les **outils informatiques** et les **jeux** pourraient être contrastées avec les notes basses dans les autres réseaux sociaux et les outils techniques de partage.

Il faut toutefois prendre ces informations avec une certaine réserve et seulement comme des indications de tendance étant donné le nombre limité de sites pour la plupart des thèmes.

Le tableau suivant présente l'ensemble des résultats de synthèse pour les 15 langues les mieux notées dans l'Internet en prenant comme notation la moyenne de 6 indicateurs mesurés. Les résultats, qui sont tous des pourcentages mondiaux, sont présentés, séparés le cas échéant entre langues premières (L1) et langue première plus seconde (L1+L2), pour les paramètres suivants: **internautes, trafic, usages, contenus, indexes et interfaces**, ainsi que pour la moyenne des précédents qui constitue le macro-indicateur **puissance**.

TABLE 8 **RESULTATS POUR 15 LANGUES**

| | INTERNAUTES | | TRAFIC | | USAGES | | CONTENUS | INTERFACES | INDEXES | PUISSANCE |
|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L1+L2 | L1 | L1+L2 | L1 | L1+L2 | L1+L2 | L1+L2 | L1+L2 | L1+L2 |
| Anglais | 0,085 | 0,222 | 0,283 | 0,568 | 0,192 | 0,437 | 0,499 | 0,309 | 0,305 | **0,390** |
| Chinois | 0,218 | 0,205 | 0,081 | 0,062 | 0,125 | 0,108 | 0,030 | 0,248 | 0,194 | **0,141** |
| Espagnol | 0,080 | 0,091 | 0,109 | 0,094 | 0,095 | 0,095 | 0,053 | 0,121 | 0,090 | **0,091** |
| Français | 0,020 | 0,056 | 0,036 | 0,080 | 0,029 | 0,074 | 0,093 | 0,074 | 0,073 | **0,075** |
| Russe | 0,028 | 0,050 | 0,015 | 0,022 | 0,025 | 0,041 | 0,054 | 0,066 | 0,051 | **0,047** |
| Allemand | 0,019 | 0,031 | 0,028 | 0,037 | 0,026 | 0,039 | 0,084 | 0,042 | 0,039 | **0,046** |
| Portugais | 0,040 | 0,040 | 0,026 | 0,020 | 0,048 | 0,043 | 0,033 | 0,053 | 0,039 | **0,038** |
| Japonais | 0,036 | 0,034 | 0,031 | 0,024 | 0,038 | 0,032 | 0,030 | 0,043 | 0,046 | **0,035** |
| Arabe | 0,044 | 0,042 | 0,042 | 0,031 | 0,032 | 0,028 | 0,013 | 0,048 | 0,032 | **0,032** |
| Hindi | 0,031 | 0,039 | 0,020 | 0,020 | 0,024 | 0,027 | 0,003 | 0,035 | 0,033 | **0,026** |
| Malais | 0,006 | 0,026 | 0,005 | 0,015 | 0,006 | 0,022 | 0,013 | 0,027 | 0,021 | **0,021** |
| Italien | 0,009 | 0,009 | 0,014 | 0,011 | 0,011 | 0,010 | 0,041 | 0,012 | 0,010 | **0,016** |
| Coréen | 0,015 | 0,014 | 0,012 | 0,009 | 0,011 | 0,009 | 0,009 | 0,016 | 0,016 | **0,012** |
| Polonais | 0,009 | 0,008 | 0,010 | 0,008 | 0,009 | 0,008 | 0,030 | 0,009 | 0,008 | **0,012** |
| Ourdou | 0,009 | 0,017 | 0,009 | 0,013 | 0,008 | 0,013 | 0,001 | 0,011 | 0,014 | **0,011** |
| **RESTE** | **0,351** | **0,368** | **0,279** | **0,236** | **0,321** | **0,265** | **0,264** | **0,135** | **0,279** | **0,258** |
| **TOTAL** | **1,000** | **1,250** | **1,000** | **1,250** | **1,000** | **1,250** | **1,250** | **1,250** | **1,250** | **1,250** |

Il faut lire le tableau de la manière suivante, pour le français, colonne par colonne, séquentiellement:

- **Colonne *Internautes* L1**: le français langue première, représente **2,0%** des personnes connectées à l'Internet.
- **Colonne *Internautes* L1+L2**: **5,6%** des personnes connectées à l'Internet possèdent le français comme langue première ou seconde[47].
- **Colonne *trafic* L1**: **3,6%** du trafic vers un large échantillon de sites web représentatif d'une variété équilibrée de thématiques provient de personnes ayant le français comme première langue.
- **Colonne *trafic* L1+L2: 7,8%** du trafic vers un large échantillon de sites web représentatif d'une variété équilibrée de thématiques provient de personnes ayant le français comme première ou deuxième langue.
- **Colonne *usages* L1**: **3,0%** des usagers de services liés à l'Internet ont le français comme langue maternelle.
- **Colonne *usages* L1+L2**: 7**,3%** des usagers de services liés à l'Internet ont le français comme langue maternelle ou seconde.
- **Colonne *contenus***: **9,3%** des contenus sont en français dans l'Internet[48].
- **Colonne *indexes***: **7,4%** représente la quote-part du français dans la société de l'information en pondérant ces scores dans les indexes avec sa population connectée dans l'Internet.

[47] A noter que le total est supérieur à 100% pour tenir compte du multilinguisme.
[48] Cet énoncé, qui prend en compte le multilinguisme existant dans les contenus, doit être reçu avec prudence, sachant qu'il porte principalement sur les contenus liés aux différents services de Wikimedia.

- **Colonne *interfaces***: **7,4%** représente la quote-part du français en tant que langue d'interface ou de traduction en pondérant son pourcentage de présence avec sa population connectée dans l'Internet.
- **Colonne *puissance* L1+L2**: **7,5%** représente une approximation générale de la quote-part de puissance du français dans l'Internet en faisant la moyenne des éléments précédents (3,1% si l'on considère seulement la langue maternelle).

La ligne RESTE représente les résultats concernant l'ensemble complet de toutes les langues du monde à l'exception des 15 langues mentionnées dans le tableau.

Les résultats montrent de manière cohérente une 4ième place pour le français (L1+L2) dans l'Internet derrière l'anglais, le chinois et l'espagnol et avec une confortable avance devant le russe et l'allemand (très proches) puis le portugais. Si on ne prenait en considération que la langue première, le français serait en 7ième position.

Des hypothèses différentes sur les valeurs des populations des langues secondes, une méthode plus précise pour les langues secondes, ou encore le choix d'une autre métrique pour la détermination des indicateurs (moyennes, moyennes réduites, médianes ou regroupements différents des micro-indicateurs) conduiraient évidemment à des valeurs présentant certaines différences, relativement mineures[49] mais susceptibles de changer l'ordre des langues assez proches. Cependant des analyses de sensibilité des facteurs montrent que l'avance forte du français (L1+L2) sur les positions suivantes garantissent cette quatrième place, quelque soit la méthode utilisée pour le macro-indicateur puissance et de même en changeant sensiblement les hypothèses pour les L2. Des simulations pratiquées sur les Li (un concept de langue de l'Internet qui serait la somme L1 + L2 + Li population possédant une maitrise suffisante de la langue pour faire un usage complet et compétent des services de l'Internet) montent également que cette quatrième place est très solide.

Le tableau suivant complémente les données relative aux 15 langues les plus puissantes de l'Internet avec les concepts de **capacité** et de **gradient.**

**La capacité** mesure par le ratio entre la puissance et le pourcentage de locuteurs par rapport à la population mondiale. Ce ratio exprime la capacité propre à la langue dans l'Internet, indépendamment de la taille de sa population. Entre d'autres termes la **puissance** de la langue française dans l'Internet exprimée par la note de 7,5% est le produit de sa **capacité** dans l'Internet (3) par la taille de sa **population mondiale** (2,5%).

Le **gradient** mesure le ratio entre la puissance générique et le pourcentage de locuteurs connectés par rapport à la population mondiale d'internautes. Ce ratio exprime le potentiel d'impact des locuteurs connectés indépendamment de leur nombre.

Le tableau présente également la puissance exprimée sur la base de 100% à des fins de comparaisons.

[49] Probablement affectant le chiffre après la virgule dans une amplitude maximum de + ou - 0,3% si l'on considère seulement les méthodes de regroupement des micro-indicateurs ou de calculs et bien plus si les écarts sur les L2 sont importants.

TABLE 9 **Capacité et gradient dans l'Internet**

| | PUISSANCE L1+L2 | CAPACITE L1+L2 | % POP. MOND. L1+L2 | PUISS./100% L1+L2 | GRADIENT L1+L2 |
|---|---|---|---|---|---|
| Anglais | **0,390** | 3,72 | 0,105 | 0,312 | 1,76 |
| Chinois | **0,141** | 0,96 | 0,148 | 0,113 | 0,69 |
| Espagnol | **0,091** | 1,58 | 0,058 | 0,073 | 1,00 |
| Français | **0,075** | 2,96 | 0,025 | 0,060 | 1,35 |
| Russe | **0,047** | 1,77 | 0,027 | 0,038 | 0,95 |
| Allemand | **0,046** | 3,40 | 0,013 | 0,036 | 1,45 |
| Portugais | **0,038** | 1,52 | 0,025 | 0,030 | 0,95 |
| Japonais | **0,035** | 2,54 | 0,014 | 0,028 | 1,03 |
| Arabe | **0,032** | 0,88 | 0,037 | 0,026 | 0,77 |
| Hindi | **0,026** | 0,48 | 0,054 | 0,021 | 0,67 |
| Malais | **0,021** | 0,82 | 0,025 | 0,016 | 0,79 |
| Italien | **0,016** | 3,16 | 0,005 | 0,012 | 1,73 |
| Coréen | **0,012** | 1,41 | 0,009 | 0,010 | 0,87 |
| Polonais | **0,012** | 2,64 | 0,004 | 0,009 | 1,41 |
| Ourdou | **0,011** | 0,54 | 0,021 | 0,009 | 0,69 |
| **RESTE** | **0,258** | **0,38** | **0,680** | **0,206** | **0,70** |
| **TOTAL** | **1,250** | | **1,250** | **1,000** | |

Un classement à partir du critère de capacité donnerait, à partir des mêmes données, des résultats très différents de celui de la puissance.

TABLE 10 **Classement à partir du critère capacité**

| | | Classement Puissance | Capacité |
|---|---|---|---|
| 1 | Hébreu | 35 | 5,40 |
| 2 | Finnois | 38 | 5,40 |
| 3 | Néerlandais | 19 | 4,81 |
| 4 | Suédois | 28 | 4,46 |
| 5 | Anglais | 1 | 3,72 |
| 6 | Allemand | 6 | 3,40 |
| 7 | Danois | 49 | 3,30 |
| 8 | Italien | 12 | 3,16 |
| 9 | Tchèque | 27 | 3,13 |
| 10 | Français | 4 | 2,96 |

Les langues nationales de pays reconnues pour leurs politiques volontaristes pour la société de l'information apparaissent dans les premières positions. Le français n'est pas mal positionné mais il est clair que les langues dont les populations sont majoritairement très centrées sur des pays industrialisés ont un avantage stratégique dans ce classement para rapport à des langues comme le français qui ont une partie importante de locuteurs dans des pays en développement. Il est d'ailleurs remarquable que plusieurs langues dépassent l'anglais qui a d'autres avantages stratégiques dans l'Internet liées à la croyance persistante de nombreuses personnes en un prétendu rôle de lingua franca.

Il est clair qu'il y a une forte et logique corrélation entre la capacité et le pourcentage de personnes connectées à l'Internet; cependant la corrélation n'est pas totale comme le montre le classement suivant trié par le pourcentage de personnes connectées:

TABLE 11 **Classement à partir du critère pourcentage de personnes connectées**

| | | % Personnes Connectées | Classement Puissance |
|---|---|---|---|
| 1 | Danois | 95,67% | 49 |
| 2 | Finnois | 92,30% | 38 |
| 3 | Néerlandais | 92,27% | 19 |
| 4 | Suédois | 90,54% | 28 |
| 5 | Japonais | 90,43% | 8 |
| 6 | Allemand | 86,43% | 6 |
| 7 | Alémanique | 86,41% | 51 |
| 8 | Bavarois | 84,42% | 39 |
| 9 | Flamand | 83,60% | 59 |
| 10 | Tchèque | 81,17% | 27 |
| 11 | Français | 81,09% | 4 |
| 12 | Anglais | 78,05% | 1 |

Lors des essais incluant des langues de France, dans ce classement certaines sont apparues logiquement au dessus du français pour les raisons déjà évoquées (voir *7.2.5 Le cas des langues locales de France*).

TABLE 12 **Classement à partir du critère de gradient**

| | | Classement Puissance | Gradient |
|---|---|---|---|
| 1 | Hébreu | 35 | 2,62 |
| 2 | Finnois | 38 | 2,16 |
| 3 | Néerlandais | 19 | 1,93 |
| 4 | Suédois | 28 | 1,81 |
| 5 | Anglais | 1 | 1,76 |
| 6 | Italien | 12 | 1,73 |
| 7 | Serbo-croate | 22 | 1,54 |
| 8 | Hongrois | 32 | 1,47 |
| 9 | Allemand | 6 | 1,45 |
| 10 | Tchèque | 27 | 1,42 |
| 11 | Polonais | 14 | 1,41 |
| 12 | Français | 4 | 1,35 |

## 5.2 Comparaison avec les résultats de la précédente étude

Les deux études précédentes ne possédaient pas suffisamment de micro-indicateurs pour construire des indicateurs qui permettent de travailler directement en termes de pourcentage (ou quote-part). Les calculs de pondération étaient réalisés sur la base des classements avec un autre système de

pondération plus simple basé sur les classements. La dernière étude de 2013 indiquaient un classement pondéré de 7,4[50] pour le français en L1 et de 4,3 pour le français (L1+L2).

De ce point de vue, on peut seulement affirmer que la place de quatrième de la langue française est consolidée et rajouter que cette place est devenu extrêmement solide avec une avance forte sur les suivantes. Il faut noter que le classement langue maternelle seulement (L1) maintient également le français à la 7ième place. Il est difficile d'évaluer la progression à partir de métriques différentes mais les chiffres annoncés plaident pour diagnostiquer un progrès qui se manifeste par une très solide consolidation dans les respectives 7ième place (L1) et 4ième place (L1+L2) et que la méthode va plus loin en mesurant la puissance du français dans l'Internet, ce qui pourrait représenter une évaluation indirecte des contenus en français (7,7%). L'étude ouvre la porte à une nouvelle mesure qui serait la plus proche de la réalité, celle des Li, **langue d'usage dans l'Internet** pour laquelle la première simulation (voir plus loin) donnerait toujours le français en 4ième position.

## 5.3 Analyse des résultats

Pour revenir au premier classement général, il est important de savoir que deux des six indicateurs (le trafic mesuré par Alexa et les statistiques de Wikipédia) peuvent présenter, chacun à sa manière, des biais très importants, en faveur de l'anglais et au détriment, en premier lieu, des langues non occidentales (en particulier le chinois, l'hindi, l'ourdou, le bengali et en moindre mesure l'arabe) et en second lieu, dans une moindre proportion, de certaines autres langues occidentales, dans le cas d'Alexa (le portugais et dans une moindre mesure l'espagnol).

Quant à Wikipédia, ses statistiques sont impeccables; par contre il faut comprendre que l'application la plus mondialisée de l'Internet connait pour certaines langues asiatiques une utilisation proportionnellement très en dessous de sa présence dans l'Internet.

TABLE 13 **Ratios articles Wikipédia/internautes**

| LANGUE | % DE PERSONNES CONNECTÉES | % D'ARTICLES WIKIPEDIA | RATIO ARTICLES / INTERNAUTES |
|---|---|---|---|
| Suédois | 0,2% | 8,8% | 44 |
| Néerlandais | 0,4% | 4,5% | 11 |
| Français | 5,6% | 4,3% | 0,8 |
| Anglais | 22,6% | 12,6% | 0,6 |
| Chinois | 20,9% | 2,8% | 0,13 |
| Ourdou | 2,0% | 0,3% | 0,15 |
| Hindi | 3,9% | 0,3% | 0,08 |
| Bengali | 1,3% | 0,1% | 0,08 |

Si on tentait de corriger ces biais à partir d'un classement purement spéculatif basé sur le meilleur jugement, à partir des biais analysés dans les sources (voir *7.4.6 Biais potentiels de Alexa.com et W3Techs*), les langues citées remonteraient puissamment dans les classements mais le français conserverait cette avantageuse 4ième position et sa confortable avance sur les suivants. Il y a tout lieu de penser que cette avance va se maintenir avec les autres langues occidentales (hors anglais); par contre à long terme, dans la mesure où la fracture numérique continue de se combler au niveau planétaire, la puissance dans l'Internet tendant naturellement vers la puissance démographique, l'avance du français pourra se réduire avec les grandes langues en terme démographique: hindi, arabe, bengali, russe et ourdou, sauf bien sûr si la démographie francophone en Afrique compense ces différences.

[50] Ce qui porte le sens de "classé en moyenne pondéré entre la 7ième et la 8ième place, plus près de la 7ième."

Il est aussi important de savoir qu'il existe une sensibilité relative de ces classements à partir des données démo-linguistiques pour lesquelles les données peuvent varier dans des proportions considérables (surtout pour l'anglais).

De fait, la méthode simple adoptée pour les L2 pourrait s'étendre à un concept de Li (langue de travail sur l'Internet) qui comptabiliserait en plus des L1 et des L2 les "*locuteurs capables de naviguer aisément dans l'ensemble des fonctions de l'Internet*" (courriel, chat, navigation web, réseaux sociaux..), ce qui serait une sorte de L1+ L2 (défini avec rigueur) + L3 (qui serait une langue apprise avec un niveau suffisant pour répondre à la définition). L'instrument mis au point pourrait même servir pour réaliser des simulations prospectives (un scénario est donné plus loin avec les Li, à titre d'exemple, qui montre une fois de plus la solidité de la quatrième place du français dans l'Internet).

En conclusion, en plus des résultats basés sur des hypothèse démo-linguistiques et la collection de micro-indicateurs, cette étude offre un **instrument de travail et de simulation** qui permet d'obtenir des résultats immédiats à partir d'autres hypothèses démo-linguistiques, ce qui en fait probablement un instrument unique et sans précédent dans le domaine de la *cybermétrie* linguistique.

### 5.3.1 Comparaison des résultats avec ceux de InternetWorldStats

Il est intéressant de comparer les résultats de cette étude avec ceux de InternetWorldStats, source la plus citée depuis des années pour des chiffres d'internautes par langues (limitée seulement aux 10 langues ayant le plus grand nombre de locuteurs connectés).

Les deux tableaux ci-dessous indiquent des différences notables: elles proviennent d'hypothèses différentes pour les L2 mais pas seulement. Il y a une forte sous-estimation des personnes connectées pour le français, le russe et dans un degré moindre pour l'espagnol et l'allemand. Comme il est impossible de savoir comment c'est faite la construction des résultats de IWS la comparaison est également faite avec une reprise simulée des données démographiques de IWS. Il faut noter que pour établir ces comparaisons le choix imposé par IWS est de travailler sur des données ramenées à 100. En effet, IWS semble partir de données démographiques L1+L2+L3 (au moins pour l'anglais) et maintenir ses totaux à 100% ce qui conduit à des incohérences si la vue s'élargit vers le reste des langues. En effet, dans la simulation cela amène un reste des langues négatif signal que le multilinguisme n'est pas correctement pris en compte. La simulation avec les valeurs démo-linguistiques de IWS montre des chiffres beaucoup plus grands pour la plupart des langues (parmi les 10 mesurées) mais il reste que cela révèle une sous-estimation du français encore plus forte par IWS qui ne serait donc pas due à ses hypothèses mais probablement à une erreur de calcul.

Le tableau suivant recopie telle quelle les données de InternetWorldStats[51].

[51] http://www.internetworldstats.com/stats7.htm

TABLE 14 **Données de InternetWorldStats**

| LANGUE | INTERNAUTES | % LOCUTEURS CONNECTES | % MONDIAL | POPULATION |
|---|---|---|---|---|
| Anglais | 948 608 782 | 67,76% | 26,27% | 1 400 052 373 |
| Chinois | 751 985 224 | 53,12% | 20,82% | 1 415 572 934 |
| Espagnol | 277 125 947 | 61,55% | 7,67% | 450 235 963 |
| Arabe | 168 426 690 | 43,37% | 4,66% | 388 332 877 |
| Portugais | 154 525 606 | 57,93% | 4,28% | 266 757 744 |
| Japonais | 115 111 595 | 91,02% | 3,19% | 126 464 583 |
| Malais | 109 400 982 | 37,76% | 3,03% | 289 702 633 |
| Russe | 103 147 691 | 70,48% | 2,86% | 146 358 055 |
| Français | 102 171 481 | 25,94% | 2,83% | 393 892 299 |
| Allemand | 83 825 134 | 88,26% | 2,32% | 94 973 855 |
| Reste | 797 046 681 | 33,66% | 22,07% | 2 367 750 664 |
| TOTAL | 3 611 375 813 | 49,20% | 100,00% | 7 340 093 980 |

Le tableau suivant établit la comparaison avec les données obtenues dans cette étude et l'on voit que les langues suivantes subissent une différence marquées, par ordre décroissant: le reste des langes (+56%), le français (-53%), le russe (-47%), l'allemand (-31%), l'espagnol (-22%). La différence sur le reste provient de la manière de gérer le multilinguisme de IWS.

TABLE 15 **Comparaison InternetWorldStats vs. résultats de l'étude ramenés à 100%**

| LANGUE | INTERNAUTES | % LOCUTEURS CONNECTES | % MONDIAL | POPULATION | DIFFÉRENCES | |
|---|---|---|---|---|---|---|
| | | | | | | en % |
| Anglais | 746 575 851 | 78,18% | 23,94% | 954 971 600 | -2,32% | -9,71% |
| Chinois | 692 230 352 | 51,36% | 22,20% | 1 347 846 000 | 1,38% | 6,21% |
| Espagnol | 306 111 437 | 58,21% | 9,82% | 525 877 800 | 2,14% | 21,83% |
| Arabe | 140 122 838 | 41,97% | 4,49% | 333 842 070 | -0,17% | -3,78% |
| Portugais | 134 142 584 | 58,74% | 4,30% | 228 372 370 | 0,02% | 0,54% |
| Japonais | 113 372 837 | 90,43% | 3,64% | 125 376 800 | 0,45% | 12,33% |
| Malais | 87 701 961 | 38,30% | 2,81% | 228 998 640 | -0,22% | -7,70% |
| Russe | 166 904 263 | 68,60% | 5,35% | 243 287 120 | 2,50% | 46,64% |
| Français | 186 641 894 | 81,10% | 5,99% | 230 146 420 | 3,16% | 52,73% |
| Allemand | 105 539 688 | 86,43% | 3,38% | 122 108 000 | 1,06% | 31,42% |
| **Reste** | **438 771 072** | **16,05%** | **14,07%** | **2 733 112 497** | **-8,00%** | -56,84% |
| **TOTAL** | **3 118 114 777** | **44,08%** | **100,00%** | **7 073 939 317** | **0,00%** | |

Le tableau suivant indique les résultats si on appliquait les mêmes hypothèses démo-linguistiques que IWS. Sur ces hypothèses le modèle établit par cette étude trouverait des valeurs supérieures de 25% à celle que IWS propose pour l'anglais mais surtout des valeurs supérieures de 72% pour le français. Il

montre aussi un reste négatif, symptôme de la mauvaise prise en compte du multilinguisme. Les différences sont explicables par la différence de traitement du multilinguisme (le considérer pour les locuteurs par langue mais pas pour le total conduit à une contradiction chez IWS qui est cachée par le fait qu'il traite un nombre limité de langues). Si cela était corrigé, il resterait une erreur forte dans le traitement du français par IWS. Notre hypothèse explicative est que probablement IWS a calculé le nombre d'internautes pour le français sur la base des L1 et il a ensuite appliqué les pourcentages sur la base des L1+L2+L3 ce qui donne des pourcentages de personnes connectées largement sous-estimés pour le français.

TABLE 16 **Simulation avec les données de InternetWorldStats**

| LANGUE | INTERNAUTES | % LOCUTEURS CONNECTES | % MONDIAL | POPULATION | DIFFÉRENCES |
|---|---|---|---|---|---|
| Anglais | 1 094 530 237 | 78,18% | 35,10% | 1 400 052 373 | 25% |
| Chinois | 727 013 732 | 51,36% | 23,32% | 1 415 572 934 | 11% |
| Espagnol | 262 080 616 | 58,21% | 8,41% | 450 235 963 | 9% |
| Arabe | 162 994 152 | 41,97% | 5,23% | 388 332 877 | 11% |
| Portugais | 156 689 591 | 58,74% | 5,03% | 266 757 744 | 15% |
| Japonais | 114 356 472 | 90,43% | 3,67% | 126 464 583 | 13% |
| Malais | 110 950 392 | 38,30% | 3,56% | 289 702 633 | 15% |
| Russe | 100 407 220 | 68,60% | 3,22% | 146 358 055 | 11% |
| Français | 319 434 926 | 81,10% | 10,24% | 393 892 299 | 72% |
| Allemand | 82 087 259 | 86,43% | 2,63% | 94 973 855 | 12% |
| **Reste** | **-12 429 819** | **-0,59%** | **-0,40%** | **2 101 596 001** | |
| **TOTAL** | **3 118 114 777** | **44,08%** | **100,00%** | **7 073 939 317** | |

### 5.3.2 Simulation pour Li

Il est tentant quand on a sa disposition un outil rapide de simulation de l'utiliser, par exemple, pour s'approcher encore plus de la réalité sur le terrain, en introduisant des hypothèses pour le concept de Li, lequel regroupe les personnes qui ont une connaissance suffisante d'une langue d'enseignement (ni L1, ni L2) pour prétendre faire un usage plein et agile de l'Internet. Pour ce scénario de simulation les données proposées par une page de Wikipédia[52] sont utilisées quoiqu'elle semble généreuse avec l'anglais et beaucoup moins avec le français et les langues de l'Inde.

Il s'agit d'une simple simulation et l'OIF, qui possède une ample expertise linguistique en la matière, pourrait poursuivre les simulations avec des données plus crédibles. Les hypothèses pour L1, L2 et Li seraient dans ce cas:

[52] https://fr.Wikipédia.org/wiki/Liste_de_langues_par_nombre_total_de_locuteurs

TABLE 17 **Hypothèses pour Li**

| LANGUE | L1 | L2 | Li | TOTAL |
|---|---|---|---|---|
| Anglais | 372 | 611 | 600 | 1500 |
| Chinois | 898 | 194 | 40 | 1092 |
| Espagnol | 437 | 91 | 21 | 567 |
| Hindi | 260 | 121 | | 381 |
| Arabe | 101 | 246 | 5 | 352 |
| Français | 76 | 153 | 62 | 274 |
| Russe | 154 | 113 | 15 | 268 |
| Bengali | 242 | 19 | | 267 |
| Portugais | 219 | 11 | 5 | 240 |
| Malais | 23 | 175 | | 198 |
| Ourdou | 68 | 94 | | 162 |
| Allemand | 77 | 53 | 14,5 | 143 |
| Japonais | 128 | | 1 | 128 |

Note: Le tableau est repris tel quel avec les valeurs de la référence de la source, y compris les incongruités.

La simulation est conduite en conservant les valeurs de L1 présentes et en cadrant les valeurs de L2 sur les totaux du tableau précédent. On peut noter qu'avec ces données le pourcentage de multilinguisme monte à près de 1/3.

Le résultat de cette simulation est résumé dans la table suivante:

TABLE 18 **Simulation Li**

| | | % Internautes L1+L2 | puissance L1+L2 |
|---|---|---|---|
| 1 | Anglais | 0,3362 | 0,544 |
| 2 | Chinois | 0,1985 | 0,109 |
| 3 | Espagnol | 0,0946 | 0,084 |
| 4 | Français | 0,0637 | 0,080 |
| 5 | Allemand | 0,0354 | 0,048 |
| 6 | Russe | 0,0527 | 0,044 |
| 7 | Portugais | 0,0404 | 0,034 |
| 8 | Japonais | 0,0332 | 0,029 |
| 9 | Arabe | 0,0424 | 0,027 |
| 10 | Hindi | 0,0287 | 0,018 |
| 11 | Italien | 0,0087 | 0,017 |
| 12 | Malais | 0,0217 | 0,016 |
| 13 | Polonais | 0,0081 | 0,012 |
| 14 | Coréen | 0,0134 | 0,011 |
| 15 | Ourdou | 0,0130 | 0,008 |
| | RESTE | 0,3315 | 0,242 |
| | TOTAL | 1,3223 | 1,322 |

La simulation montre une quote-part de l'anglais supérieure à 50% et le français, toujours en 4ième position, maintenant très proche de l'espagnol. Dans cette simulation les langues de l'Inde descendent logiquement dans le classement.

# 6. Considérations prospectives

Ce chapitre examine les possibilités d'utilisation, suivi et amélioration de la méthode.

Comme il a été indiqué a plusieurs reprises, cette étude repose principalement en termes de données sur les indicateurs de personnes connectées à l'Internet par pays proposés par l'UIT.

En ce qui concerne un suivi de l'observation, il est ainsi recommandé de rééditer cette mesure **une fois par an après publication des nouvelles données par l'IUT.**

En ce qui concerne les possibilités d'amélioration des sources pour cette étude plusieurs directions sont à considérer.

- **Pour les données démo-linguistiques L1**: il est recommandé de tenter une alliance entre l'OIF, MAAYA et Ethnologue qui pourrait permettre de bénéficier du gigantesque travail de Ethnologue et de mises à jour annuelles. Une telle alliance pourrait donner une promotion mondiale des résultats à la hauteur des enjeux et permettre de luter efficacement contre la désinformation persistante au sujet de la place de l'anglais dans l'Internet, toujours largement surestimée par rapport à la réalité.

- **Pour les données démo-linguistiques L2**: il est recommandé de réaliser une nouvelle itération avec des données plus fines qui pourrait permettre grâce à la méthode des quartiles une meilleure approximation.

- **Pour l'indicateur INDEX**, il serait judicieux de l'étendre avec l'ensemble des sources identifiées. Cela n'aura probablement que très peu d'effet sur les résultats du français mais cela donnerait une meilleure assise à l'ensemble.

- **Pour l'indicateur USAGES**, il serait important de rajouter des micro-indicateurs en maintenant une veille sur les données d'abonnement dans des applications les plus utilisées de l'Internet.

- **Pour l'indicateur TRAFIC**, il pourrait être utile de pousser le nombre d'applications jusqu'à une valeur autour de 500 en faisant un effort d'identification de thématiques qui n'ont pas été considérées (un bon exemple serait une thématique sur l'ouverture) et en augmentant les sites pour les thématiques existantes pour lesquelles le nombre de site est trop faible pour en tirer des enseignements différentiés.

- **Pour l'indicateur CONTENUS**, le défi de trouver des sources additionnelles est très difficile à surmonter quoiqu'il est important de maintenir une veille. Une option possible de réduction des biais serait de systématiser la recherche des alternatives existantes à Wikipédia et de les intégrer dans le classement de manière à réduire les biais des résultats.

En ce qui concerne les possibilités d'amélioration des traitements voici des directions possibles:

- **La saisie des sources par pays** est pénible et comporte des risques d'erreur; la plupart des sources mentionnent les pays en anglais. Il serait utile de préparer un module qui serait capable d'intégrer les données avec un certain niveau d'automatisme.

- La nature structurelle du problème traité est celle d'une matrice à 3 dimensions (langue, pays, sources de données) pour laquelle Excel n'est pas l'outil le plus indiqué. Il pourrait être utile de **programmer la méthode en langage APL** ce qui en ferait un outil de simulation beaucoup plus souple.

En ce qui concerne l'utilisation de l'étude comme moyen de simulation, il serait intéressant de conduire quelques scénarios prospectifs à partir d'une prévision d'évolution des données démographiques et des données de connexion à l'Internet.

## 7. Limites méthodologiques, analyse des biais et contrôles réalisés

La démo-linguistique est loin d'être une science exacte et les divergences entre les spécialistes sur les quantités L1, L2 et L3 sont importantes. D'un autre côté, l'Internet est un univers en expansion qui atteint des chiffres impressionnants (et dans certains cas impossibles de déterminer) en nombre de personnes connectées, en serveurs et encore plus en pages web et pour lequel, soit il n'y a pas toujours de sources disponibles, soit certaines sources prennent de grandes libertés avec les données. Dans une étude qui va pondérer les données démo-linguistiques avec des données propres à l'Internet pour construire de indicateurs, il convient donc d'être extrêmement prudent et de bien discerner les biais qui portent sur les données ou sur la méthode Ainsi il sera possible de valider la fiabilité de la méthode et/ou d'entrevoir les implications possibles des biais sur les résultats et les corrections possibles.

Ce chapitre expose tous les biais ou limitations qui peuvent être engendrés directement ou indirectement par les choix méthodologiques et/ou les sources et montre les contrôles qui ont été réalisés dans la manipulation des données pour garantir des résultats fiables même s'il est impossible d'atteindre la perfection dans ce domaine.

### 7.1 Limitations propres à la méthode

La méthode s'est consacrée à chercher, identifier, évaluer puis utiliser des indicateurs quantitatifs des langues dans l'Internet pour construire des indicateurs susceptibles de caractériser convenablement la présence des langues dans l'Internet. Les indicateurs sur les langues étant en nombre très limité, l'option d'utiliser des indicateurs sur les pays a été prise pour dépasser cette limite et parvenir à obtenir une vision assise sur un nombre plus grand de micro-indicateurs. Cette option propose un défi qui est celui de la transformation de données s'appliquant à des pays en données s'appliquant à des langues. Ce défi a été surmonté par la méthode de pondération des données entre la répartition des locuteurs des différentes langues par pays et des données sur l'Internet concernant les pays. Il convient ici de bien appréhender les critères de validité et les limites de cette méthode.

Une matrice a été constituée qui contient les 7500 langues et les 192 pays identifiées. Par une simple opération arithmétique de pondération il est ainsi possible de répartir les données par pays sur les langues parlées dans chaque pays (au prorata de leur présence) et d'obtenir par cumul des données linguistiques mondiales pour chacun des critères mesurés par pays.

Quelle est l'hypothèse qui soutient cette démarche et est-elle correcte? L'hypothèse implicite est que les données concernant l'Internet dans un pays donné se répartissent de la même manière dans toutes les langues parlées dans ce pays. Est-ce que cela correspond à la réalité ou au moins s'en approche?

Pour prendre l'exemple le plus évident, le pourcentage de personnes connectées à l'Internet dans un pays donné est-il indépendant de la langue maternelle de ces personnes? C'est bien entendu une hypothèse simplificatrice qui sous-entend, dans l'exemple choisi, que la fracture numérique nationale n'est pas corrélée à la langue. C'est probablement faux: il est très possible que les locuteurs de

certaines langues aient un taux de connexion à l'Internet inférieur ou supérieur à la moyenne nationale. Cela peut être le cas parce qu'ils représentent une population avec des données socio-économiques ou culturelles ou encore de niveau d'éducation différents. Cependant, c'est une approximation acceptable qui ne va pas introduire de déformation violente des résultats, à condition de ne pas l'appliquer sur des espaces ou applications qui pourraient se prêter à une aggravation de ce biais méthodologique (par exemple, on ne pourrait pas appliquer cette simplification pour une application qui consiste à se perfectionner dans une des langues concernées) et à condition également ne pas la considérer valide pour différencier la situation respective des langues à l'intérieur d'un pays.

Cette question ne concerne que les calculs pour les langues maternelles et sa portée est à l'échelle d'un pays; la question du traitement des L2, qui elle s'applique à l'échelle de tous les pays, est susceptible de déformation majeure par rapport à la réalité et mérite une autre analyse qui est faite dans le prochain chapitre.

L'autre question que l'on doit se poser est celle de la complétude des données. Que se passe-t-il si, pour un critère mesuré le tous les pays ne sont pas renseignés? Quelle déduction peut-on faire sur les données dans cette situation? Comment peut-on surmonter d'éventuelles limitations? Si l'on conserve la même équation générale, la technique de pondération adoptée traite les pays non renseignés comme si les valeurs correspondantes au micro-indicateur concerné était nulle dans ces pays et cela va pénaliser fortement les langues présentes dans ces pays. Si on réalisait la pondération seulement à partir de ce nombre limité de pays en ignorant les pays non renseignés cela reviendrait exactement au même résultat. La solution trouvée a été celle de l'extrapolation des données pour les pays non renseignés, soit directement au prorata direct de leurs taux de connectivités respectifs soit par seuil, avec la méthode des quartiles. Cette méthode résout convenablement la situation et on peut considérer que les biais introduits (dans le cas où des pays aient dans la réalité une valeur pour ce micro-indicateur très au dessus ou très au dessous de ce que l'extrapolation a calculé pour des raisons particulières tenant à la nature de ce micro-indicateur[53]) se compensent en moyenne et deviennent marginaux quant le nombre de cas traités est suffisamment grand.

Il convient maintenant de se poser la question des limitations ou des biais pour chacun des critères qui correspondent aux micro-indicateurs c'est à dire chacun des éléments du complexe mis en œuvre: langues, pays et sources.

## 7.2 Les langues.

### 7.2.1 Choix de la source pour le calcul des L1

Dans cette édition le projet Joshua a été la source principale pour les données démo-linguistiques en langue maternelle (L1). Dans une prochaine édition le choix devrait être fait d'utiliser les données de Ethnologue.

### 7.2.2 Le cas des L2

La méthode choisie dans ce rapport pour les L2 fait abstraction des pays et donc de la correction fine que procure la pondération des locuteurs par pays en fonction du pourcentage de personnes connectées dans chaque pays. Appliquer le même taux de croissance de L1 vers L1+L2 aux résultats des calculs basés sur des L1, sous-entend en fait la même hypothèse simplificatrice propre à la

[53] Il doit rester clair que parfois l'extrapolation pourra attribuer une note pondérée à la présence dans l'Internet pour des langues parlées dans un pays où réellement le critère n'a aucune présence.

méthode de transformation des données pays en données langue (les données concernant l'Internet dans un pays donné se répartissent de la même manière dans toutes les langues parlées dans ce pays), mais cette fois, non pas à l'intérieur d'un pays, mais à l'échelle de tous les pays. L'hypothèse simplificatrice dans ce cas est que le taux de connexion à l'Internet des personnes d'une langue maternelle donnée (qui est le résultat d'une pondération calculée entre tous les pays où cette langue est présente) est également valide pour les langues secondes dans tous les pays où elles ont des locuteurs. Si pour une langue donnée les personnes de langue seconde (L2) se trouvent majoritairement dans des pays avec des pourcentages faibles de personnes connectées alors que les personnes de langues maternelles (L1) se trouvent dans des pays avec des pourcentages forts, alors la méthode va surestimer la population L1+L2 connectée à l'Internet pour cette langue.

Dans le cas imaginaire d'une langue L qui aurait 100 millions de locuteurs L1 dans un seul pays P1, avec un pourcentage de connectivité à l'Internet de 80%, et que les 100 millions de personnes de langue seconde résidaient toutes dans un seul pays P2 avec un taux de connectivité de 40%, alors l'erreur de la méthode serait de surestimer de 50% le nombre de personnes connectées.

Nc = nombre de personnes connectées calculé
Nr = Nombre de personnes connectées réelles

Nc(L1) = Nr(L1) = 100 millions x 80% = 80 millions
Nc(L1+L2) = 80 millions x 2 = 160 millions
Nr(L1+L2) = 100 millions x 80% + 100 millions x 40% = 120 millions
Taille de l'erreur = 40 millions soit 66%

Il s'agit bien sûr d'un cas totalement imaginaire et farfelu pris pour les besoins de pédagogie et la réalité ne comporte pas d'écarts aussi violents.

Il serait possible cependant de palier à cet inconvénient potentiel en utilisant la méthode des quartiles s'il était possible de séparer les populations L2 en sous-groupe en fonction du critère de connectivité à l'Interne:

- P0: la proportion qui résident dans des pays où moins de 15% de la population connectée : premier quartile, (Q1= note la plus basse des % de personnes connectées: 0,08 avec les valeurs actuelles)
- P1: la proportion qui résident dans des pays où plus de 15% mais moins de 35% de la population connectée : deuxième quartile (Q2= 0,34)
- P2: la proportion qui résident dans des pays où plus de 35% mais moins de 65% de la population connectée : troisième quartile (Q3= 0,56)
- P3: la proportion qui résident dans des pays où plus de 65% mais moins de 85% de la population connectée : quatrième quartile (Q4=0,76)
- P4: la proportion qui résident dans des le pays où plus de 85% de sa population connectée : dernier quartile (note la plus haute: Q5= 1).

et la formule pour recalculer le taux de croissance serait:

$$Tr(j) = (T(j) \times P_{L1}(j) / \sum_{i=0}^{i=4} Pi \times Qi)$$

où Tr(j) est le taux recalculé pour la langue j et $P_{L1}(j)$ est le % de personnes connectées en L1 pour la langue j.

Dans l'exemple du français, avec L2 = 153 485 770 et T= 3 et $P_{L1}$= 0,81, si la répartition était de 20% par quartile, alors le taux de croissance appliqué aux résultats serait recalculé de la manière suivante:
Tr(français) = 3 x 0,81 / (20% x 0,08 + 20% x 0,34 + 20% x 0,56 + 20% x 0,76 + 20% x 1 = 0,548)
Le résultat baisserait alors à 2.

Il est bien sûr possible aussi de simplifier cette méthode et de répartir seulement en 2 catégories, une haute et une basse, tout dépend en fait de la clarté que l'on peut avoir sur une idée plus ou moins précise des pays de résidence des locuteurs L2.

La qualité des résultats seraient certainement mieux garantie avec l'application cette méthode simple. Dans l'attente de cette méthode plus fine, il est raisonnable de considérer que la méthode pourrait provoquer une surestimation des résultats de l'anglais et du français qui ont des populations L2 importantes dans des régions de moindre taux de connectivité que respectivement les États-Unis (ou le Royaume Uni) et la France (ou la Belgique).

Cela amène à une question d'importance, celle de **la sensibilité du facteur L2** dans les résultats du français. Si on remplaçait la valeur du rapport L1+L2/L1 qui est dans cette version légèrement supérieure à 3, par la valeur de 2, ce qui reviendrait à réduire le pourcentage de personnes connectées francophone dans un rapport important (il passerait de 5,6% à 3,8% ), le français resterait la 4ième langue avec une quote-part qui passerait de 7,7% à 5,8% et il conserverait un écart encore important sur les suivants. Avec un changement à 1,5 de ce ratio (juste une hypothèse pour vérifier la sensibilité) le pourcentage de francophones connectés passerait à 2,9% et le français resterait encore en 4ième position avec une quote-part de 4,8%. Mais dans ce cas le pourcentage qui le séparerait du suivant serait alors réduit à 0,01%.

Cet exercice de simulation permet d'affirmer avec sécurité que la position favorable du français en quatrième position est extrêmement solide.

### 7.2.3 Réduction du nombre de langues

Le principe fondamental de l'indépendance des traitements par rapport au nombre de langue a été vérifié ce qui garantit l'invariance des résultats quelque soit le nombre langues traitées (à condition bien sûr de bien confectionner la ligne "Reste des langues"). Il est donc théoriquement possible de traiter l'ensemble des langues avec la méthode. Dans la pratique cela représente un volume de calcul impressionnant (7500 x 200 = 1 million et demi de cellules dans l'onglet LOC1) et des fichiers de volumes lourd à gérer (plus de 100 MB) donc un coût marginal gigantesque pour un bénéfice marginal nul, car seulement 500 langues parmi les 7500 ont un existence virtuelle. De plus les hypothèses simplificatrices vont invalider les résultats pour les langues avec des quantités de locuteurs qui ne se mesurent pas en millions.

Après plusieurs essais le choix s'est finalement arrêté avec la liste des langues ayant plus de 5 millions de locuteurs (140 langues).

### 7.2.4 Vérification de l'invariance des calculs par rapport au nombre de langues

Il était bien entendu indispensable de bien s'assurer que, comme l'intuition le dictait, les opérations arithmétiques réalisées pour les transformations pays ---> langues conservaient, étant de simples pondérations, les mêmes valeur quelque soit la taille de la matrice. Cela a été vérifié plusieurs fois en

passant de 7500 à 390, 100, 28, 85 langues et pour la liste des langues de plus de 10 millions de locuteurs puis finalement de plus de 5 millions. Ces tests en permis parfois de découvrir des erreurs dans les données et ont représenté un bon investissement pour la fiabilité des résultats.

### 7.2.5 Le cas des langues locales de France

La liste retenue ne comporte pas les langues locales de France. Parrni les essais réalisés il y a eu des échantillons avec certaines langues locales de France et des observations ont pu être faites. Est ce que ce travail permet de tirer des enseignements comparatifs précis entre les différentes langues de France? La réponse est catégoriquement: non! L'hypothèse simplificatrice sur laquelle repose la méthode ne permet pas de différencier entre elles les langues dont la présence est totale où extrêmement forte dans un seul pays. En effet, par définition, elles se voient toutes attribuées la valeur de la moyenne nationale en termes de personnes connectées. Si une langue était présente à 100% sur le territoire français (c'est à dire que tous ses locuteurs sont comptabilisés en France) les résultats en terme de capacité (c'est à dire indépendants du nombre de locuteurs) correspondants aux indicateurs par pays seraient exactement ceux de la France. Cela donnerait à cette langue locale un avantage par rapport aux autres langues qui auraient des locuteurs dans des pays avec des notes moindres. L'hypothèse simplificatrice qui sous-tend la méthode ne permet pas de différencier les pourcentages relatifs de personnes connectées à l'Internet par langue a l'intérieur d'un pays, elle n'est donc pas assez fine pour que la comparaison entre langues locales de France fasse sens et il faut se garder de tirer des conclusions hâtives sur les différences de notation entre les langues locales. Cependant, pour les données par langues, Wikipédia étant l'indicateur essentiel, la différentiation entre langues locales pourrait faire sens (mais seulement pour cet indicateur) lequel ne l'oublions pas toutefois, élimine avec des notes nulles, les langues absentes de Wikipédia[54]. Pour une vision plus fine de la place des langues de France dans l'Internet, le lecteur est invité à consulter le travail réalisé[55] par la même équipe pour la Délégation générale au français eu langue de France (DGLFLF[56]), qui va bientôt être proposé sous la forme d'une base de données publique.

## 7.3 Les pays

Après quelques itérations sur des questions de principe, la liste finale des pays à considérer a été arrêtée sur une base de simplification (suppression de certains pays avec très peu d'habitants et/ou pour lesquels l'UIT ne fournit pas de données, regroupement d'états satellites dans les pays principaux).

La possibilité d'utiliser la source essentielle de ce travail: les données fournies par l'UIT de pourcentage de personnes connectées à l'Internet par pays est un facteur essentiel à la fois du choix et des conséquences à en tirer. Une fois prise cette décision les données de l'UIT ont été modifiées pour tenir compte des changements dans la liste des pays retenue, en rattachant les données de population et en modifiant les données de pourcentage de personnes connectées, au prorata. Un bon exemple est celui de la Chine qui a un pourcentage de connexion à l'Internet bien inférieur à celui de Hong Kong ou Macao et donc l'intégration de ces deux états dans les données de la Chine a été faite au prorata de ces valeurs. Certains pays ont été supprimés (comme par exemple la Corée du Nord, Saint Marin ou

[54] Ce qui est le cas des créoles francophones à l'exception du créole haïtien mais pas celui de beaucoup d'autres langues locales comme le corse ou le breton.

[55] http://www.culturecommunication.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/La-diversite-linguistique-et-la-creation-artistique-dans-le-domaine-numerique/Etude-sur-la-place-des-langues-de-France-sur-l-internet

[56] http://www.culturecommunication.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France

les Malouines), d'autres ont été rattachés à un autre pays (comme par exemple les Iles Féroé et le Groenland au Danemark). Le Royaume Uni est le pays qui a fait l'objet du plus grand nombre de rattachements. Pour quelques rares pays importants l'UIT ne fournit pas de données et les locuteurs de ces pays ne seront pas comptabilisés: c'est le cas de la Corée du Nord (qui est très peu connectée à l'Internet), du Sahara occidental, du Kosovo et du Soudan du Sud. Enfin il y a une liste de petits états qui n'ont pas été pris en compte, soit parce que l'UIT ne fournit pas de données (en italique) soit parce que leur population est extrêmement limitée: *Bonaire, Saint-Eustache et Saba, Curaçao, Île Christmas, Îles Cocos, Île de Man, Île Norfolk, Îles Pitcairn,* Malouines, *Mariannes du Nord, Mayotte, Nauru, Palaos, Saint-Barthélemy,* Saint Marin, *Saint-Martin, Saint Pierre et Miquelon, Svalbard et île Jan Mayen, Territoire britannique de l'océan Indien, Vatican (État de la Cité du Vatican).*

TABLE 19 **Pays de rattachement**

| PAYS DE RATTACHEMENT | LISTE DES PAYS RATTACHÉS |
|---|---|
| Chine | Hong Kong et Macao |
| États-Unis | Guam, Îles Vierges américaines, Porto Rico et Samoa américaines |
| Danemark | Groenland et Iles Féroé |
| France | Guadeloupe, Guyane, Martinique, Nouvelle Calédonie, Polynésie, la Réunion, et Wallis et Futuna |
| Nouvelle Zélande | Iles Cook, *Niue et Tokelau* |
| Pays-Bas | Aruba, *Antilles néerlandaises* |
| Royaume Uni | Anguilla, Antigua-et-Barbuda, *Ascension*, Bermudes, Gibraltar, Guernesey, Îles Caïman, Îles Vierges, Jersey, Montserrat, *Saint Hélène*, Saint-Christophe-et-Niévès et Turks-et-Caïcos |

## 7.4 Les sources

### 7.4.1 Principes de base

Les trois principes suivant ont été respectés en règle générale:

1. aller systématiquement à la source première[57];
2. préférer toujours les sources les plus récentes;
3. ne pas se responsabiliser pour corriger les approximations détectées dans les sources (sauf exceptions décrites plus loin).

Le premier principe s'applique bien entendu aux sources qui offrent le panorama mondial par pays, pas aux sources individuelles pour chaque pays (cela reste le travail de la source première qui a composé le panorama par pays). La tenue du premier principe implique souvent le deuxième principe mais il doit rester clair que les calculs peuvent combiner des sources de dates différentes à l'intérieur d'un panorama par pays mais aussi entre les éléments mis en équation (voir plus loin).

### 7.4.2 Exceptions aux principes de base

Il existe cependant quelques cas pour lesquels l'amélioration des sources premières s'est imposée parce qu'elles constituaient des données essentielles dans tout le processus: c'est le cas des données démographiques par pays et des données de pourcentages de personnes connectées par pays[58]. Le deuxième cas est décrit en détail dans l'onglet UIT.

[57] En d'autres termes, toujours préférer les sources premières plutôt que les sources faisant référence aux sources premières.

[58] Dans le second cas il ne s'agissait pas de changer les valeurs proposées par l'UIT mais seulement de tenir compte correctement du découpage par pays

Une autre exception a été faite dans le cas des données d'abonnement à Netflix offertes par Statista. Il y a deux séries: les données actuelles et une projection à 2021. La découverte, dans la projection à 2021, d'une diminution du nombre d'abonnés aux États-Unis, pays de loin leader dans cette application, a fait vérifier dans la source que les États-Unis n'étaient pas renseignés dans la projection à 2021. L'extrapolation a pu limiter les dégâts mais le résultat restait peu satisfaisant avec une diminution des abonnements entre 2016 et 2021 aux États-Unis. Cette situation aller entrainer l'élimination de cette source, sauf à obtenir une donnée fiable par ailleurs sur la croissance des abonnements aux États-Unis dans ce marché, ce qui fut obtenu grâce à l'assistance de Statista. L'échantillon de données avec la projection pour 2021 a ainsi été complémenté avec une valeur pour la projection des abonnements aux États-Unis de manière à conserver une donnée exploitable.

### 7.4.3 La question des dates

La question des dates des sources est complexe; il est commun que certaines sources utilisent des dates différentes selon les pays, une situation rendue obligatoire par les différences de niveau d'actualisation des sources dans les pays. La présence dans la même opération arithmétique de données de sources avec des dates différentes ne représente pas un obstacle, même si cela peut avoir un impact marginal sur les résultats. Ainsi les données en terme de taux de connectivité à l'Internet par pays sont de 2015 mais les données démographiques par pays peuvent être de dates distinctes; il faut en être conscient lors de la manipulation des données mais cela ne les invalide pas.

Dans le cas des sources qui combinent plusieurs sources extérieures pour construire des indicateurs composites, il est malheureusement fréquent qu'elles n'utilisent pas les données les plus actualisées; se charger de cette actualisation reviendrait à actualiser ces sources composites ce qui sort du cadre de cette étude.

C'est le cas de l'indicateur "Government online services index" que WebIndex utilise pour créer des indicateurs composites qui donne la France en première position avec la note de 100. La source première a été vérifiée tardivement; cette vérification a permis de découvrir que WebIndex utilise une source datée de 2013 alors que la source première[59] offre des données de 2016 où la France est passée à la cinquième place avec une note de 94,20. Les résultats ne tiennent pas compte de cette évolution qui a un effet tout à fait marginal sur les résultats (voir plus loin).

### 7.4.4 La question du sens de la transformation pays > langues

Si le sens de cette transformation est intuitivement évident pour des données quantitatives par pays il peut devenir plus difficile à saisir, voire incompréhensible pour des pourcentages qui ne sont pas exprimés par pays.

L'exemple qui fait le plus de sens est celui du nombre d'internautes par pays. Que cette donnée soit exprimée en terme de quantité par pays ou en terme de pourcentage de personnes connectées par pays (il s'agit de la source première offerte par l'UIT) ou encore en terme de pourcentage d'internautes dans un pays par rapport aux total des internautes dans le monde, de simples calculs arithmétiques permettent de passer d'une valeur à l'autre. Le produit matriciel (langues x pays) avec quantité par pays va donc permettre par simple calcul arithmétique d'obtenir le nombre d'internautes par langues (ou le pourcentage de personnes connectées pour chaque langue ou encore le pourcentage de

[59] https://publicadministration.un.org/egovkb/en-us/reports/un-e-government-survey-2016

personnes connectées par langue par rapport au total des internautes, selon le mode de calcul sélectionné[60]).

Dans les cas de données exprimées en terme de pourcentages nationaux, il devient parfois très difficile de donner un sens à la transformation par langue ou encore impossible de trouver une technique d'extrapolation pour ce micro-indicateur (les deux étant probablement liés). Dans ces cas, l'option choisie a été de renoncer à cette transformation et donc à l'intégration de ce micro-indicateur dans la confection des indicateurs. C'est la cas par exemple du pourcentage de requête d'effacement vers Google (droit à l'oubli) en provenance d'un pays. L'extrapolation n'est pas possible car la donnée qui est liée à la culture juridique du pays (et à l'existence ou non d'une législation) n'a aucun rapport avec le pourcentage de personnes connectées dans les pays. Le sens ramené à la langue de cet indicateur serait, de plus, une source de confusion car on pourrait en déduire que l'effacement concerne des données dans la langue ramenée par le calcul alors que si c'est en France il concerne surtout le français. Si l'on cherchait à donner un sens strict et cohérent à la transformation par langue dans ce cas, cela donnerait "*pourcentage des locuteurs de telle langue faisant une requête d'effacements à Google dans la langue nationale de son pays de résidence*"... ce qui n'est ni d'une compréhension évidente ni d'un intérêt évident!

### 7.4.5 Limitations dues aux sources

Elles tiennent bien sûr d'abord aux limites et possibles biais des données à traiter. Dans le cas de l'indicateur essentiel concernant les personnes connectées à l'Internet par pays, l'analyse de la méthode utilisée par la source[61], l'UIT, montre :

- qu'il s'agit de l'indicateur HH7: *Proportion de particuliers utilisant l'Internet* avec description précise de la méthode et de ses limitations[62];
- que les valeurs ont été établies à partir de sondages de population;
- que la définition établit qu'il s'agit de personnes se connectant à l'Internet, quel que soit le dispositif utilisé (qui n'est pas forcément un ordinateur et peut être un téléphone mobile, une tablette, un assistant numérique personnel, une console de jeux, un téléviseur numérique, etc.) et que l'accès peut se faire par le réseau fixe ou mobile;
- que le critère de comptabilité adopté est celui de personnes de 16 à 75 ans qui se sont connectées au moins une fois dans les 3 derniers mois.

Une analyse plus fine montre que:
- environ 60% des sources sont des estimations propres de l'UIT et 15% sont des données fournies par Eurostat en utilisant les mêmes définitions et critères[63];
- par contre, pour le reste des pays la donnée est fournie par différentes autorités, selon les pays. Les autorités en question utilisent souvent des critères différents; ainsi, la fourchette d'âge peut démarrer, selon le cas, à 3, 5, 6, 7, 10, 12 , 15 ou 20 ans et il peut parfois ne pas y avoir d'âge limite;
- pour le cas de la France les données de la Martinique, la Guadeloupe, la Guyane et la Réunion et de sont incluses dans celle de la France (avant 2010 elles étaient estimées par l'UIT), ce qui n'est pas le

[60] La feuille LOC2 a été utilisée pour présenter ces différentes option de manière à obtenir une lecture aisée
[61] http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals_Internet_2000-2015.xls
[62] Manuel sur la mesure de l'accès des ménages et des particuliers aux technologies de l'information et de la communication (TIC) et de l'utilisation de ces technologies, 2014 (voir Page 74) - https://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ITCMEAS-2014-PDF-F.pdf
[63] http://ec.europa.eu/eurostat/statistics-explained/index.php/Information_society_statistics_-_households_and_individuals/fr

cas de la Polynésie française, de la Nouvelle Calédonie et de Wallis et Futuna et de Saint Pierre et Miquelon.

L'UIT reste la source la plus fiable de loin pour ce type de donnée et souvent d'autres sources l'utilisent sans la citer et parfois en procédant à des aménagements non documentés. La donnée commentée dans cet exemple étant, de plus, un élément essentiel dans les traitements qui vont être réalisés, un soin extrême a été pris pour bien prendre en compte tous les facteurs et en particulier la répartition des chiffres exacts dans le cas de pays avec des territoires associés.

Quels sont les biais possibles des données UIT? Les pays européens, pour lesquels l'UIT a repris les données de Eurostat, et les pays pour lesquels l'UIT a placé ses propres estimations peuvent être considérés comme renseignés de manière particulièrement fiable; par contre il est à craindre une légère surestimation de certains des pays renseignés par les autorités locales, soit parce que les critères de définition sont plus large, comme déjà mentionné, soit aussi parce que ces pays pourrait avoir fait une estimation généreuse pour montrer le bien fondé de leurs politiques de lutte contre la fracture numérique. Il est bon de connaitre ce biais potentiel mais il n'est pas rédhibitoire et peut être considéré en général comme tout à fait marginal.

Un des facteurs de décision du découpage finalement adopté pour les pays a été dans la manière choisie par l'UIT pour présenter ses données pays de manière à obtenir la meilleur cohérence dans les calculs. Chaque fois que cela a été possible les données concernant les territoires rattachés à des pays ont été soigneusement rajoutés au pays de rattachement, et la pondération les prendra correctement en compte (voir l'onglet UIT). Quelques rares pays importants n'ont pas de données pour les internautes de la part de l'UIT et les locuteurs de ces pays ne seront pas comptabilisés dans les calculs (voir le *chapitre 7.3 Pays*).

### 7.4.6 Biais potentiels de Alexa.com et W3Techs

La majorité des micro-indicateurs sont des données de trafic vers des sites web pour lesquels Alexa.com a été utilisé comme source. Quels sont les biais potentiels de cette source?

Alexa fournit les statistiques des usages du web grâce à une barre d'outils qu'un supposé large échantillon d'internautes[64] a accepté d'installer sur l'application de navigation. Alexa comptabilise les accès aux sites qui lui sont rapportés par cette barre d'outils et à partir de cette comptabilité effectue un classement des 25 millions de sites les plus fréquentés de la toile. Alexa, comme la plupart des entreprises de l'Internet, est très peu transparent sur sa méthode et les biais qui peuvent en résulter; en particulier, il n'existe malheureusement pas de statistiques sur les millions de barres d'outils installées qui permettraient de vérifier leur répartition par pays ou par langue.

Bien entendu, Alexa ne rapporte des données de trafic que pour les sites qui font partie de la partie haute de ce classement. La sélection de sites pour l'étude est faite à partir de sites bien placés dans ce classement (en général avec un classement Alexa inférieur à 10,000 sauf pour certaines thématiques moins fréquentées pour lesquels des classements jusqu'à 50,000 ont pu être sélectionnés). Les biais qui peuvent découler de la méthode Alexa sont les suivants:

- une concentration d'attention sur les sites les plus visités (par définition);

[64] Alexa reste flou avec la mention "de millions de barres d'outils installés" (http://www.alexa.com/about) et un article Wikipédia mentionne le chiffre de 10 millions en 2005 (https://en.Wikipédia.org/wiki/Alexa_Internet#Toolbar).

- une sous-estimation de sites dans des pays où la barre d'outils Alexa a été peu installée (en particulier une forte sous-estimation du trafic intérieur chinois car 80% des sites ne sont pas atteignables par le système de nommage de l'ICANN[65]);

- une surestimation relative des trafics en provenance des pays qui ont le plus installé la barre de Alexa (en toute probabilité, plus particulièrement les pays occidentaux et encore plus particulièrement les États-Unis[66]).

Quelles sont les conséquences de ces biais sur les résultats de l'étude?

Il y a une surreprésentation des pays occidentaux et indirectement des langues occidentales par rapport au reste des langues et une surreprésentation de l'anglais par rapport à l'ensemble des langues, cela dans une très forte proportion par rapport aux langues non occidentales.

Une étude comparative faite sur les résultats de Facebook, Twitter, YouTube, Netflix et Linkedin[67] confirme ces hypothèses.

L'étude compare les données en termes d'abonnements pour ces applications avec les données en termes de trafic fournies par Alexa. Les écarts entre les différents résultats (après avoir ramenés en pourcentages les quantités) sont relativement importants et, ce qui était prévisible, en général, en faveur des données de trafic de Alexa dans le cas des langues occidentales. Les écarts entre les pourcentage d'abonnés et les trafics se révèlent extrêmement importants pour la plupart des pays non occidentaux.

Les écarts sont moindres pour Linkedin ce qui est logique: étant un réseau social professionnel les abonnés passeront moins de temps à flâner et la sensibilité du temps passé par rapport au prix de connexion sera moindre.

A noter que le portugais se comporte comme les langues non occidentales dans ces comparaisons et que ce cas n'a pas encore était élucidé (existe-t-il une raison spécifique au Brésil qui ferait que le téléchargement de la barre d'outils de Alexa ne se réalise pas ou bien que les temps de connexion aux différents services soient beaucoup plus courts à cause des tarifs?). Cette étude fait ressortir des scores avantageux pour le français, ce qui n'est pas le cas de l'espagnol et encore moins du finnois (ce dernier cas montre probablement que la technique d'extrapolation ne parvient pas toujours à proportionner les niveaux réels).

Le tableau suivant montre le résultat de la division des valeurs proposées par Alexa (en termes de trafic) par les valeurs proposées par l'intermédiaire de Statista (en termes d'abonnements).

[65] Selon des experts de OpenRoot cela concerne 80% des sites chinois ce qui expliquerait d'ores et déjà une sous-estimation des contenus (ou du trafic) chinois dans un facteur 5!
[66] Cette affirmation provient du simple raisonnement et ne peut être soutenue par des données concrètes en l'absence de données statistiques concernant cette barre d'outils; Cependant il sera montré plus loin comment les statistiques de contenus par langue de W3Techs qui s'appuient sur le classements de Alexa permettent d'induire ce diagnostic.
[67] Voir le fichier Contrôles en pièces jointes.

TABLE 20 **Ratios de surestimation/sous-estimation trafic/abonnés**

| | Facebook | Twitter | Youtube | Linkedin | Netflix | MEDIANE |
|---|---|---|---|---|---|---|
| **Anglais** | 3,11 | 2,65 | 1,72 | 1,63 | 1,19 | 1,72 |
| **Espagnol** | 0,80 | 0,72 | 0,81 | 0,62 | 1,32 | 0,80 |
| **Français** | 2,81 | 4,40 | 1,78 | 2,49 | 1,41 | 2,49 |
| **Portugais** | 0,33 | 0,19 | 0,40 | 0,27 | 2,25 | 0,33 |
| **Finnois** | 0,41 | 0,35 | 1,03 | 0,41 | 0,09 | 0,41 |
| **Chinois** | 0,55 | 0,23 | 0,24 | 0,44 | 0,23 | 0,24 |
| **Hindi** | 0,21 | 0,19 | 1,46 | 0,74 | 0,21 | 0,21 |
| **Arabe** | 0,51 | 0,48 | 1,13 | 0,57 | 0,50 | 0,51 |

Un facteur de 1 indiquerait l'équivalence des résultats, le résultat est d'autant supérieur à 1 que les valeurs de proposées par Alexa sont supérieures à celles de Statista.

L'explication des différences entre les données comparées est due à la conjugaison de deux facteurs:

1) Celui, déjà mentionné: Alexa va favoriser nettement les pays occidentaux comme un reflet naturel du fait que la barre d'outils y est plus représentée.

2) Mais aussi un facteur additionnel doit jouer qui est lié au temps moyen de connexion des utilisateurs, lequel, pour des raisons de tarif, doit être en général supérieur dans les pays industrialisés.

Cependant, le deuxième facteur ne peut expliquer à lui seul la taille des différences constatées et c'est sans aucun doute le premier facteur, le biais de Alexa en faveur de l'anglais[68], qui en est le principal responsable.

La place du portugais pose question. Le Brésil et le Portugal sont dans le milieu du classement mondial en termes de prix de connexion à l'Internet[69] mais le Brésil est en queue pour les vitesses de connexion[70]. Si l'on regarde l'ensemble des facteurs pour le portugais on voit bien que celui du trafic est singulier (les 5 autres facteurs dans la fenêtre, centrée sur 4%, entre 3% et 5% alors que le trafic est en dessous de 2%). A noter que si l'on simulait une note de 4% pour ce facteur cela améliorerait le score du portugais mais ne modifierait pas le classement.

La conclusion générale à tirer de ces comparaisons est que les résultats pour l'indicateur "trafic" vont exagérer les valeurs pour les langues occidentales et à l'intérieur de ces langues, pour l'anglais[71]. Il faut garder en perspective cette réalité lors de l'utilisation des résultats. De quel ordre de grandeur est le décalage au détriment des langues non occidentales?

Pour en avoir une idée, il faut consulter W3Techs qui en utilisant une méthode systématique de détection des langues appliquée sur les 10 millions de sites les mieux classés selon Alexa, va

[68] Avec les scores très haut de l'anglais dans les classements de Alexa, une simple facteur de surestimation de 1,7 comme dans le cas de Linkedin, représenterait plus de 20% de surestimation à répartir sur les autres langues, ce qui est énorme.
[69] https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=33
[70] https://en.Wikipédia.org/wiki/List_of_countries_by_Internet_connection_speeds
[71] Et peut être également pour le français à en juger par les tests de comparaison réalisés.

naturellement refléter ces décalages dans son résultat[72]. En réalité, W3Techs va non seulement refléter ces décalages mais également les amplifier légèrement car d'autres facteurs vont s'additionner:

- les algorithmes de détection de langue ont eux aussi un biais en faveur de l'anglais, quoique cela pourrait être un facteur marginal de l'ordre de 10%;

- le choix de la détection de la langue sur la page d'entrée des sites (home page), qui est souvent multilingue, avec l'anglais comme langue additionnelle, va augmenter substantiellement le score de l'anglais (là on peut craindre une surestimation supérieure à 1/3).

Déterminer en conséquence quelle est la part de chacun des facteurs dans le biais est un exercice très difficile; cependant il existe un indicateur qui va permettre d'appréhender la situation jusqu'à un certain point. Lors des études réalisées par Funredes et l'Union Latine pendant la période 1998-2007 le ratio entre pourcentage de contenus et pourcentage d'internautes pour une langue avait été étudié et suivi de près. Ce ratio qui indique la productivité dans la création de contenus dans une langue donnée possédait une certaine stabilité et cohérence dans le temps. Une très forte corrélation quasi linéaire entre les deux facteurs avait été observée ce qui se traduisait par des indicateurs très clairement centrés autour de 1. Les écarts étaient de l'ordre de plus au moins 50% selon que la langue faisait l'objet d'un effort important de production de contenus (facteur 1,5) ou au contraire qu'il y avait faiblesse de ce côté (facteur 0,5). Il avait été également remarqué que cette note faiblissait avec le temps, les nouveaux internautes semblant moins enclins à produire des contenus ce qui impliquait une baisse de cette productivité après les premières phases de développement dans l'Internet.

Voici les valeurs de cet indicateur en 2005 et 2008 (source FUNREDES[73]):

| | **Français** | **Anglais** | **Espagnol** | **Portugais** | **Allemand** | **Italien** | **Roumain** | **Catalan** |
|---|---|---|---|---|---|---|---|---|
| **2005** | 1,14 | 1,57 | 0,66 | 0,55 | 1,06 | 0,81 | 0,70 | |
| **2008** | 0,87 | 1,42 | 0,43 | 0,34 | 1,16 | 0,98 | 0,66 | 0,74 |

On voit que cet indicateur restait dans la fenêtre 0,3 - 1,5. Voici maintenant, à titre de comparaison les mêmes données aujourd'hui, en utilisant les données de l'UIT (converties en termes de langue par l'étude) pour les internautes et celle de W3Techs pour les contenus.

| | **Français** | **Anglais** | **Espagnol** | **Portugais** | **Allemand** | **Italien** | **Roumain** | **Catalan** |
|---|---|---|---|---|---|---|---|---|
| **2017** | 0,71 | 2,34 | 0,53 | 0,62 | 1,72 | 2,41 | 0,92 | 0,44 |

Les résultats pour l'anglais, l'allemand et l'italien ne sont pas crédibles. Ils montrent une grande surestimation des valeurs attribuées par W3Techs aux contenus pour ces langues. Les deux tableaux qui suivent montre les mêmes indicateurs pour un échantillon d'autres langues.

| | **Chinois** | **Hindi** | **Arabe** | **Malais** | **Ourdou** | **Bengali** | **Coréen** | **Vietnamien** | **Thaï** |
|---|---|---|---|---|---|---|---|---|---|
| **2017** | 0,09 | 0,02 | 0,19 | 0,23 | 0,05 | 0,07 | 0,57 | 0,45 | 0,36 |

| | **Russe** | **Japonais** | **Polonais** | **Hébreu** | **Hongrois** | **Suédois** | **Turc** | **Ukrainien** | **Tchèque** |
|---|---|---|---|---|---|---|---|---|---|
| **2017** | 1,27 | 1,60 | 2,00 | 1,65 | 1,59 | 2,2 | 1,58 | 0,17 | 3,22 |

[72] https://W3Techs.com/technologies/overview/content_language/all
[73] http://funredes.org/lc

On constate que la sous-estimation de certaines langues atteint des proportions gigantesques (chinois, hindi, ourdou, bengali) ce qui invalide absolument les chiffres présentés par W3Techs pour ces langues ou encore trop grandes pour être crédibles (arabe, malais, ukrainien, thaï) tandis que d'autres sont clairement surestimées (italien, tchèque, suédois, polonais) .

Ces comparaisons montrent que W3Techs n'est pas crédible comme source sur les contenus. Il ne s'agit pas de douter de la capacité de cette entreprise de conduire ces mesures, mais de bien comprendre que deux biais interviennent pour fausser les résultats:

- ✓ Les biais propres à l'échantillon de Alexa qui sert de support aux mesures.
- ✓ Le biais propre à la méthode de W3Techs qui favorise l'anglais assez nettement en mesurant les pages d'entrée des sites et ne prenant pas en compte le multilinguisme.

L'analyse des valeurs montrent que si le biais propre à W3Techs existe et doit gonfler les chiffres de l'anglais dans une certaine proportion (celle des sites multilingues avec de l'anglais dans les pages d'entrée, peut être 30% de surestimation), le biais principal, et de très loin, est celui induit par Alexa. Dans une certaine mesure, W3Techs révèle les biais de Alexa! En effet, il semblerait qu'il y ait des pays dont la présence dans la base de Alexa soit sous-estimée dans un facteur de l'ordre de 500% : la Chine et l'Inde ont beau représenter ensemble plus d'un milliard de personnes connectées à l'Internet (c'est à dire près du tiers des internautes), ils n'auraient selon W3Techs que 6,63% des contenus de l'Internet. Des chiffres inférieurs à 0,1 montrent que le déficit est gigantesque; des chiffres entre 0,1 et 0,3 montrent qu'il reste énorme et peu crédible, enfin des chiffres supérieurs à 2 montrent un surplus difficile à croire. Il faut noter en passant que le français, l'espagnol, le portugais et le catalan font partie des mal loties par les statistiques de W3Techs avec des chiffres qui resteraient dans la zone de crédibilité mais qui ne correspondent pas aux positions que les résultats de cette étude permette de déduire.

Il existe une technique pour contrôler ce genre de résultats, utilisées avec fréquence dans les premières études de FUNREDES/Union latine dans la période 1998-2007 pour déterminer le poids de l'anglais[74], c'est celle qui tente de rendre cohérents les indicateurs pour le reste des langues. Si l'on considère les 15 premières langues classées dans cette étude, le pourcentage de personnes connectées restant est d'environ 35%. A cet ensemble de langues W3Techs attribue à peine un peu plus de 10% des contenus ce qui n'est pas cohérent. En partant de l'hypothèse logique d'une très forte surestimation des contenus en anglais par W3Techs (qui les place à 53%), de la forte sous-estimation de certaines langues asiatiques et de l'indicateur pour le reste des langues, il est possible de tenter, par recoupement, une approximation raisonnable... quoique **purement spéculative** de ce qui est le plus probablement la situation des contenus dans l'Internet.

Un des buts de cet exercice spéculatif est d'essayer de comprendre si les biais importants, qui portent sur des éléments sur lesquels reposent les indicateurs, sont susceptibles de modifier considérablement la quatrième place donnée au français. La réponse est non: le diagnostic de quatrième place du français est très solide: après tentative de correction des biais le résultat conserve le français en quatrième position même s'il redistribue les cartes derrière pour voir les langues de la Chine et de l'Inde prendre les places qu'elles méritent probablement.

[74] En effet, la méthode produisait des résultats en pourcentage relatif, par rapport à l'anglais, et la détermination du pourcentage absolu des contenus en anglais (qui permettait, en cascade, de déduire les autres pourcentages absolus) était une étape délicate qui ne pouvait se résoudre que par recoupement.

Le tableau suivant propose les données suivantes, par colonne:
1: Macro indicateur générique (quote-part)
2: Pourcentage de personnes connectées
3: Données proposées par W3Techs pour les contenus
4: Ratio des données de W3Techs par personnes connectées (division de la donnéno3 par la no2)
5: Capacité mesurée (1/7)
6: Capacité spéculative (8/7)
7: Pourcentage de la population mondiale
8: Spéculation sur le pourcentage de contenus
9: Productivité contenus résultant de la spéculation (8/2)
10: Spéculation sur le pourcentage de contenus ramené sur la base de 100%

On constate sur ce tableau que les productivités spéculatives de contenus ont été ramenées à des valeurs crédibles, en particulier pour le reste des langues. La spéculation est basée justement sur un équilibre raisonnable entre la productivité du reste et celle des langues mentionnés en maintenant l'ensemble de ces valeurs dans une fenêtre raisonnable entre 0,7 et 1,3 (sauf pour l'anglais qui peut atteindre une valeur supérieure). Il y a une assez forte sensibilité sur ces deux facteurs; ainsi si on passait l'anglais à 30% (sans changer le reste) la productivité du reste des langues dépasserait 1 ce qui ne correspond pas à ce qu'il faut attendre; si l'on passait à l'anglais à 40%, alors la productivité de l'anglais passerait à 1,8. Une valeur spéculative de 1/3 pour la proportion des contenus en anglais donnerait une productivité pour le reste supérieure à 0,9 ce qui est trop important. C'est ainsi que la spéculation s'est arrêté sur cette valeur pour l'anglais : 36%.

TABLE 21 **Classement spéculatif**

| | | 1 QP GÉN. L1+L2 | 2 % INTERN. L1+L2 | 3 W3T | 4 W3TECHS/ INTERN. L1+L2 | 5 CAP. L1+L2 | 6 CAP. Spéc. | 7 POP. MONDIAL | 8 CONTENUS SPEC. L1+L2 | 9 PRODUCT. CONT L1+L2 | 10 CONTENUS SPEC. /100% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Anglais | 0,404 | 0,226 | 0,519 | 2,299 | 3,74 | 3,33 | 0,108 | **0,360** | 1,59 | 29% |
| 2 | Chinois | 0,140 | 0,209 | 0,020 | 0,096 | 0,92 | 1,18 | 0,152 | **0,180** | 0,86 | 14% |
| 3 | Espagnol | 0,093 | 0,093 | 0,051 | 0,551 | 1,56 | 1,35 | 0,059 | **0,080** | 0,86 | 6% |
| **4** | **Français** | **0,076** | **0,056** | **0,041** | **0,726** | **2,93** | **2,31** | **0,026** | **0,060** | **1,06** | **5%** |
| 5 | Russe | 0,047 | 0,050 | 0,065 | 1,288 | 1,71 | 1,45 | 0,028 | **0,040** | 0,79 | 3,2% |
| 6 | Portugais | 0,038 | 0,041 | 0,026 | 0,641 | 1,49 | 1,36 | 0,026 | **0,035** | 0,86 | 2,8% |
| 6 | Allemand | 0,046 | 0,032 | 0,055 | 1,723 | 3,31 | 2,53 | 0,014 | **0,035** | 1,10 | 2,8% |
| 6 | Japonais | 0,033 | 0,034 | 0,056 | 1,633 | 2,34 | 2,47 | 0,014 | **0,035** | 1,02 | 2,8% |
| 9 | Hindi | 0,027 | 0,039 | 0,001 | 0,023 | 0,48 | 0,54 | 0,056 | **0,030** | 0,76 | 2,4% |
| 9 | Arabe | 0,032 | 0,042 | 0,007 | 0,165 | 0,86 | 0,79 | 0,038 | **0,030** | 0,71 | 2,4% |
| 11 | Malais | 0,021 | 0,027 | 0,060 | 2,262 | 0,80 | 0,77 | 0,026 | **0,020** | 0,75 | 1,6% |
| 12 | Coréen | 0,013 | 0,014 | 0,009 | 0,637 | 1,41 | 1,69 | 0,009 | **0,015** | 1,06 | 1,2% |
| 12 | Ourdou | 0,014 | 0,020 | 0,001 | 0,044 | 0,51 | 0,55 | 0,027 | **0,015** | 0,74 | 1,2% |
| 14 | Italien | 0,016 | 0,009 | 0,023 | 2,515 | 3,22 | 1,97 | 0,005 | **0,010** | 1,09 | 0,8% |
| 15 | Bengali | 0,008 | 0,013 | 0,001 | 0,069 | 0,29 | 0,37 | 0,027 | **0,010** | 0,76 | 0,8% |
| 16 | Polonais | 0,012 | 0,009 | 0,017 | 1,996 | 2,55 | 1,94 | 0,005 | **0,009** | 1,06 | 0,7% |
| | RESTE | 0,230 | 0,335 | 0,048 | 0,144 | 0,48 | 0,45 | 0,630 | **0,286** | **0,85** | 23% |
| | TOTAL | 1,250 | 1,250 | 1,000 | | | | 1,2500 | 1,2500 | | 100% |

La rareté de sources d'information sur les langues dans l'Internet a créé une situation préoccupante dans laquelle les seules sources existantes ne sont pas fiables mais sont reprises par une grande quantité d'interlocuteurs sérieux qui bâtissent des discours faussés par des prémisses biaisés dans lesquels la sous-estimation de la présence du français va de pair avec une surestimation grossière de la place de l'anglais dans l'Internet[75]!

### 7.4.7 Limitations/biais liés au degré de localité des sources

De nombreux internautes des pays occidentaux pourraient penser que des applications aussi célèbres que Wikipédia, Facebook, Google, Twitter ou YouTube sont des standards internationaux dont la célébrité et l'utilisation est égale dans tous les pays du monde. La réalité est différente: certains pays (comme par exemple la Chine ou la Russie) ont su développer des applications alternatives qui drainent une partie importante du trafic potentiel national vers ses grandes applications de l'Internet. C'est le cas par exemple de Vkontakte qui remplace Facebook en Russie ou de Badiu qui remplace avantageusement Google en Chine. Si on pouvait attribuer un qualitatif de degré de mondialité (ou inversement de degré de localité) aux applications dans l'Internet et il faut être conscient que ce critère peut varier de manière importante et qu'il est très difficile qu'il arrive à son horizon de 100%. Parmi les applications aujourd'hui les plus "mondiales" on trouve très probablement Twitter et certainement Wikipédia[76] (et ce n'est pas un hasard si elle est la plus diverse linguistiquement et de loin) et peut-être Google, avec certaines réserves, mais c'est moins le cas de Facebook, YouTube, ou Linkedin même si ces applications restent beaucoup plus "mondiales" que leurs concurrents asiatiques Vkontakte ou Badiu.

Dans ce contexte quelle serait alors la validité de l'analyse faite à l'aide de Alexa des différents sites? S'il est clair que chaque indicateurs-pays pour une application ou un espace donné doit être pris avec précaution car le classement du français, par exemple, pourrait être déterminé par un fort de degré de localité des sites sélectionnés, il reste néanmoins vrai que la multiplication des sites mesurés (plus de 300) permet d'obtenir une neutralisation de tels effets grâce à l'utilisation de la moyenne réduite à 20% qui permet d'éliminer les données en provenance de sites à forte localité. Le lissage de la courbe est par ailleurs amplifié par la technique d'extrapolation qui va fournir des valeurs pour les pays manquants en se déterminant seulement par rapport à des critères mondiaux.

A cet égard il est intéressant d'analyser la position du français par rapport aux centaines de sites web mesurés en trafic avec Alexa. A première vue, on note que la majorité des scores se répartissent dans une fenêtre entre 4% et 6%. Si on recherche les écarts importants par rapport à cette fenêtre médiane on va découvrir des sites qui ont, soit une forte localité francophone, soit une forte localité dans une autre langue.

[75] Voir, comme un exemple caractéristique, le rapport récent de Mozilla sur la *santé de l'Internet* qui porte une attention particulière aux aspects linguistiques, malheureusement à partir d'hypothèses biaisées (https://internethealthreport.org/v01/).

[76] Quoique une analyse plus profonde a montré qu'il faut tempérer cette évaluation car la représentation des grandes langues asiatiques dans Wikipédia est extrêmement en dessous de leur présence démographique.

TABLE 22 **Sites à forte localité francophone**

| **SITE** | **% TRAFIC MONDIAL pour français** |
|---|---|
| Archives-ouvertes.fr | 21 |
| Bnf.fr | 45 |
| Cairn.info | 41 |
| Deezer.com | 36 |
| Exalead.com | 36 |
| Fnac.com | 70 |
| Fun-mooc .fr | 48 |
| Gameblog.com | 69 |
| Openclassrooms.com | 52 |
| Qwant.com | 68 |
| Skyrock.com | 47 |
| Theses.fr | 51 |
| Viadeo.com | 51 |
| Yacy.com | 34 |

A noter que Dailymotion ne figure pas dans ce classement mais dans celui des sites à préférence francophone, ce qui prouve le caractère international de ce site français. Selon ce classement, le site le plus localisé dans la francophonie, parmi les sites sélectionnés, serait, après le libraire fnac.com (et malgré l'existence d'un fnac.fr), un site pour les communautés de joueurs (Gameblog), puis le moteur de recherche Qwant (le moteur de recherche Exalead par contre a gagné une assise plus large en dehors de la francophonie).

Pour les sites localisés dans d'autres communautés linguistiques, on constate que le degré de localité peut arriver jusqu'à des pourcentage entre 80 et 92%.

TABLE 23 **Sites à forte localité non francophone**

| SITE | % TRAFIC MONDIAL pour français | Langue dominante | % TRAFIC MONDIAL pour langue dominante |
|---|---|---|---|
| 4shared.com | 0,8 | Portugais<br>Arabe<br>Espagnol | 22<br>16<br>11 |
| Anobii.com | 0,3 | Italien | 38 |
| Baidu.com | 0,07 | Chinois | 88 |
| Cyword.com | 0,4 | Coréen | 78 |
| Douban.com | 0,3 | Chinois | 85 |
| Filmaffinity.com | 0,4 | Espagnol | 72 |
| Filmow.com | 0,7 | Portugais | 83 |
| Fotolog | 0,6 | Espagnol | 52 |
| Gigasize.com | 0,9 | Espagnol<br>Hindi | 21<br>17 |
| Gmx.net | 0,08 | Allemand | 57 |
| Icq.com | 1,1 | Russe | 31 |
| Ikiyi.com | 0,1 | Chinois | 88 |
| Jurn.org | 0,9 | Arabe | 13 |
| Kaixin001.com | 0,1 | Chinois | 85 |
| Kakao.com | 0,2 | Coréen | 84 |
| Mail.ru | 1,1 | Russe | 41 |
| Mamba.ru | 0,4 | Russe | 43 |
| Megaupload.com | 1,4 | Espagnol | 21 |
| Mouthshut.com | 0,2 | Hindi | 43 |

| | | | |
|---|---|---|---|
| Naver.com | 0,2 | Coréen | 86 |
| Nicovideo.jp | 0,05 | Japonais | 90 |
| Novoed.com | 0,09 | Hindi | 14 |
| Odnosklassniki.ru | 0,7 | Chinois | 75 |
| Playstore.com | 0,07 | Turc | 41 |
| Plurk.com | 0,3 | Chinois | 64 |
| Qq.com | 0,3 | Chinois | 79 |
| Rapidshare.com | 1,1 | Espagnol | 17 |
| Rediff.com | 0,3 | Hindi | 40 |
| Rediffmail.com | 1,1 | Hindi | 44 |
| Renren.com | 0,2 | Chinois | 84 |
| Sapo.pt | 1,1 | Portugais | 71 |
| Scielo.org | 0,5 | Portugais<br>Espagnol | 52<br>21 |
| Sogou.com | 0,05 | Chinois | 92 |
| Spaces.ru | 0,3 | Russe | 47 |
| Somuch.com | 0,4 | Hindi | 36 |
| Taringa.net | 0,4 | Espagnol | 78 |
| Tuenti.com | 0,05 | Espagnol | 65 |
| Weibo.com | 0,08 | Chinois | 89 |
| Xing.com | 0,08 | Allemand | 53 |
| Youku.com | 0,08 | Chinois | 89 |
| Zihu.com | 0,1 | Chinois | 86 |

Puis il y a des sites qui ne sont pas centrés sur la francophonie mais pur lesquels le score francophone s'éloigne sensiblement de la moyenne, marquant une certaine préférence.

**TABLE 24 Sites à préférence francophone**

| SITE | % TRAFIC MONDIAL pour français |
|---|---|
| Badoo.com | 7,9 |
| BaseSearch.net | 7,3 |
| Blackle.com | 8,9 |
| Codeanywhere.com | 7,6 |
| Dailymotion.com | 7,0 |
| Duckduckgo.com | 7,6 |
| Europeana.eu | 6,6 |
| Infinit.io | 9,1 |
| Ixquick.com | 8,2 |
| Ixquick.eu | 10,6 |
| Napster.com | 11 |
| Netlog.com | 11,2 |
| Openoffice.org | 11 |
| Periscope.tv | 6,3 |
| StartPage.com | 7,5 |
| Uploaded.net | 7,8 |
| Wetransfer.com | 8,1 |
| Yammer.com | 6,7 |

Pour terminer cette analyse, il reste à noter que les sites centrés sur la communauté anglophone n'atteignent jamais les 70%, signe de la capacité indéniable de l'anglais dans l'Internet a attiré plus loin que la langue ou, ce qui revient au même, de l'importance probable de la population Li des internautes qui n'ont l'anglais ni comme langue première, ni comme langue seconde mais comme une langue suffisamment maitrisée pour permettre la navigation dans l'Internet.

TABLE 25 **Sites à préférence anglophone**

| SITE | % TRAFIC MONDIAL pour anglais |
|---|---|
| Adam4adam.com | 50 |
| Adictinggames.com | 53 |
| Aim.com | 62 |
| Amazon.com[77] | 50 |
| Avvo.com | 67 |
| Beyond.com | 68 |
| Cafemom.com | 51 |
| Caringbridge.org | 68 |
| Chacha.com | 48 |
| Classmates.com | 67 |
| Commonsensemedia.org | 59 |
| Crunchyroll.com | 50 |
| Eharmony.com | 62 |
| Excite.com | 53 |
| Fetlife.com | 52 |
| Flixter.com | 50 |
| Gaiaonline.com | 56 |
| Gfycat.com | 55 |
| Influenster.com | 64 |
| Joinhouse.party | 57 |
| Justanswers.com | 53 |
| Match.com | 53 |
| Metafilter.com | 54 |
| Mocospace.com | 55 |
| Mylife.com | 66 |
| Oovoo.com | 51 |
| Ravelry.com | 57 |
| Rumble.com | 64 |
| Sharecare.com | 51 |
| Smugmug.com | 54 |
| Straightdope.com | 55 |
| Telegram.com | 54 |
| Yelp.com | 65 |

TABLE 26 **Répartition des sites à préférence linguistiques**

| LANGUE | NOMBRE DE SITES |
|---|---|
| Allemand | 2 |
| Anglais | 29 |
| Arabe | 2 |
| Chinois | 10 |
| Espagnol | 9 |
| Français | 15 |
| Hindi | 6 |
| Portugais | 3 |
| Russe | 4 |

Il reste à vérifier si la sensibilité de ce facteur sur les résultats. La sensibilité est forte si l'on utilise la moyenne simple ou la médiane. Par exemple, avec la moyenne le score spécifique trafic du français

[77] Conséquence la politique de domaines par pays comme amazon.fr.

est de 10,31 et le score générique puissance serait alors de 8,07. Si on supprime fnac.com la moyenne tombe à 10,00; si on supprime aussi gameblog.com la moyenne tombe à 9,69; si on supprime également Fun-mooc.fr, Openclassrooms.com, Qwant.com, Skyrock.com,Theses.fr et Viadeo.com alors la moyenne de l'indicateur trafic tombe à 8,09 et le score générique passe à 7,70. Cette analyse a également démontré que si la médiane n'est pas effective la moyenne non plus et qu'il faudrait écarter les valeurs avec les écarts les plus forts. La solution finalement adoptée a été pour l'indicateur du trafic celui de la **moyenne réduite** en supprimant les plus extrêmes de chaque côté (10% des valeurs extrêmes de chaque côté, ce que réalise la moyenne réduite à 20%).

Cette situation affecte aussi et d'autant plus l'effort de différentiation par thématiques; il ne sera appliqué qu'avec les données de trafic, qui sont les plus nombreuses, et les résultats ne seront pas pris en considération dans le cas des thématiques avec trop peu de micro-indicateurs car il ne ferait que refléter le biais de la sélection de sites.

### 7.4.8 A propos du principe de pondération

Si le principe de la pondération est facile à assimiler pour des concepts comme contenus, trafic, usages et surtout personnes connectées il peut devenir assez abscons, voire mystérieux, quand il est appliqué à des notations de 0 à 100, comme dans le cas des indexes, ou à des pourcentages relatifs à un total d'applications comme dans le cas des interfaces. Cela mérite donc quelques explications.

Par quel mécanisme une note de 0 à 100 attribuée à une langue comme le français, résultat d'une évaluation sur un critère donné, au prorata des pays où les francophones résident, peut-elle se transformer en pourcentage mondial et quelle est la signification de cette nouvelle valeur? Par quel mécanisme un pourcentage représentant le nombre de fois que le français est présent dans les 24 interfaces ou langues de traduction peut-il se transformer en pourcentage mondial et quelle est la signification de ce pourcentage?

Le mécanisme et la signification sont les mêmes dans les deux cas. Le mécanisme est celui de la pondération des valeurs obtenus après transformation des données de pays vers langues par le pourcentage de personnes connectées. Cela se calcule de cette manière:

$$P_{inter}(j) = Tg \times INTER(j) \times UIT(j) / Po$$

avec $$Po = \sum_{i=1}^{i=L} UIT(i) \times INTER(i)$$ (facteur de pondération)

où

$P_{inter}(j)$ est la pondération de l'indicateur interface calculé pour la langue j
Tg est le taux global de multilinguisme (pour normaliser les totaux sur L1+L2)
INTER(j) est le pourcentage trouvé pour la langue j pour les interfaces
UIT(i) le pourcentage de personne connectées pour la langue i.
L le nombre total de langue (y inclus le reste).

La signification de ce pourcentage mondial est celle de ramener la valeur des pourcentages d'interfaces au pourcentage des personnes connectées dans l'Internet par langue, avec des variations autour de cette valeur dues à la répartition de ces pourcentages pour l'ensemble des langues. Si cette valeur est exactement la moyenne pondérée des notations (pour les indexes) ou des pourcentages (pour les interfaces), le résultat serait identique au pourcentage de personnes connectées. Si cette

valeur est très supérieure à cette moyenne, alors le résultat sera de renforcer ce pourcentage; dans le cas contraire de le déprimer. Il faut bien comprendre que l'effet ne dépend pas seulement de la notation ou du pourcentage en question pour une langue donnée, mais en fait de la distribution des valeurs et de celle des pourcentages de personnes connectées pour l'ensemble des langues.

Pour développer une approximation plus intuitive de ce mécanisme, le mieux est de simuler la situation sur le cas le plus simple, celui des interfaces. Le français est présent dans 23 des 24 éléments mesurés ce qui lui donne un pourcentage de 23/24 = 95,83% qui porte vers le haut son pourcentage mondial déduit par rapport à son pourcentage de personnes connectées. Effectivement, on passe ainsi pour cet indicateur de 5,64% de personnes connectées à une quote-part de 7,37%. La simulation en jouant avec des valeurs différentes permet de sentir intuitivement ce qui se passe avec cette valeur.

TABLE 27 **Simulation pour interfaces**

| **Nombre d'interfaces** | 0 | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Pourcentage d'interfaces** | 0% | 12,5% | 25% | 37.5% | 50% | 62,5% | 75% | 87,5% | 95,8% | 100% |
| **Quote-part interface calculée par pondération** | 0% | 1,01% | 2,01% | 2,99% | 3,96% | 4,9% | 5,8% | 6,8% | 7,4% | 7,7% |
| **Quote-part générique** | 6,4% | 6,6% | 6,7% | 6,9% | 7,1% | 7,2% | 7,4% | 7,5% | 7,6% | 7,7% |

Ainsi, si aucun interface parmi les 24 était en français l'indicateur "interface" prendrait la valeur de 0 et cela affecterait la note générique qui baisserait de 1,2%. Par contre si les 24 interfaces existaient en français l'indicateur gagnerait 0,3% et la note globale 0,1%. On voit que la valeur de 5,6% de personnes connectées pour le français se trouve très proche de la valeur de 17 pour le nombre d'interface en français et donc c'est là que se situe le seuil de basculement (qui dépend de la distribution de ces valeurs) qui fera passer l'indicateur en dessus ou en dessous de la valeur de référence du pourcentage de francophones connectés.

Le même exercice avec l'indicateur de gouvernement électronique donnerait les résultats suivants.

TABLE 28 **Simulation pour index**

| **Note de la France dans l'index** | 0 | 25 | 50 | 75 | 96,1 | 100 |
|---|---|---|---|---|---|---|
| **Note du français** | 19 | 38 | 57 | 76 | 91 | 94 |
| **Quote-part du français pour l'index e.gov** | 0,18% | 3,5% | 5,1% | 6,7% | 8,0% | 8,3% |

L'effet sur la note global générique est assez marginal puisque il s'agit d'un parmi 5 micro-indicateurs dont la moyenne donne la note pour l'indicateur index: avec une note de 50, la quote-part puissance diminue à 7,5% et avec une note de 0 à 7,4%. Dans ce cas le seuil de basculement est sur la note de 57. Ces deux exemples montrent intuitivement comment le mécanisme de pondération fonctionne et quelle est sa sensibilité aux conditions initiales.

# Annexe I. Liste des micro-indicateurs

| MICRO-INDICATEURS USAGES |
|---|
| 3G pénétration |
| Abonnements mobiles |
| Appel vidéo via mobile |
| Appel vidéo via mobile - Selon Instant messaging |
| Appel vidéo via mobile - Selon Statista |
| Comptes haut-débit fixe |
| Comptes haut-débit mobile |
| Coût haut-débit |
| Foyers connectés à Internet |
| Guichets Bitcoin |
| Infections PC |
| Ligne fixe téléphone |
| Marché e.commerce |
| Messagerie instantanée via mobile |
| Monétique |
| Porte-monnaie électronique |
| Requêtes officielles vers Google (données) |
| Requêtes officielles vers Google (effacement) |
| Serveurs Internet sécurisés |
| Smartphones |
| Téléchargement OpenOffice |
| Téléphonie IP via mobile |
| Trafic mobile |
| Twitter via mobile |
| Usage serveurs par site Web |
| Utilisateurs Facebook |
| Utilisateurs Facebook selon Owloo |
| Utilisateurs Facebook selon Statista |
| Utilisateurs Linkedin |
| Utilisateurs Netflix- selon Netflix |
| Utilisateurs Netflix- selon Statista |
| Utilisateurs réseaux sociaux |
| Utilisateurs réseaux sociaux (projection 2021) |
| Utilisateurs Twitter |
| Utilisateurs Youtube |
| Utilisateurs Youtube selon Alexa |
| Utilisateurs Youtube selon Statista |
| Utilisation navigateur Google |

| MICRO-INDICATEURS INDEXES |
|---|
| Indicateur accès universel |
| Indicateur e.gov |
| Indicateur e.participation |
| Macro Indicateur WebIndex |
| Macro-indicateur infrastructure |

| MICRO-INDICATEURS CONTENUS |
|---|
| Amazon US - nombre de livres |
| W3Techs - contenus |
| Wikilivres – interface |
| Wikilivres - Nombre d'utilisateurs |
| Wikilivres Nombre d'articles |
| Wikilivres Nombre d'éditeurs |
| Wikipédia - Éditeurs occasionnels |
| Wikipédia - Articles par langues |
| Wikipédia - Éditeurs confirmés |
| Wikipédia – interface |
| Wikipédia - profondeur |
| Wikiquote - articles |
| Wikiquote - interface |
| Wikisource - interface |
| Wikisource articles |
| Wikiversité - interface |
| Wikiversité articles |
| Wiktionnaire - interface |
| Wiktionnaire - articles |

| MICRO-INDICATEURS INTERFACES |
|---|
| Android |
| Apple |
| Bing - Langues de traduction |
| Cortana– interface |
| Dictionnary.com - Langues de traduction |
| DMOZ - Langues des contenus |
| Duckduckgo - Interface |
| Duolingo - Langues de traduction |
| Facebook - Interface |
| Free-translator - Langues de traduction |
| Google - interface |
| Google Scholar - interface |
| Google Traduction - Langues de traduction |
| IM Translator - Langues de traduction |
| Ios |
| Linux |
| On-Line - Langues de traduction |
| Préférence Bing |
| Préférence Google |
| Préférence Yahoo |
| Reverso - Langues de traduction |
| SDL - langues de traduction |
| Skype - interface |
| Systran - Langues de traduction |
| Telegram– interface |
| Windows |

| MICRO-INDICATEURS TRAFIC | | | | |
|---|---|---|---|---|
| 1and1.com | Dmoz.org | Jurn.org | Reverbnation.com | Yelp.com |
| 4shared.com | Doaj.org | Justanswer.com | Rumble.com | Youku.com |
| 500px.com | Douban.com | Kaixin001.com | Rutube.ru | Youtube.com |
| A2hosting.com | Draugiem.lv | Kakao.com | Sapo.pt | Zhihu.com |
| Abilogic.com | Dreamhost.com | Kongregate.com | Scielo.org | Zoosk.com |
| About.me | Dreamwidth.org | Last.fm | Scienceopen.com | |
| Academia.edu | Dropbox.com | Librarything.com | Search.com | |
| Acfun.tv | Drupal.org | Linkedin.com | Secondlife.com | |
| Adam4Adam.com | Duckduckgo.com | Liquidweb.com | Semanticscholar.org | |
| Addictinggames.com | DXY.cn | Liveleak.com | Sharecare.com | |
| Adobe.com | Eclipse.org | Logoslibrary.eu | Similarweb.com | |
| Adultfriendfinder.com | Edx.org | Mamba.ru | Sitebuilder.com | |
| Alexa.com | Egnyte.com | Match.com | Skyrock.com | |
| Alivedirectory.com | Eharmony.com | Mediafire.com | Slideshare.net | |
| Anastasiadate.com | Etoro.com | Meetic.fr | Smugmug.com | |
| Angel.co | Europa.eu/ | Meetup.com | Snapchat.com | |
| Anobii.com | Exalead.com | Mega.nz | Socolar.com | |
| Answers.com | Excite.com | Mendeley.com | Sogou.com | |
| Apple.com | Experienceproject.com | Metacafe.com | Somuch.com | |
| Archive.org | Fetlife.com | Metafilter.com | Sony.com | |
| Archives-ouvertes.fr | Filefactory.com | Microsoft.com | Soso.com | |
| Armorgames.com | Fileserve.com | Miniclip.com | Soundcloud.com | |
| Arvixe.com | Filmaffinity.com | Mixi.jp | Spaces.ru | |
| Arxiv.org | Filmow.com | Mocospace.com | Spip.net | |
| Ashleymadison.com | Flickr.com | Moodle.org | Spotify.com | |
| Ask.com | Flipboard.com | Mouthshut.com | Squarespace.com | |
| Ask.fm | Flixster.com | Mozilla.org | Stackexchange.com | |
| Atom.io | Fotki.com | Mubi.com | Startpage.com | |
| Avvo.com | Fotolog.com | Myheritage.com | Steampowered.com | |
| Badoo.com | Foursquare.com | Mylife.com | Straightdope.com | |
| Baidu.com | Fun-mooc.fr | Myspace.com | Stumbleupon.com | |
| Bandcamp.com | Funnyordie.com | Napster.com | Sublimetext.com | |
| Base-search.net | Futurelearn.com | Naver.com | Tagged.com | |
| Beyond.com | G2a.com | Netbeans.org | Taringa.net | |
| Bilibili.com | Gaiaonline.com | Netcraft.com | Theses.fr | |
| Bing.com | Gameblog.com | Netflix.com | Tinyurl.com | |
| Bit.ly | Gamefaqs.com | Netlog.com | Trombi.com | |
| Bitbucket.org | Geni.com | Newgrounds.com | Tudou.com | |
| Bitshare.com | Gfycat.com | Nicovideo.jp | Twitch.tv | |
| Blackle.com | Gigablast.com | Ning.com | Twoo.com | |
| Bluehost.com | Gigasize.com | Notepad-plus-plus.org | Udacity.com | |
| Blurtit.com | Girlsaskguys.com | Novoed.com | Udemy.com | |
| Box.com | Github.com | Oatd.org | Uploaded.net | |
| Brackets.io | Godaddy.com | Odnoklassniki.ru | Uploading.com | |
| Business.com | GOG.com | Office.com | Veoh.com | |
| Busuu.com | Goodreads.com | Okcupid.com | Viadeo.com | |
| C9.io | Google.com | Openclassrooms.com | Vimeo.com | |
| Cafemom.com | Gotinder.com | Opengrey.eu | Vine.co | |
| Cairn.info | Grindr.com | Openlibrary.org | Visualstudio.com | |
| Care2.com | Hi5.com | Openoffice.org | Vk.com | |
| Caringbridge.org | Hightail.com | Openthesis.org | Wattpad.com | |
| Chacha.com | Hostgator.com | Opera.com | Wayn.com | |
| Chrome.com | Hulu.com | Origin.com | Wdl.org | |
| Classmates.com | Icloud.com | Paypal.com | Webcrawler.com | |
| Codeanywhere.com | Infinit.io | Periscope.com | Webometrics.info | |
| Codepen.io | Influenster.com | Periscope.tv | Weebly.com | |
| Commonsensemedia.org | Inmotionhosting.com | Photobucket.com | Weheartit.com | |
| Contentful.com | Instagram.com | Pinterest.com | Wetransfer.com | |
| Couchsurfing.com | Iqiyi.com | Playstation.com | Wikimedia.org | |
| Coursera.org | Italki.com | Plurk.com | Wistia.com | |
| Crunchyroll.com | Itch.io | Qq.com | Wix.com | |
| Cyworld.com | Ixquick.com | Quora.com | Wolframalpha.com | |
| Dailymotion.com | Ixquick.eu | Qwant.com | Wordpress.com | |
| Dart-europe.eu | Jasminedirectory.com | Raptr.com | Worldcat.org | |
| Daum.net | Jetbrains.com | Ravelry.com | Worldwidescience.org | |
| Deezer.com | Joinhouse.party | Reddit.com | Xbox.com | |
| Del.icio.us | Joomla.com | Rediff.com | Xing.com | |
| Depositfiles.com | Journalseek.net | Renren.com | Yacy.net | |
| Deviantart.com | Jstor.org | Researchgate.net | Yahoo.com | |

## Annexe II. Sources retenues

| NOM DE LA SOURCE | ADRESSE PRINCIPALE (ACCUEIL) | NOMBRE D'INDIC. FOURNIS |
|---|---|---|
| Alexa. Données sur le trafic des sites.[78] | http://www.alexa.com/ | 310 |
| Broad band Commission Report 2016. Données sur la connectivité et accès à l'information | http://broadbandcommission.org/ | 5 |
| CIA. Données démographiques et sur l'accès à l'information | https://www.cia.gov/ | 1 |
| Owloo. Pourcentage d'utilisation de Facebook | https://www.Owloo.com/ | 1 |
| Social Progress Imperative. | http://socialprogressimperative.org/ | 1 |
| Statista. Données sur la consultation de sites[79] | https://www.statista.com/statistics/ | 31 |
| Apache OpenOffice. Nombre de téléchargements d'OpenOffice | http://openoffice.org/stats/countries | 1 |
| The Internet World Stats. Données sur le nombre d'internautes par pays et par langue | http://www.internetworldstats.com/ | 2 |
| Translated. Index sur l'intérêt économique par langue | http://www.translated.net/ | 1 |
| W3Techs. Langues les mieux placées sur l'Internet | https://W3Techs.com/ | 1 |
| WebIndexData 2014. Indexes société de l'information | http://thewebindex.org/ | 9 |
| Wikipédia. Données démographiques et démo-linguistiques | https://wikipedia.org/ | 2 |
| Amazon. Données sur Amazon | https://www.amazon.com/ | 2 |
| Wikimédia. Données sur les différents Wikis de l'univers Wikimédia | https://stats.wikimedia.org/ | 11 |
| Projet Joshua. Données démo-linguistiques et démographiques | http://legacy.joshuaproject.net/ | 1 |
| Ethnologue. Données démo-linguistiques. | https://www.ethnologue.com/ | 1 |

[78] Donnée payante.
[79] Statista a été choisi comme la donnée payante principale après étude des sources possibles.

## PIÈCES JOINTES

- Fichier Excel de travail restructuré avec 140 langues plus reste des langues (fichier Excel L1L2-5M-RESTRUCT.xlxs)
- Fichier de contrôles (fichier Excel ControlesF)