maaya

réseau mondial pour la diversité linguistique

Étude sur la présence de la langue française dans le cyberespace

Rapport final – mars 2017

Par Daniel Prado.

Note: Le travail de collecte et analyse de données, un grand nombre d'aspects méthodologiques et un nombre très important de concepts contenus dans ce document ont été élaborés de manière conjointe avec Daniel Pimienta"

Contenu

Introduction

Ce rapport concerne le Protocole d'accord du 19/10/2015 (référence DNFDL F/ITF/OBS/AW/ph/20151015-004) signé entre l'Organisation internationale de la francophonie et MAAYA.

Pour rappel, ce projet réalisé entre octobre 2015 et mars 2017 propose de mesurer la présence de la langue française sur 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'Observatoire de la langue française de l'Organisation internationale de la francophonie, avec les résultats des différentes mesures, analyses et synthèses.

Rappelons également que le travail d'analyse se base sur la collecte d'informations concernant l'usage des outils informatiques et des langues dans de nombreuses applications, espaces et usages de l'Internet. La compilation et organisation finales de ces données permettent une évaluation par croisement avec différentes études ponctuelles et données démolinguistiques et leur mise en perspective.

Pour ce faire, les auteurs ont repris la base méthodologique de l'approche de 2013 appliquée à la langue française, la comparant aux autres grandes langues utilisées sur l'Internet, tout l'élargissant au plus grand nombre possible d'espaces et applications et en éliminant ceux qui seraient périmés. Les sources de l'étude[1] ont été renouvelées, tout en conservant les précédentes –rafraîchies- de manière à constituer des séries chronologiques qui permettront le suivi des évolutions.

Ce document résume le travail réalisé à la date du 31 mars 2017 et se veut le témoignage du travail réalisé à cette date et la photographie de l'état de la langue française dans l'Internet fin 2016.

[1] Beaucoup de sources étant éphémères ou non-actualisées, le travail de renouvellement requiert une attention particulière afin de trouver des résultats comparables.

Mesurer la place de la langue française

Lors de cette étude, il a été produit donc une première mesure[2] sur la place de la langue française – par rapport aux autres langues de la planète- qui permet de situer, outre sa place globale, celle qu'elle occupe dans un nombre importants de secteurs concernant le cyberespace.

La langue française figurerait entre les 4e et 5e places parmi les langues les plus utilisées de l'Internet selon les différentes estimations et angles de vue de cette étude.

Il a été utilisé le plus grand nombre possible de sources numériques pour quantifier la place des langues dans l'Internet soit directement lorsque les statistiques accessibles concernent la langue[3], soit indirectement, en convertissant des données par pays[4] en données par langue. Le nombre très important de sources étudiées oblige à une classification et à une certaine rigueur dans la terminologie et surtout pour en tirer, à la fin, des enseignements en fonction des types de pratiques et usages de l'internet, raison de proposer de consulter la section ***Terminologie utilisée***, dans ce chapitre.

Contexte et précédents

Cette étude est la troisième d'une série de mesures réalisées avec des méthodologies comparables, pour obtenir un classement du français dans l'Internet. Même si les auteurs collaboraient, depuis 1998 (voir encadré), afin de mesurer la présence des langues (notamment romanes) dans l'Internet, c'est depuis 2010 qu'ils ont mis en place cette nouvelle méthodologie, qui connaît année après année, des évolutions qualitatives. Elle a été implémentée en 2012, dans le cadre d'une étude soutenue par l'OIF, et en 2013, elle a nourri le chapitre Internet du rapport 2014 *« Le Français dans le Monde »* (francophonie, 2013).

Pendant la période 1998-2007, les auteurs de cette étude collaboraient, à partir de leurs institutions respectives, l'Union Latine et l'Association Réseaux & Développement (FUNREDES), pour des méthodes de mesure de la places des langues dans l'Internet susceptibles de fournir des indicateurs reproductibles et fiables en même temps que d'autres initiatives similaires existaient . A partir de 2007, l'évolution de la taille du Web et des politiques des moteurs de recherche a rendu obsolètes ces méthodes et a créé un vide dans la production d'indicateurs des langues dans l'Internet. 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 et susceptibles d'être financés par l'Union Européenne de manière à combler ce vide (projet DILINET). Cependant cet effort n'a pu aboutir malgré l'obstination et la qualité des équipes

[2] Première dans le cadre de cette étude. D'autres mesures ont été réalisées par le passé avec une méthodologie similaire.

[3] Par exemple le « nombre d'articles de Wikipédia » ou le nombre de livres chez Amazon, informations concernant directement la langue des articles ou des livres et non pas des chiffres concernant les utilisateurs, les connexions, etc.

[4] Par exemple, la répartition en pourcentage du trafic généré par pays du moteur de recherche Google ou le nombre d'utilisateurs de Facebook.

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 les auteurs en 2012.

Si le cadre méthodologique est commun aux trois mesures historiques, il y a un dégrée de complexité dû à l'augmentation des éléments mesurés et aux perfectionnements méthodologiques. Cette troisième étude atteint un niveau de maturité justifiant la mise en place d'un système informatique de gestion qui devrait permettre d'en faciliter la pérennité.

Important : 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, notamment par l'instabilité et/ou absence d'informations tant démolinguistiques (nombre de locuteurs, tant de langue maternelle que de langue seconde), que sociolinguistiques (choix de la langue utilisée au moment d'utiliser l'Internet) et des statistiques d'usage de l'Internet.

Cette transformation de données par pays en données par langue, avec toutes les limitations que cela comporte, fait de cette méthode une approche 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 l'usage de la langue sont devenus très peu fiables et le plus souvent inexistants.

La discussion sur les limites et les contrôles qui ont été réalisés pour garantir la fiabilité d'une méthode qui fait appel à une grande complexité tant dans les calculs réalisés que dans la compréhension des concepts qui en résulte, occupera en conséquence une part importante de ce rapport et de ses annexes.

Terminologie utilisée

Il est important de préciser un certain nombre de termes utilisés dans ce document car parfois certains peuvent être ambigus ou être utilisés différemment dans la littérature usuelle (polysémie), ou parfois c'est le processus inverse : un même concept peut porter différentes dénominations selon différentes sources (synonymie). Les auteurs ont également dû prendre parfois certaines *libertés* au moment de nommer des axes d'étude quand leur dénomination ne s'ajustait pas à leur approche méthodologique.

Les concepts de « L1 » (langue maternelle), « L2 » (langue seconde) et de « macrolangue » ne présentent pas de problèmes terminologiques majeurs, mais le comptage des locuteurs L1 et L2 ainsi que la présence de la distinction des « macrolangues » dans cette étude nécessitent quelques précisions.

Langue maternelle ou natale [L1].

Nous parlerons des « locuteurs L1 » ou des expressions du type « français L1 » (ou « anglais L1 », « espagnol L1 », « russe L1 ») pour définir les locuteurs dont la langue maternelle est celle indiquée par L1. Bien évidemment, les auteurs sont conscients qu'un locuteur peut avoir plus d'une L1, mais pour des raisons de manque de disponibilité des informations, et de cohérence des statistiques démographiques avec celles démolinguistiques, nous n'avons tenu compte que d'une seule L1 par individu, en général celle qui est mentionné dans la source consulté, en général, celle socialement dominante.

Langue seconde [L2]

Nous prendrons la définition usuelle en Europe et dans le monde francophone de « langue seconde [L2] »[5], à savoir la langue la plus importante après la langue maternelle. Dans nos statistiques, nous avons tenu compte des chiffres concernant ces locuteurs L2, mais seulement dans les pays où ces langues sont officielles (de jure ou de facto), soit au niveau national, soit au niveau régional.

Par exemple, ont été comptabilisés comme locuteurs de langue seconde [L2] ceux de langue française au Gabon, de langue anglaise en Inde ou de langue italienne en Croatie (italien langue officielle en Istrie), mais nous n'avons pas tenu compte des « francophones L2 habitant en Allemagne » ou des « anglophones L2 habitant en Suède », par exemple, pour une question d'harmonisation des statistiques prises en compte. ***Voir page 14.***

Langues LA

Pour des questions pratiques nous évitant des périphrases dans ce document, nous appellerons langues LA, celles ayant une faible ou très faible quantité de locuteurs L2 (locuteurs non comptabilisés), ainsi que celles qui pourraient disposer d'un nombre relativement important de locuteurs L2, mais pouvant jouer un second rôle de fait de la présence d'une L2 de prestige majeur, notamment dans l'usage des nouvelles technologies[6].

Langues LS

Inversement, et également pour des questions pratiques, les langues LS seront, dans ce document, les langues pouvant comptabiliser un nombre très important de locuteurs L2 ou ayant un rôle majeur vis-à-vis d'autres langues L2 dans l'accès aux

[5] D'après Wikipédia (https://fr.wikipedia.org/wiki/Langue_seconde), une autre définition serait valable dans le monde anglo-saxon, à savoir « langue apprise en deuxième, chronologiquement. »

[6] Ainsi, par exemple, nous n'avons pas retenu des langues comme le marathi ou télougou en Inde, l'afrikaans en Afrique du Sud ou le quechua dans les pays andins, toutes des langues ayant des locuteurs L2, mais se trouvant en situation de faiblesse face à une autre langue L2 en matière d'accès aux technologies internet.

technologies de l'Internet. Les auteurs n'ont gardé que 17 langues pouvant répondre à ces conditions. ***Voir page 14.***

Nous utiliserons également dans ce document les expressions **« unités de mesure » ou « UM »**, **« segments »** et **« catégories »** avec un sens précis qui répond à une classification établie par nos soins. Chaque **« catégorie »** réunit plusieurs **« segments »** qui, à leur tour, réunissent chacun d'entre eux, des **« unités de mesure »**.

***La liste des UM (368 en tout), des segments (36 en tout) et de catégories (11) est présentée en* Annexe. Unités de mesures (« UM ») retenues. Description.**

Unités de mesure (UM).

Nous appellerons ainsi toutes les applications, espaces ou usages de l'Internet qui ont été mesurées, un par un, par une source unique. Ainsi il peut y avoir plus d'une unité de mesure pour la même application. **Ces unités son en nombre de 368.**

Par exemple, seront des unités de mesure (UM) aussi bien les statistiques que donne *Alexa* de l'utilisation de *Facebook*, comme celle qu'en donne *Statista*, ou bien le nombre d'internautes selon l'*UIT* ou le nombre d'éditeurs qu'a *Wikipédia*.

Nous appellerons *« UM linguistiques »* les unités qui prennent en compte exclusivement des paramètres de langue (nombre d'articles de Wikipédia, nombre d'éditeurs du Wiktionnaire, nombre de livres chez Amazon, pourcentages de consultation par langue d'après W3Techs, etc.). **En nombre de 36**, **seulement 12** ont été utilisées à des fins de statistiques, **les 24 autres** étant des valeurs logiques apportant l'information sur la présence ou absence de **langue d'interface ou de traduction** d'un nombre équivalent d'applications ou espaces.

Et nous appellerons « UM d'usage » ou « UM d'usage d'Internet » les unités qui, **en nombre de 332** prennent en compte des chiffres par pays (pourcentage ou nombre d'usagers, pourcentage ou nombre d'internautes, pourcentage ou nombre de visites, etc.).

Segments

Nous appellerons *« segments »* **(en nombre de 36)** un ensemble d'unités de mesures ayant des caractéristiques proches, complémentaires, voire concurrentes, et présentant un intérêt particulier d'être mesurées ensemble. Par exemple, le segment *« Encyclopédie »* (de la catégorie *« Univers Wikimédia »*) réunit, parmi d'autres, les UM *Éditeurs occasionnels, Articles par langues* ou *Éditeurs confirmés* de Wikipédia.
Ou, pour donner un autre exemple, le segment *« images dominant »* (de la catégorie *« Réseau sociaux »*) comprend une cinquantaine d'applications telles, par

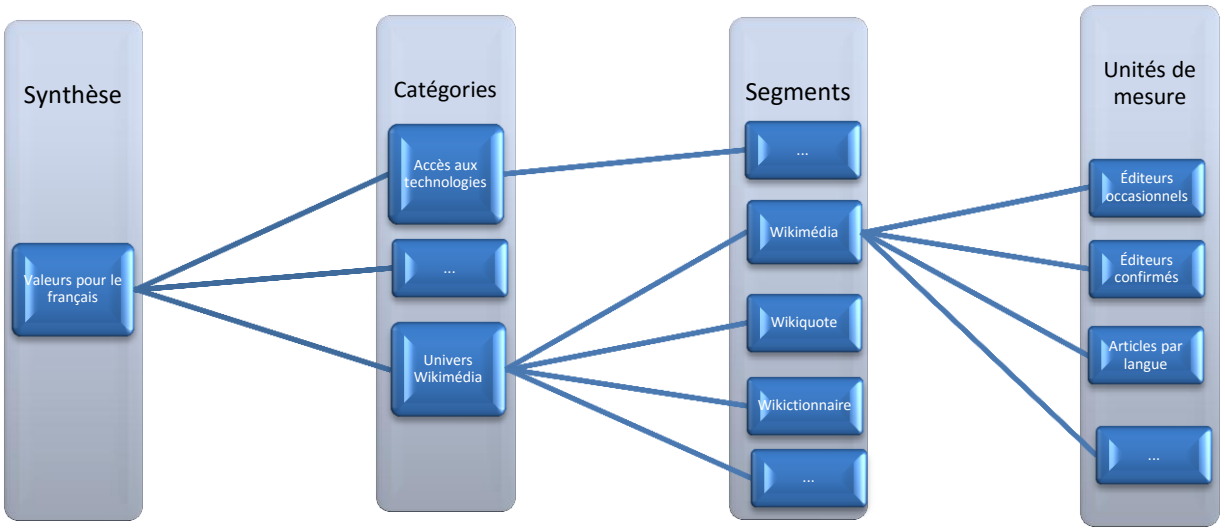
exemple, Netflix, Youtube ou Pinterest. Voir la liste des « segments » en ***Annexe - Unités de mesure, segments et catégories.***

Catégories

Nous appellerons *« catégories »* **les 11 ensembles** de *« segments »* ayant des traits distincts des autres UM et donc susceptibles de faire l'objet d'une mesure additionnée commune. L'ensemble complet des catégories établies est le suivant :

- Accès aux technologies
- Commerce électronique
- Communication électronique
- Contenus
- Environnement informatique
- Générale
- Jeux électroniques
- Outils informatiques
- Recherche scientifique
- Réseaux sociaux
- Téléchargement, partage et Autres outils techniques
- Univers Wikimédia

Voici une représentation graphique de la hierarchie décrite ci-dessus :



Travail réalisé

L'équipe de Maaya constitué par Daniel Pimienta et Daniel Prado, aidée par une assistante de saisie et une compagnie de développement informatique a réalisé, au cours de l'année 2016 et début 2017, les activités suivantes :

1. Reprise et amélioration de la méthode de travail concernant la mesure des langues sur l'Internet avec focalisation sur la langue française :
    - Extension et systématisation des sources et des applications. Plus de 500 sources ont été analysées, ainsi que prévu d'analyser plus de 500 unités de mesure (applications, espaces ou usages). Finalement ont été retenues une petite dizaine de sources (voir ***page 11***) et 368 unités de mesure. Voir ***page 7***.
    - Extension du matériel linguistique de base : la totalité des langues recensées ont été prises en compte dès le départ comme matériel d'étude, réduit après calculs. Voir ***page 16***.
    - Perfectionnement des méthodes de croisement de données démolinguistiques et d'usage avec techniques d'extrapolation pour parer au manque d'information par pays. Voir ***page 27***.
2. Recherche et sélection de sources ouvertes et payantes concernant le cyberespace. Voir ***page 20***.
3. Recherche et sélection de sources démolinguistiques idoines. Voir ***page 12***.
4. Création d'un système bureautique permettant la saisie, l'importation et la génération de données.
    - Optimisation des outils bureautiques de travail : en 2012, près d'une centaine de feuilles de calcul avaient été utilisées, une petite dizaine aujourd'hui pour l'analyse et cinq pour la synthèse. Voir ***page 34***.
    - Une bonne partie des informations a été directement importée, lorsque le format le permettait (informations démolinguistiques, démographiques et d'accès à la technologie, notamment)
    - Il a été fait appel à du personnel de saisie pour la majorité des cas où l'information n'était pas dans un format permettant une importation, notamment les sources « Alexa » et « Statista », principales sources d'information concernant les usages.
    - Création de routines de génération automatique de graphiques et tableaux.
5. Analyse finale et synthèse des résultats. Voir ***page 36.***
6. Réalisation (par une compagnie extérieure) d'un site Web permettant son insertion sur le site de l'OIF (Observatoire de la langue française)
    - Élaboration de la maquette du site dynamique avec graphiques, cadres et notes
    - Contrôle de l'opérabilité du site
    - Mise en ligne de graphiques, cadres et textes
    - Création de pages textuelles détaillés
7. Analyse évolutive concernant deux périodes (l'une concernant la recherche proposée à laquelle s'ajoute la recherche réalisée en 2013). Voir ci-dessous pour les ***actions manquantes***.
8. Réunion de cadrage avec l'OIF lors de la rencontre sur la préparation du rapport *« État de la Francophonie numérique ».*
9. Deux rapports intermédiaires et un rapport final détaillé

Produits devant être rendu avec ce rapport

Outre le présent rapport, il est prévu de rendre également :

- Les classeurs contenant les feuilles de calcul explicitées en Structure de la feuille de calcul utilisée
- La page Internet ad hoc décrite en Structure de la feuille de calcul utilisée
- Un manuel d'utilisation du système des classeurs et de son intégration à la page Internet.

Actions manquantes

Les retards pris en cours de route -qui ont motivé la demande d'un avenant au Protocole d'accord- et d'autres problèmes internes survenus en cours de route ont empêché ces actions que les auteurs se compromettent sur l'honneur de mener à bien après le rendu du rapport, et dans un délai non supérieur à 4 mois :

i. Rédaction d'un manuel d'utilisation du système informatique, contenant les informations nécessaires pour permettre à l'OIF de réaliser de manière autonome les campagnes de mesure postérieures à ce contrat.
ii. Une deuxième mesure permettant ainsi une courbe d'évolution de la langue française sur 3 périodes (et non pas 2 comme aujourd'hui) avec mise à jour des données sur le site.
iii. La version finale de la page Internet

Sources et indicateurs

Les estimations sur le nombre des locuteurs d'une langue et son usage présentent, tels des organismes vivants et complexes, un bon nombre de difficultés et ne peuvent pas être considérées comme des segments étanches au moment de faire de statistiques d'usage ou de contenus. Trop d'études de marché ou d'études statistiques prétendent donner des chiffres absolus sur la pondération par langue alors qu'ils ne prennent en compte que les langues officielles des États qu'ils étudient (comme semble le faire InternetWorldStats[7], à quelques nuances près) ou la langue déclarée dans les pages HTML, ou bien font des estimations de marché en fonction de paramètres de notoriété ou d'usage de sites (W3Techs[8]).

La réalité étant bien plus complexe et n'ayant aucune étude sérieuse et mondiale sur l'usage de la langue des individus dans leur relation avec le cyberespace, les auteurs préfèrent (et ceci depuis 2010) réaliser leurs calculs sur la base de :

- D'une part, tous les indicateurs de contenu linguistique accessibles
- D'autre part, le maximum des indicateurs d'usage d'internet accessibles –et crédibles- et les croiser avec les locuteurs d'une langue accédant au cyberespace et/ou en capacité d'utiliser l'ensemble d'applications et, de ce fait, **étant susceptibles de produire ou de créer du contenu linguistique**.

Cette capacité de création de contenus ne peut bien évidemment pas être mesurée de manière exacte, mais sur la base d'une fourchette dont les auteurs estiment que la partie basse serait donnée par des chiffres tenant en compte les locuteurs natifs des langues (L1) et la partie haute serait donnée, dans le cas spécifique du français, par des chiffres tentant en compte l'ensemble des locuteurs natifs et de ceux qui pourraient l'utiliser en tant que langue seconde (L1+L2)

C'est ainsi que le corpus d'étude se base, en plus des 36 UM fournissant des résultats par langue (p. ex. articles de Wiktionnaire, estimations W3Techs sur les langues des sites, langues d'interface et de traduction, etc.), d'un ensemble de 332 UM fournissant des informations d'usage d'internet par pays qui ont été croisées avec les informations démolinguistiques par pays, afin d'obtenir ainsi des estimations d'usage par langue.

Nous avons ainsi cherché des sources nous permettant d'avoir de chiffres de :

[7] http://www.internetworldstats.com/stats7.htm
[8] https://w3techs.com/technologies/overview/content_language/all

- Population par pays
- Locuteurs de langue maternelle (L1) pour toutes les langues
- Locuteurs de langue seconde (L2) pour l'ensemble des langues possédant un bassin de locuteurs significatif
- Indicateurs de contenus par langue
- Indicateurs d'usage d'Internet par pays

Afin de traiter ces informations il a été prévu de :

- recourir systématiquement à la source première. En d'autres termes, toujours préférer les sources réelles plutôt que les sources faisant référence aux sources premières, lorsque c'était possible[9]
- préférer toujours les sources les plus récentes (voir détails en ***note I***)
- ne pas corriger les approximations détectées dans les sources (sauf exceptions décrites en ***note II***)
- donner priorité aux sources présentant une certaine homogénéité dans les traitements
- préférer les sources pouvant facilement être automatisables pour un travail aisé de veille et d'actualisation des données.

Sur les données démolinguistiques

Pour pouvoir comparer le français à d'autres langues, il faut d'abord avoir des statistiques fiables de l'ensemble des langues de la planète ou, du moins, de celles qui sont parlées par une large majorité.

À l'heure actuelle, il n'y a pas une source **satisfaisante** qui soit **accessible et automatisable** concernant les chiffres démolinguistiques acceptables et validées, que ce soit pour la langue maternelle comme pour la langue seconde. Voici les décisions prises par les auteurs afin de se rapprocher le plus finement possible à des estimations crédibles tout en veillant au confort de l'utilisation des données. Voir **note III**

Chiffres sur les locuteurs L1 (langue maternelle)

Pour éviter les écarts de méthodologie entre études et de prendre une **typologie stable et facilement accessible**, les auteurs de cette étude ont dû prendre la décision de se référer à une source unique des données démo linguistiques pour la langue maternelle.

Les auteurs ont analysé la possibilité d'utilisation de sources pouvant fournir des statistiques démolinguistiques accessibles et actualisées pour l'ensemble de la

[9] Ce principe s'applique bien entendu aux sources qui offrent un panorama mondial mesuré par pays, mais pas aux sources consultées séparément pour chaque pays qui serait plutôt un travail de compétence de la source que nous avons consultée.

planète et ont particulièrement retenu, dans un premier temps, Ethnologue, Wikipédia, le CEFAN (ex-CIRAL)[10] et le Projet Josué[11].

Wikipédia aurait pu être la source à prendre en compte, du fait de la grande quantité d'informations disponibles actuellement. Mais la célèbre encyclopédie n'a pas une méthodologie unique pour le comptage des locuteurs. Alors qu'Ethnologue est souvent mentionné comme la source principale (outre CEFAN, également) pour la plupart des langues inventoriées, elle prend, pour d'autres, des études distinctes offrant des méthodologies différentes et donc pas comparables entre elles, ce qui aurait posé des problèmes statistiques sévères. Qui plus est, pour certaines langues, Wikipédia reste prudent et donne des fourchettes et non pas de chiffres précis. C'est notamment le cas de l'anglais ("360–400 millions en 2006 »[12]) se référant à Crystal[13] dans la version française. Un inconvénient supplémentaire est qu'il faut parcourir un nombre non négligeable de pages pour accéder à l'ensemble des données et qu'on trouve souvent des informations contradictoires entre versions linguistiques de Wikipédia.

Le ***CEFAN*** ne fournit pas des chiffres pour l'ensemble des langues de la planète et base beaucoup de ses chiffres à partir d'Ethnologue. Également, l'importation des données n'est pas aisée.

Ethnologue, aurait pu, comme par le passé, être la source choisie, malgré les réserves sur un certain nombre de statistiques et de typologie linguistique, mais pour une importation des données aisée, il nous aurait fallu opter par une option payante qui n'était pas à la portée financière de cette étude. Qui plus est, Ethnologue peut compter plus d'une langue L1 par individu (ce qui est logique, car nombre de populations peuvent avoir plus d'une langue native), ce qui provoque un problème statistique au moment du croisement des informations démolinguistiques avec les informations démographiques par pays. Également, l'addition des chiffres donnés par Ethnologue des locuteurs de toutes les langues par pays peut aussi être inférieure au chiffre démographique global qu'il indique.

Le ***Projet Josué,*** dont la vocation première n'est pas du tout un recensement des langues, mais vise des objectifs d'évangélisation (voir note [IV], met à disposition une base de données démolinguistiques en différentes versions téléchargeables. Parmi les sources du Projet Josué figure Ethnologue, mais aussi un nombre important d'autres références.[14] Il semble avoir une actualisation périodique (nous avons trouvé des différences importantes à différentes dates de 2016) et une analyse en

[10] http://www.axl.cefan.ulaval.ca/
[11] http://legacy.joshuaproject.net/international/fr/languages.php
[12] https://fr.wikipedia.org/wiki/Anglais
[13] David Crystal, chap. 9 « English worldwide », in David Denison, Richard M. A Hogg (dir.), *History of the English language*, Cambridge University Press, 2006, pp. 420–439.
[14] http://legacy.joshuaproject.net/data-sources.php

diagonal et aléatoire permet, en général, de trouver des chiffres proches à Ethnologue ou Wikipédia. Néanmoins, nous avons détecté un nombre important d'erreurs ou de contradictions que nous avons corrigées au fur et à mesure, sur la base d'Ethnologue et de Wikipédia. Certaines de ces modifications sont mentionnées en note [V]. Voir également l'***Annexe – Langues retenues***.

Ces erreurs du Projet Josué, dont certaines ont pu être détectées par nous-mêmes –mais il est impossible de tout vérifier dans le cadre de cette étude, permettent d'affirmer que le plus raisonnable en termes de futur est d'acquérir l'accès à la base d'Ethnologue qui, malgré nos réserves, semble être la plus stable en matière d'informations démolinguistiques, d'autant plus que c'est là sa vocation première. Mais, en attendant, la décision a été prise, et ce pour des raisons de praticité, d'utiliser la base démolinguistique du Projet Josué, tout en apportant les modifications nécessaires lorsque des erreurs sont constatées.

Chiffres sur les locuteurs L2 (langue seconde)

Si les chiffres sur le nombre de locuteurs L1 posent problème, ceux concernant la langue seconde en posent davantage. Les locuteurs « L2 », à savoir, ces individus qui n'ont pas la langue en question en tant que langue maternelle, mais acquise, sont supposés en avoir une maîtrise convenable.

Si pour la plupart des langues de la planète, ces locuteurs peuvent ne pas avoir une incidence directe sur ce type d'étude, il n'en va pas de même lorsqu'il s'agit d'étudier spécifiquement de langues « véhiculaires » comme le français, ainsi que l'anglais, l'espagnol[15], l'arabe, le russe, etc.

Les problèmes trouvés en matière de traitement de la langue seconde sont de différente sorte :

- Bien qu'il existe certaines données fiables sur les locuteurs L2 du français ou d'espagnol dans les pays où ces langues sont officielles (de facto ou de jure), il devient plus difficile de trouver des statistiques homogènes et comparables pour d'autres langues comme l'anglais, le chinois, l'arabe, le malais ou le hindi. Il est par exemple notoire la grande différence des chiffres existant sur le même article Wikipédia (Langue seconde ou L2) lorsqu'il est

[15] En effet, et pour ne prendre qu'un exemple non francophone, si la population du Paraguay parle majoritairement le guarani (langue amérindienne), la maîtrise de l'espagnol L2 par l'ensemble de la population semble bien reléguer le guarani dans l'Internet en faveur de la langue espagnole.

consulté en différentes versions linguistiques (français, anglais, espagnol, russe, etc.).

- Il est aussi difficile de trouver des statistiques à la fois précises et comparables, sur l'usage des grandes langues de communication[16] ailleurs que dans les pays où ces langues sont officielles. Or, il est courant de voir des sites (ou de pages) en version anglaise dans de nombreux pays non officiellement anglophones ainsi que voir beaucoup de sites (ou de pages) en langue française en Espagne, en Allemagne ou aux États-Unis[17].

- Mais il y a aussi beaucoup de cas où plus d'une langue a un rôle véhiculaire pour certaines populations (l'anglais et le swahili en Afrique de l'Est, l'anglais et le hindi en Inde, le français et le haoussa en Afrique francophone, etc.). Dans ce cas, quelle est la langue véhiculaire préférée pour l'utilisation par Internet? L'haoussa et le swahili paraissent se voir relégués derrière le français et l'anglais si l'on en croit les anciennes études du LOP[18] et de Marcel Diki-Kidiri (Diki-Kidiri, 2007), mais ce serait de moins en moins le cas pour le hindi, par exemple.

- Et dernier écueil, celui de l'interprétation de chaque source sur ce qu'est un locuteur L2 concernant le niveau de maîtrise, le niveau d'usage et l'étendue des populations étudiées. Si les chiffres de l'OIF sont clairs et documentés concernant la langue française, d'autres le sont beaucoup moins ou ne sont pas avares en mises en garde, comme le fait Ethnologue[19]

*Après une minutieuse étude des différentes sources et de ces questions, nous avons pris comme sources définitives pour les chiffres sur l'usage des langues secondes, l'**Observatoire de la langue française** pour la langue française et **Ethnologue** et **Wikipédia** pour les autres langues prises en compte.*

La raison du premier choix allait de soi, c'est sans doute la meilleure source et la plus documentée sur la langue de référence de cette étude.

[16] Tels l'anglais, le français, l'espagnol, le russe, le chinois, le portugais, l'arabe, etc.
[17] A tel point que nos précédentes études avant 2010 avaient montré que l'on produisait plus d'information en ligne en français dans des pays comme l'Espagne ou les États-Unis qu'en Afrique francophone.
[18] Le site du LOP est hors service aujourd'hui. Des archives peuvent être consultées à cette adresse : https://web-beta.archive.org/web/20161008061226/http://gii2.nagaokaut.ac.jp/gii/blog/lopdiary.php
[19] Voir *Bilingualism remarks* dans https://www.ethnologue.com/about/language-info

Nous aurions pu prendre pour certaines autres langues L2 les chiffres de leurs institutions de tutelle (Institut Cervantès pour l'espagnol, Institut Camoens et CPLP pour le portugais, etc.), mais non seulement leurs méthodes de comptage étaient incompatibles, aussi la plupart des langues étudiées n'ont simplement pas d'organisation de référence.

Le choix pour ces autres langues s'est donc porté sur Ethnologue quand le site avait des statistiques de L2 par pays[20] et Wikipédia dans certains cas où le premier ne satisfaisait pas à cette condition.

Par contre, afin de comparer des univers similaires et suivant la logique de l'OIF, nous n'avons pas pris les chiffres d'Ethnologue et de Wikipédia de locuteurs L2 concernant **seulement les pays où la langue en question n'est pas officielle ou usuelle**.

Corpus de langue retenu

Dans un souci d'objectivité et de cohérence, les résultats présentés dans ce document proviennent de calculs sur la base d'un corpus où **toutes** les langues recensées par le code ISO 639-3[21] ont été intégrées dès le départ. Ceci a permis de manière non arbitraire consigner toute référence trouvée à l'ensemble des langues dans les sources à contenu linguistique.

Un premier filtre avait enlevé les langues « appartenant » à une macrolangue comme expliqué ci-dessous, ne laissant qu'environ 7500 langues. Un deuxième filtre, une fois consignées toutes les données d'intérêt linguistique (croisement avec l'univers Wikimédia et Amazon, particulièrement), ont été enlevés les langues non vivantes, dont le traitement n'avait pas d'intérêt pour cette étude si ce n'est que la possibilité de constituer un fichier historique sur ces sources linguistiques pour une comparaison future.

A la fin des traitements, un ensemble des 6781 langues et macrolangues ont été retenues. Voir les classeurs Excel en annexe.

Appliquer nos algorithmes sur l'ensemble des langues sans distinction préalable a permis de tester les calculs nécessaires à la fusion des données démolinguistiques avec les données d'usage d'Internet sans a priori sur les langues devant être traitées.

[20] Sachant que des lacunes importantes existent, et que les chiffres globaux annoncés ne correspondent pas toujours à l'addition des chiffres par pays, le site a néanmoins une ligne méthodologique comparable entre les langues.

[21] Code de représentation des langues à trois chiffres. https://fr.wikipedia.org/wiki/ISO_639-3

En effet, la tentation était grande de ne considérer qu'un corpus réduit de langues pour la comparaison, mais ils auraient tous été sujets à caution (langues officielles ?, langues connues ?, langues démographiquement importantes ?, langues de Wikipédia ?, etc.), alors que le critère d'appliquer à toutes les langues nos calculs de croisement, sans distinction et en aveugle, a permis de mettre en évidence des réalités linguistiques qui nous auraient échappée au départ. Voir **note** [VI]

Le tableau final résultant de cette fusion a donné *in fine* une fourchette large sur les langues qui auraient une capacité à avoir une présence sur l'Internet.:

- Selon nos calculs, **292 langues ont une présence avérée** et/ou figurent parmi les **10 premières places** dans au moins l'un des usages de l'Internet, alors que seulement **278 langues naturelles vivantes** auraient une présence avérée dans l'Internet selon les sources qui ont été consultées[22]

- Selon ce qui se dégage sur une mesure fixée aux langues occupant au moins une fois l'une des **100 premières places** dans l'une des 344 UM, **620 langues** y figureraient, et **992** occuperaient au moins une fois l'une des **300 premières places**.

Sans pouvoir nous prononcer autrement que par des intuitions, ce premier chiffre **(620)** pourrait préfigurer un ordre de grandeur du nombre approximatif des langues ayant des perspectives de présence dans l'Internet à l'heure actuelle (au moins utilisation d'applications et quelques contenus), le deuxième pouvant préfigurer une présence à plus long terme.

Macrolangues

L'un des problèmes majeurs de croisement d'informations a été celui de la typologie linguistique utilisée par les différentes sources consultées. En effet, il est difficile de croiser ou d'additionner des informations sur l'usage qui indiqueraient, par exemple :

- le serbo-croate (y incluant aussi l'usage de ce qui est appelé aujourd'hui le « bosnien »[23]) et d'autres qui séparent l'un de l'autre
- ou bien certaines sources mentionnent le chinois (sans autre spécification) et d'autres séparent le mandarin du cantonais (et parfois aussi du yue).
- l'arabe aussi est parfois considéré dans ses formes dialectales, parfois seulement l'*arabe littéral*
- le malais et l'indonésien[24] figurent parfois séparés, parfois additionnés.

[22] Soit par le fait d'avoir une langue d'interface dans au moins l'une des applications étudiées et/ou par leur présence, au moins partielle, dans l'univers Wikimédia.
[23] Plus pour des raisons politiques, semble-t-il, que linguistiques
[24] Indonésien étant le nom donné en Indonésie à la variante malaise parlée dans le pays, où elle serait majoritairement langue seconde. https://fr.wikipedia.org/wiki/Indon%C3%A9sien

Nous avons donc dans ce cas été contraints de prendre la *macrolangue* comme unique référence, car autrement nous n'aurions pas pu les comparer aux différentes mesures. Ainsi, lorsqu'on parle de chinois, ce sera l'ensemble des langues chinoises (mandarin, hakka, yue, etc.), de même que pour l'arabe, l'allemand, le malais, le peul et d'autres groupes où seules les macrolangues (et non pas leurs composantes) sont prises en compte. Ces considérations ont une faible incidence sur les résultats de cette étude qui concernent notamment le français, mais nous les mentionnons car un regard critique pourrait pointer ce choix comme contraire à une logique saine de traitement des variantes linguistiques.

Bien évidemment, les chiffres que nous avons pu trouver pour des langues appartenant à une macrolangue ont été additionnés aux chiffres de celle-ci avant de l'éliminer de nos calculs.

★★★★★★★★★★

D'après Wikipédia , *« Une macrolangue désigne une catégorie de langue dans la norme internationale ISO 639-3 »* et d'après le Wiktionnaire , une *« macrolangue est un Groupe linguistique, d'importance variable, composé de plusieurs langues, dialectes ou parlers divers »*. Et rajoute : *« D'après Ethnologue, il s'agit d'une macrolangue, car elle comporte plusieurs groupes de dialectes pouvant chacun être considéré comme une langue distincte »*
https://fr.wikipedia.org/wiki/Macro-langue
https://fr.wiktionary.org/wiki/macrolangue

Sur les données démographiques

De nombreuses sources proposent des chiffres actualisés sur la population mondiale et des habitants par pays et si bien, pas toujours concordantes, elles sont assez proches les unes des autres de manière globale, ne posant des problèmes majeurs que quand il s'agit de petits pays de faible démographie. La *Banque mondiale*[25], le *CIA World Factbook*[26] et, dans le cas des pays européens, *Eurostat*[27] seraient les sources les plus fiables.

Bien entendu, *Wikipédia* offre aussi un ensemble riche d'informations démographiques (pays par pays, notamment) et sa page sur l'ensemble de la planète[28] semble choisir les meilleurs statistiques par pays, que ce soit des trois sources précédemment mentionnées, comme celles émanant officiellement des pays concernés. Mais une simple comparaison entre les versions linguistiques montrent des différences de l'ordre de 0,92 %.

[25] http://donnees.banquemondiale.org/indicateur/SP.POP.TOTL
[26] https://www.cia.gov/library/publications/the-world-factbook/fields/2119.html
[27] http://ec.europa.eu/eurostat/fr/web/population-demography-migration-projections/population-data
[28] https://fr.wikipedia.org/wiki/Liste_des_pays_par_population

Or, un premier exercice d'utiliser des chiffres démographiques de Wikipédia et des données démolinguistiques en provenance du Projet Josué (ou de l'OIF, Ethnologue ou Wikipédia lorsque des corrections s'imposaient), posait des problèmes de calculs récurrents, notamment pour les pays le plus faiblement peuplés.

C'est ainsi qu'il a été décidé, afin d'empêcher toute collision statistique, d'appliquer les chiffres démographiques émanant des statistiques démolinguistiques (de langue L1, cela s'entend), d'autant plus que la différence en termes globaux du projet Josué avec la version française de Wikipédia était acceptable (0,93) par rapport aux différences existantes entre sources (allant de 0,88 à 2,92 entre celles considérées).

Une raison supplémentaire à la prise en compte de valeurs coïncidentes avec les chiffres démolinguistiques est que la plupart d'indicateurs d'usage de l'Internet ne concernaient pas de chiffres en nombre mais en pourcentages, à l'exception de 10 UM sur les 344 traitées.

Les pays retenus

Afin de réaliser les calculs en croisant ´les données d'usage par pays avec les données démolinguistiques, nous avions considéré travailler, sur un ensemble de pays et des territoires que nombre de statistiques tant démolinguistiques que d'usage considéraient souvent de manière séparée. Beaucoup de ces territoires appartenaient –ou avaient passé un accord de protection- à un État souverain[29] y figurant aussi.

Nous avons progressivement restreint ces pays et territoires à ceux indiqués par la norme ISO-3166[30], soit 253[31] pays et territoires.

Finalement, il a été décidé de réunir les chiffres par État, en fusionnant les différentes statistiques tant démolinguistiques que d'usage. Même si cela complique le nombre de manipulations des fichiers, ça a permis de rendre un ensemble cohérent, d'autant plus que certaines études donnaient des chiffres sur des territoires que d'autres incluaient directement dans ceux de l'État d'appartenance[32] .

Par contre, a été prise la décision d'exclure certains pays qui n'avaient pas officiellement de statistiques à jour du nombre d'internautes (d'après l'UIT) ou bien le nombre d'utilisateurs était non signifiant. C'est le cas de la Corée du Nord, du Kosovo, de Nauru, des Palaos, de Saint-Marin, du Sahara occidental, du Soudan du Sud et du Vatican, qui seront considérés à l'intérieur d'une même catégorie comme

[29] A titre d'exemple, la Nouvelle Calédonie ou La Réunion, ou d'autres territoires français présentaient des statistiques différenciées de la France.
[30] https://fr.wikipedia.org/wiki/Code_pays#Norme_ISO_3166. Aujourd'hui, 249 États et territoires sont compris par la norme
[31] Au départ de nos études, ont été considérés jusqu'à 273 pays et territoires dont des statistiques démolinguistiques et d'usage pouvaient être trouvées.
[32] Bien sûr, nous n'entrerons pas dans la complexité des relations d'appartenance ou d'autonomie qu'ont les territoires d'outre-mer avec ce que nous pourrions appeler leur *métropole* de référence.

s'il s'agissait d'une même unité. Tous les chiffres ou pourcentages démolinguistiques et d'usage les concernant seront regroupés.

À la fin de l'étude n'ont été retenus que 191 pays (dont le groupement de plusieurs pays n'ayant pas de statistiques sur les internautes dans une seule unité). Voir ***note*** [VII] pour les pays et territoires concernés.

Sur les donnés d'information d'usage de l'Internet

Les types de pratiques ou usages de l'Internet se rapportent à des *applications*[33] ou des *espaces* ou *usages,* au sens large du terme[34]. Ce sont ces espaces et applications qui nous permettront par des **mesures** du positionnement des langues d'évaluer leur place dans l'Internet. Pour obtenir ces mesures ont été identifiées des **sources** quantitatives de données considérées comme suffisamment fiables pour témoigner raisonnablement de la place relative des langues dans une application ou un espace donné.

Concernant les sources

En prime abord, nous avons exploré la possibilité de consulter majoritairement des sources ouvertes au public et de considérer un nombre limité de sources commerciales payantes concernant des espaces ou applications essentielles, cela de manière à améliorer la qualité des données collectées. Hélas, le panorama existant est bien différent de celui trouvé cinq années auparavant, et il y a de moins en moins d'informations statistiques mondiales non marchandes ou n'appartenant pas à des entreprises de mercatique ou de communication d'entreprise.

S'il est aisé de trouver les statistiques concernant l'accès à l'Internet et aux nouvelles technologies[35], il est bien plus difficile, par contre, de trouver des sources concernant l'usage des applications, 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.

Pour réunir cet ensemble de données, ont été répertoriés au départ environ 500 sources d'information différentes (statistiques, sondages, évaluations, index, bases de données, etc.). Les sources ont été évaluées à propos de leur pertinence pour l'étude, de la confiance que l'on pouvait leur accorder et également de leur adéquation avec la méthode.

[33] Par exemple le moteur de recherche Google ou le réseau social Facebook
[34] Par exemple les téléphones intelligents ou la téléphonie haut-débit fixe.
[35] 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

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 sont des compagnies de marketing de l'Internet qui demandent souvent de fortes sommes pour divulguer leurs 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[36], ou de *W3Techs*.

Alexa et Statista ont été finalement sélectionnées comme sources payantes en raison de leur meilleur rapport qualité/prix[37] et ont complémenté les statistiques souhaitées pour la plupart des applications ou espaces. Cela a permis d'éliminer, par souci d'homogénéité, le nombre final de sources retenues à une quinzaine. Il est important de mentionner ci-après les sources principales et renvoyer le lecteur à l'***Annexe - Sources retenues*** pour plus de détails sur les autres sources consultées.

- L'Union Internationale des Télécommunications (UIT) est probablement la source la plus fiable et diffusée concernant l'accès aux technologies de la population mondiale. Bien souvent d'autres sources l'utilisent sans la nommer et parfois en procédant à des aménagements non documentés. Voir ***Annexe - Sources retenues*** pour plus de précisions sur l'UIT.
- Le système de statistiques de l'univers *Wikimédia*[38] (*Wikipédia, Wiktionnaire, Wikilivres*, etc.) est sans doute le plus transparent et le plus à jour de tout ce qu'il a pu être trouvé. Ce système s'alimente automatiquement et est mis à disposition au public. Il comprend un nombre important d'informations chiffrées et notamment le nombre d'articles de chaque site (et ses versions linguistiques), le nombre d'éditeurs (confirmés et occasionnels), nombre d'éditions, etc.
- *Alexa*[39] dispose d'un choix plus large[40] de statistiques, mais leurs données sont souvent approximatives de trafic par pays avec un nombre souvent limité de

[36] Appartenant aujourd'hui au groupe *Amazon*
[37] 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.
[38] https://stats.wikimedia.org/#fragment-12
[39] http://www.alexa.com/
[40] Plus de 80% de nos segments sont mesurés a l'aide de ce service.

pays, ce qui a mis de l'emphase sur la nécessité d'établir des méthodes d'extrapolations à l'ensemble des pays.

- W3Techs[41] est la source la plus diffusée concernant les contenus linguistiques par langue et probablement le site le plus commenté par les auteurs de ce rapport au long de leurs recherches *par la désinformation que ce site apporte au panorama mondial*, à cause des biais expliqués en ***Annexe - Unités de mesures (« UM ») retenues. Description***.
- Statista[42] est un généraliste des statistiques à l'écoute du client, facilitant la gestion des demandes, mais ayant finalement apporté peu de ressources utilisables.

La liste des autres sources retenues est détaillée en *Annexe - Sources retenues*

Concernant les unités de mesure d'usage (UM)

Comme expliqué auparavant, certains espaces, usages ou applications peuvent faire l'objet de différentes unités de mesure (UM) provenant de différentes sources. Il est donc possible de trouver, par exemple, des mesures d'usage de l'application Facebook trois fois (sources *Owloo*[43], *Alexa* et *Statista*) ou des chiffres sur les internautes fournis quatre fois (données de l'*UIT*, sources *WebIndex*, *InternetWorldStats* ou *Broadband Commission*[44]), même si c'est très peu fréquent. ***En cours d'étude la décision a été prise de privilégier la source première de laquelle s'inspire les 3 autres, celle de l'UIT.***

Il est aussi important de signaler que le type de mesure peut être différent selon l'espace, l'application ou, parfois, selon la source, et qu'il faut alors veiller à traiter correctement chaque type de données. Ainsi il est aisé d'obtenir des mesures en **quantités**[45], en **pourcentages nationaux** [46] ou en **pourcentages mondiaux**[47], ces deux derniers pouvant être appliqués au nombre d'habitants comme au nombre d'internautes seulement.

Concernant la datation des UM, il faut tenir compte qu'elles ont été consultées au deuxième semestre 2016 et que, dans la mesure du possible, ce sont des mesures établies pour l'année 2016, très souvent pour une date plus ancienne et parfois, pour des projections pour le futur.

[41] https://w3techs.com/
[42] https://fr.statista.com/
[43] Cette source qui offrait des données sur les utilisateurs de Facebook pour l'ensemble des pays a disparu pendant l'étude et n'est pas présenté dans les résultats finaux; cependant elle est utilisée dans le processus de révision et contrôle de ces résultats.
[44] http://www.broadbandcommission.org/Pages/default.aspx
[45] P. ex. le nombre d'internautes par pays
[46] P. ex., le pourcentage de consultation à Facebook par pays, ou le pourcentage du trafic Internet réalisé par mobile dans chaque pays.
[47] P. ex. la répartition du trafic mondial vers Facebook.com par pays.

Finalement, il est important de signaler que les sources de données par pays disponibles ne couvrent que rarement l'ensemble des pays de la planète, et très souvent, 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 les données pour les pays non renseignées par la source (Voir Traitement des données).

Les seules sources disponibles pour tous les pays retenus 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 ».

Il est également nécessaire de prendre en compte que le degré de mondialisation des applications est variable et que de plus en plus des pays ou des régions adoptent des applications qui leur sont propres, au détriment des grandes applications mondiales (come Google, Facebook, Yahoo !, Youtube, etc.).

Cette situation explique une présence excessive de la langue étudiée : par exemple, la 1ère place du français pour le moteur Exalead (appartenant au groupe français Dassault Systèmes) ou celle du coréen pour Kakao.com, application mobile coréenne permettant appeler et envoyer des messages gratuitement.

La conversion de ces différents types de données numériques, soient-elles en nombre ou en pourcentages demande une attention très particulière du type des données et de la portée de ces données. Ainsi, certaines sources offrant des résultats pour un nombre trop limité de pays et difficilement *extrapolables* (voir ***page 32***) n'ont pas été retenues. D'autres pouvant exprimer des index par pays (inutilisables dans une transformation par langue) ont été écartées.

Un ensemble d'environ 450 UM sur les usages avait été retenu au départ des calculs, mais n'ont été finalement gardés que 332 après épuration.

La liste complète des unités de mesure est présentée en ***Annexe. Unités de mesures (« UM ») retenues. Description***.

Type de données

Des sources qui ont été consultées, a été extraite une multitude de chiffres de différent type, divisés en « nombres entiers », en « pourcentages », en « données logiques » (1 / 0) et en « index » (ou notation). Quelques explications sur la nature de ces différences :

Nombres entiers (« NBE » dans la feuille de calcul): Les chiffres sont exprimés en nombres entiers. Cela peut être le nombre de « lignes téléphoniques fixes » par pays, le « nombre d'entrées dans le Wiktionnaire » ou le « nombre d'éditeurs de Wikipédia ».

Pourcentage nationaux (« PN » ou « PNE » dans la feuille de calcul). Les chiffres sont exprimés en décimales et, sauf exception[48], cela va de 0 à 1 ou de 0 à 100. P. ex. un ***« PN »*** de 44,79 % pour la Suisse dans l'***UM « Comptes haut-débit fixe »*** signifie que sur 100 habitants de la Suisse, presque 45 ont une connexion au haut-débit fixe.

Nous pouvons avoir des pourcentages sur le nombre d'habitants ou sur le nombre d'internautes simplement. Lorsque l'extrapolation s'avère nécessaire elle sera faite avec la méthode des quartiles. Nous verrons en ***page 27*** que certains cas de pourcentages nationaux s'avèrent réfractaires à être transformés en pourcentages par langue.

Pourcentage mondial (« PM » dans la feuille de calcul). Ces pourcentages indiquent la part de chaque pays dans la mesure mondial de l'UM étudiée. Ainsi, le fait que la France obtienne un chiffre de 0,05 dans l'UM « Usage serveurs par site Web (%) » signifie qu'en France se trouvent 5 % des serveurs mondiaux des sites recensés par W3Techs.

Pourcentage mondial relatif aux internautes (« PMI » dans la feuille de calcul). Ces pourcentages indiquent la part de chaque pays dans la mesure mondial de l'UM étudiée. Il est important de prendre en compte le nombre d'internautes par pays et non pas le nombre total d'habitants à cause de la source des données qui produit ces pourcentages (Alexa) qui base ses statistiques sur des usages d'Internet et non pas sur l'ensemble de la population. Ainsi, si le Canada est crédité du chiffre « 0,02 » dans l'UM « Outlook », cela signifiera que l'ensemble des internautes du Canada représentent 2 % des usagers du service de messagerie de Microsoft.

Langues d'interface (« LQ » dans la feuille de calcul). Ce type de données est seulement utilisé dans la feuille de calcul des *UM linguistiques* et a une valeur logique (1/0). Le « 1 » en Telegram signifiera qu'il y a une version française de l'application, par exemple. Ce type de données n'intervient pas du tout dans les calculs signifiant la présence du français et ne sont

[48] Certains pourcentages peuvent dépasser le plafond (1 ou 100), comme, par exemple, et c'est l'exemple le plus notoire, celui de la téléphonie mobile. En effet, dans de nombreux pays, surtout ceux étant très équipés en téléphonie mobile, on peut avoir plus de 100 téléphones pour 100 habitants.

pas comptabilisés que pour la mesure indiquée au point *Erreur ! Source du renvoi introuvable.* la ***page 43***.

Index ou notation. Il s'agit d'une notation attribuée à chaque pays pour un segment donné, issue des analyses ou du croisement de différents indicateurs. Les valeurs oscillent entre 0 à 100 ou de 0 à 1 et sont élaborées à partir d'une analyse de différents facteurs.

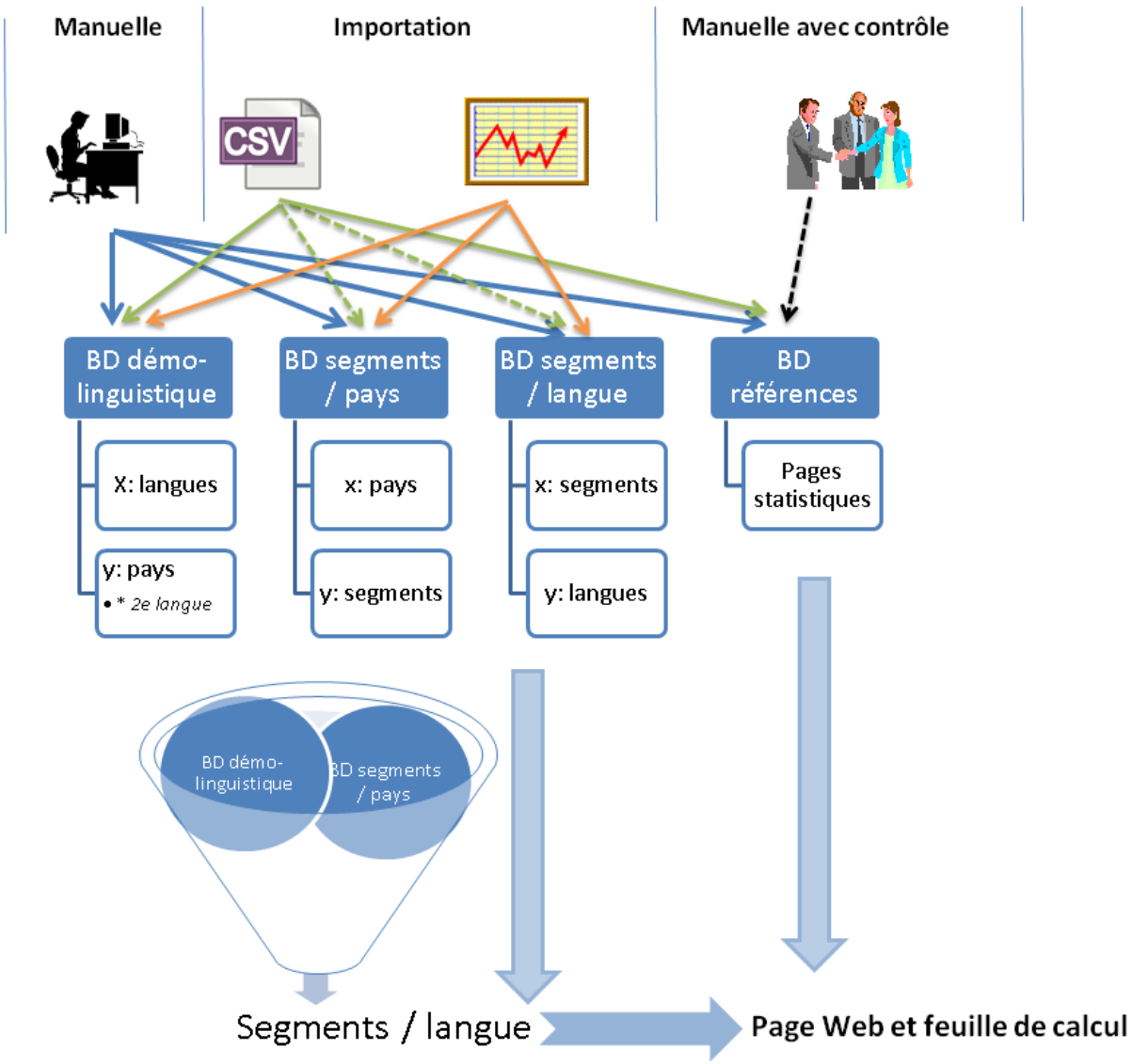
Le type de données « Index » ne figure pas dans ce document et a été exclu des calculs proposés dans ce rapport, en tout 5 UM d'usage. Voir ***note*** [VIII]

Traitement des données

Comme il a été dit, deux traitements différents ont été opérés sur les sources pour obtenir des résultats par UM, par segments, par catégories et d'ensemble :

- L'un concerne les UM linguistiques, minoritaires, issues des mesures sur la langue, indépendamment des pays et des individus.
- L'autre concerne les UM d'usage d'Internet, majoritaires, qui se basent sur les usages par pays.

L'ensemble des traitements comprend autant une saisie manuelle des informations comme une importation des données déjà traitées par d'autres sources, ainsi que, via la page Internet, une intégration d'informations spontanées. Voici le schéma qui montre l'ensemble du processus.



Traitement des UM linguistiques

Les données purement linguistiques[49] ne requièrent aucun type de traitement particulier. Elles sont traitées directement et tous les classements et pourcentages obtenus le sont par une simple formule appliquée à la feuille de calcul « Finale » (voir ***Structure de la feuille de calcul utilisée*** en ***page 34***) demandant de classer les 10 premières langues par UM étudiée et rendre des pourcentages calculés, langue par langue sur la base de l'addition de l'ensemble d'unités étudiées. Par exemple, si on trouvait 1 810 782 articles en français sur Wikipédia sur un ensemble de 41 462 882 pour l'ensemble des langues, le français, occupant la 6e place[50] aurait un contenu estimé a 4,36% du total.

Un complément de données linguistiques est donné par les indicateurs sur la présence ou non d'une interface dans les langues étudiées, donnant lieu à un traitement des plus simples, soit faire un tri descendant sur le nombre d'applications qui offrent une interface dans les langues étudiées. Voir Erreur ! Source du renvoi introuvable. ***en*** **page 43.**

Traitement des UM d'usage de l'Internet

Les données obtenues sur la mesure de l'utilisation, l'accès ou l'appropriation de l'Internet et de ses espaces et applications requièrent, par contre, un traitement d'une certaines complexité, car elles sont « croisées » avec les données démolinguistiques pays par pays, c'est-à-dire, sur le nombre estimé de locuteurs de l'ensemble des langues de la planète.

Pour cela, également et comme expliqué auparavant, et afin de créer la partie basse de la fourchette de mesure, sont prévues deux mesures, l'une qui tient en compte les locuteurs de langue maternelle (L1) et une seconde qui tient en compte le nombre de locuteurs de langue seconde (L2) également.

Les sous-chapitres suivants exposent la méthodologie utilisée dans la manipulation des données pour garantir des résultats fiables en croisant les informations d'usage (UM d'usage de l'Internet) avec les informations démolinguistiques. L'ensemble du développement s'applique autant à la partie basse de la fourchette base (locuteurs de langue maternelle) comme à la partie haute de la fourchette.

Néanmoins, nous nous devons de donner des informations préalables sur l'utilisation des données démolinguistiques, notamment concernant la langue seconde (L2).

Décisions démolinguistiques (notamment L2)

Comme nous l'avons dit, les langues étant des organismes vivants et imbriqués, nous ne pouvons pas nous contenter de prendre comme seul paramètre de calcul le

[49] Nombre de livres par langue chez Amazon, pourcentages de contenus linguistiques mesurés par W3Techs, nombre d'articles ou unités d'information, ainsi que des éditeurs et éditions dans l'univers Wikimédia.
[50] L'anglais, le suédois, le cébouano, l'allemand et le néerlandais avaient davantage d'articles que le français au mois d'octobre 2016.

nombre de locuteurs de langue maternelle et, du reste, pour beaucoup d'habitants de la planète, laquelle d'entre elles ?. Mais prendre comme seul paramètre la langue seconde serait aussi réducteur et, encore, laquelle d'entre elles ?.

La définition de la place de la langue française, selon la méthodologie ci-après expliquée, n'est donc pas donnée par une projection unique mais par une fourchette large à l'intérieur de laquelle nous prévoyons que se trouve la place de la langue française dans l'Internet. Cette fourchette, comme il a été dit, comprend une partie basse, sur des calculs qui tiennent en compte le nombre de locuteurs natifs d'une langue (L1), et la partie haute, sur des calculs qui tiennent en compte aussi bien le nombre de locuteurs natifs que ceux de langue seconde (L1+L2).

Si le choix méthodologique pour prendre le nombre de locuteurs de L1 est assez simple et a déjà été expliqué auparavant, la décision sur le nombre de locuteurs L2 n'est pas exempte de risques.

Les questions sur ce sujet sont nombreuses. À supposer que le nombre de locuteurs L2 soient bien définis autant mondialement que par pays (ce qui est loin d'être le cas comme indiqué précédemment), nous sommes confrontés à des défis d'importance, dans la manière de les comptabiliser :

- Il a été déjà énoncé que ne seront pas comptabilisés les locuteurs L2 se trouvant dans des pays où ces langues ne sont pas officielles ou d'usage. Les explications ci-dessous ne s'appliquent donc qu'aux pays concernés. Les données démolinguistiques des autres pays (tout en sachant qu'il y existe des locuteurs L2 [51] des langues en question), ne seront pas modifiées.
- Additionner simplement, dans le cadre d'un pays, tous les habitants L1 (que ce soit pour des langues que nous appelons LA[52] ou LS[53]) et les L2 ne serait pas logique. En effet, si à un moment « T », on comptabilise les L2, cela signifie que le locuteur comptabilisé abandonne à cet instant « T » sa L1 (dans le cas d'une LA). Comptabiliser les deux langues pour le même individu en même temps n'aurait pas de sens. Bien entendu, nous ne savons ni le moment qu'il utilise sa L2, ni pour quelle application, ni le contexte, et il peut aussi bien jongler de l'une à l'autre[54]. Mais nous savons qu'à un moment T il abandonne sa L1-LA au profit de sa L2, et c'est à ce moment qu'un équilibre se fait sur la base d'une utilisation alternée.
- Additionner ces L1 et les L2 ne serait pas pratique non plus : Si on additionnait simplement l'ensemble des locuteurs L1-LA avec les L2, certains pays fortement multilingues se verraient dotées d'une population relative bien supérieure à d'autres pays moins multilingues, voire monolingues,

[51] Souvent langue apprise, mais pas nécessairement pratiquée, appelée « LE ».
[52] Voir Langues LA dans le chapitre Terminologie utilisée
[53] Voir Langues LS dans le chapitre Terminologie utilisée
[54] Notamment pour les habitants qui vivent en situation de diglossie ou de bilinguisme. Voir https://fr.wikipedia.org/wiki/Diglossie

faussant les comparaisons avec les statistiques qui ne prennent en compte que les L1.

Décision 1 : le nombre d'habitants des pays ne peut pas changer, il doit rester stable pour comparer correctement avec les calculs basés exclusivement sur L1, car au moment de comptabiliser un locuteur en tant que locuteur de L2, cela signifie que c'est cette langue qu'il utilise et qu'il abandonne momentanément l'usage de sa L1.

Mais nous ne savons rien non plus sur le nombre de L2 qu'un individu utilise et laquelle il préférera. Nous pouvons nous baser sur des suppositions[55], et nous pouvons estimer que, dans un cadre comme celui de la France, la L2 majoritaire sera le français pour des locuteurs immigrés ou pour ceux dont la langue maternelle serait une langue régionale.

Mais quel choix sera celui des individus ayant une langue LA dans un pays comme le Canada (officiellement bilingue français-anglais), comme le Cameroun (idem), l'Inde (où anglais et hindi cohabitent souvent), c'est-à-dire où cohabitent des langues véhiculaires en compétition[56] ?

Quatre possibilités s'ouvrent à nous :

1. Nous pourrions comptabiliser tous les locuteurs recensés des L2 (plus le L1-LS) et l'addition se fait en détriment des L1 autres (LA). Mais rien qu'au Canada, pour ne prendre qu'un exemple mentionné, les L1+L2 d'anglais et de français additionnés dépassent la population totale du pays. Choix discutable et posant des problèmes de comparaison statistique.
2. Nous pouvons essayer de comptabiliser au prorata, en fonction des usages constatés, le nombre de locuteurs qui préféreront utiliser l'une des L2 plutôt que l'autre à un moment T. Si pour certains pays et certaines situations sociolinguistiques, ces chiffres sont disponibles, cet exercice pour l'ensemble des langues et des pays est hors de portée de cette étude.
3. Nous pourrions comptabiliser le chiffre haut de locuteurs de L2 dans des territoires où elle est seule L2, mais préférer systématiquement l'une des deux (ou trois) quand elles sont en compétition, celle qui aurait plus de « prestige » au niveau national. Choix discutable au Liban, en Algérie, au Cameroun, au Canada, en Inde, aux Philippines, etc. et parfois exercice hors de portée pour cette étude sur la décision de la langue ayant localement « plus de prestige ».
4. Prendre par défaut le chiffre haut de locuteurs L2 de français, de manière systématique dans toute circonstance, peu importe la langue LS en compétition.

[55] Et, pour certaines langues, sur des études bien précises et chiffrées, mais l'étude de ces cas pour l'ensemble des langues est hors de notre portée.

[56] Français et arabe pour les kabyles en Algérie, français et anglais pour les arabophones au Liban, anglais et swahili en Afrique de l'Est, etc.

Encore une fois, choix discutable au Liban, au Canada, au Cameroun, au Rwanda, en Haïti, etc.

Comme nous le voyons, ou les choix possibles sont discutables ou bien ils sont hors de portée pour cette étude.

La solution vient, par contre, du fait d'avoir établi à l'avance un système sur la base d'une fourchette, ou la partie basse est le chiffre nu des locuteurs L1. Nous pourrions tout simplement prendre, comme **partie haute de la fourchette, la possibilité idéale d'un usage large du français comme langue L2.** Nous savons que le prendre comme chiffre unique serait faux[57], mais dans le contexte de le déclarer comme chiffre le plus haut d'une fourchette, cela nous permet d'utiliser les chiffres L2 données par la francophonie dans toute sa grandeur et fixer une extrême limite supérieur compatible avec l'extrême limite inférieur que serait le chiffre nu de français L1. C'est un choix arbitraire et discutable, mais c'est vraisemblablement, à défaut de posséder les chiffres sociolinguistiques pour toutes les langues en situation de L2, le moins mauvais de ceux énoncés.

Décision 2 : La partie haute de la fourchette sera le potentiel maximum de la langue française en tant que L2, y compris dans des situations de compétition avec d'autres LS[58]. Donc, pour chaque pays étudié, le nombre total de francophones L1+L2 est gardé dans toute sa dimension[59].

Croisement de données démolinguistiques avec les UM d'usage d'internet

Afin de convertir des données par pays en données par langue, il a été procédé comme expliqué ci-dessous :

a) Constitution de feuilles de calculs :

*Le lecteur est conseillé de prêter une attention particulière aux feuilles **LOC, RP, EX** et **Final**, pouvant revenir sur les autres feuilles pour plus de détail par la suite.*

- Une feuille (appelé **LOC**) contenant les langues retenues sur l'axe « Y » et les pays retenus sur l'axe « X ». Au croisement se trouvent les informations du nombre de locuteurs L1 ou L2 sur la base déjà spécifiée de considérer une seule langue L1 ou L2 par individu. Cette feuille contient également :
    - en axe « Y », le code de la langue à trois chiffres correspondant à la norme ISO-639-3.

[57] Au Canada, par exemple, le français L2 réduirait à les locuteurs d'anglais à un chiffre inférieur au nombre d'anglais L1, par exemple, ce qui est vraisemblablement absurde.
[58] Notamment anglais et arabe, mais aussi certaines langues africaines ou l'allemand en Suisse et Luxembourg
[59] Pour d'autres situations de compétition de L2 (anglais-swahili, anglais-hindi, anglais-chinois à Hong-Kong, etc.), le nombre maximum des locuteurs de chaque langue est utilisé, ce qui, après contrôle, est sans risque sur les chiffres globaux par pays

- en axe « X », le code de pays à trois chiffres correspondant à la norme ISO-3166.
- en axe « X », les totaux démographiques par pays par une simple addition des colonnes respectives.
- en axe « X », les chiffres de l'UIT concernant les pourcentages d'internautes par pays et son calcul en nombre (les deux sur l'axe X).
- en axe « Y », les estimations calculées du nombre d'internautes par langue (voir feuille de calcul **« LI »**)
- d'autres informations de contrôle et de référence

- Une feuille de calcul (appelée **« RP »**) contenant les UM d'usage d'Internet sur l'axe « Y » et les pays retenus sur l'axe « X ». Cette feuille contient également :
  - en axe « Y », un code *ad hoc* permettant de donner une identification unique à l'UM, code mnémotechnique à 4 lettres et 2 chiffres
  - en axe « Y », le nom court donnée à l'UM
  - en axe « Y », le segment attribué à l'UM
  - en axe « Y », la catégorie attribuée à l'UM
  - en axe « Y », des informations textuelles concernant la source,
  - en axe « Y », son adresse (URL),
  - en axe « Y », des spécifications sur le contenu,
  - en axe « Y », le type d'information (nombre, pourcentage par pays, pourcentage mondial, etc.)
  - en axe « Y », la date,
  - en axe « Y », le nombre de pays recensés,
  - en axe « Y », des informations diverses de contrôle et de référence
- Une feuille de calcul (appelé **« PL »**) qui, identique à **« LOC »** (langues en axe « Y » et pays en « X ») calcule automatiquement le pourcentage de locuteurs de chaque langue par pays.
- Une feuille de calcul (appelé **LI**) qui, identique à LOC (langues en axe « Y » et pays en « X ») calcule automatiquement les chiffres sur le nombre d'internautes par pays
- Deux feuilles de calculs (appelées **« MP »** et **« MA »**) qui, identiques à **« RP »** (UM d'usage d'Internet en « Y » et pays en « X ») calculent automatiquement une valeur logique (0/1) permettant d'identifier l'existence ou inexistence d'informations de l'UM pour le pays concerné.
- Une feuille de calcul (appelée **« EX »**) identique à RP (UM d'usage d'Internet en « Y » et pays en « X ») permet d'obtenir des valeurs estimées pour les pays qui n'ont pas été renseignés à la base de la source consultée, par l'application d'extrapolations.

- Une feuille (appelée **Final**) contenant les langues retenues sur l'axe « Y » et l'ensemble des UM retenues[60] sur l'axe « X », **« Final »** permet d'obtenir les valeurs définitives par langue[61]. Cette feuille contient également :
  - en axe « X », le code d'identification de l'UM
  - en axe « X », le nom court donnée à l'UM
  - en axe « X », le segment attribué à l'UM
  - en axe « X », la catégorie attribuée à l'UM
  - en axe « X », d'autres informations de contrôle et de référence

b) Mise en garde 1. L'uniformité des statistiques par pays

Lorsqu'on a des statistiques par pays, l'hypothèse implicite est que les données concernant l'Internet se répartissent de la même manière dans toutes les langues parlées dans le pays. Prétendre que le pourcentage de personnes connectées à l'Internet dans un pays donné est indépendant de la langue maternelle de chaque individu est une hypothèse évidemment simplificatrice qui sous-entend que la fracture numérique nationale n'est pas corrélée à la langue.

Et c'est probablement faux: il est très probable que les locuteurs de certaines langues aient un taux de connexion à l'Internet inférieur ou supérieur à la moyenne nationale. La possibilité est grande qu'il s'agisse d'une population de niveau socio-économique, culturel ou d'éducation différents.

Cependant, cette simplification ne va pas déformer de manière violente les résultats, à condition, bien entendu, de ne pas mesurer des applications concernant exclusivement les locuteurs d'une langue spécifique (par exemple, des méthodes d'apprentissage de langue).

c) Mise en garde 2. La complétude des statistiques par pays

Sur les 332 UM d'usage de l'Internet, rares sont celles qui concernaient des statistiques sur l'ensemble des pays retenus[62]. Bien que la plupart de mesures concernent les pays les plus connectés et les plus peuplés et que ceci représente en général la majorité de la population devant être étudiée, il n'est pas moins frustrant de constater que nous avons dû parfois nous contenter de sources n'offrant des statistiques que d'une petite dizaine de pays.

Puisque nous ne pouvions pas restreindre l'étude aux très rares indicateurs complets, et pour parer à l'absence de données de certains pays, ce qui aurait été pénalisant, notamment pour les langues qui nous intéressaient étudier de plus

[60] Dans « Final » se trouvent, en axe X, autant les UM linguistiques que celles d'usage de l'Internet, mais seulement ces dernières nous intéressent dans cette section.
[61] Exprimées en nombre ou en pourcentages.
[62] Comme le « Pourcentage d'internautes par pays (UIT) », le « Nombre de lignes téléphoniques fixes par pays (UIT) », le · »nombre de téléchargements du logiciel OpenOffice (OpenOffice) », etc.

près[63], nous avons opté pour « extrapoler »[64] au mieux ce que devraient être les valeurs non renseignées à partir des valeurs existantes et à partir d'autres éléments pragmatiques dont nous pouvions tenir compte. Même si l'extrapolation n'est pas la panacée, elle permet d'obtenir des résultats, certes approximatifs, mais avec une marge d'erreur acceptable car elle se trouve à l'intérieur des marges d'erreurs des autres éléments, en particulier des sources. Voir ***note***[IX] sur la vérification des extrapolations.

Deux méthodes principales d'extrapolation avaient été adoptées, mais seul l'un d'eux est utilisé dans ce document, car les 5 UM concernées par le second (les index) ont été écartées de l'étude. Voir note[X] pour plus de précisions.

Également, dans ce document, n'ont pas été appliquées des méthodes d'extrapolation aux données suivantes :

- Aux pourcentages de requêtes d'effacement (droit à l'oubli) sollicitées aux sites Google, Facebook et Twitter. L'extrapolation n'est pas possible car la donnée est liée à la culture ou situation juridique du pays et ceci nous amène à renoncer à les traiter.
- Aux types de données exprimant des pourcentages par pays (**« PN »** dans la feuille de calcul), car la méthode se révélait être inappropriée offrant des résultats paradoxaux en matière de langues minoritaires parlées dans un seul pays. Voici les UM concernées :
    - En particulier, la mesure du pourcentage de personnes connectées par pays (UIT). Un simple calcul arithmétique permet d'obtenir le nombre d'internautes par pays et donc par langue, d'autant plus que cette mesure couvre l'ensemble des pays, à l'exception du groupement de pays réuni dans une seule catégorie, précisément par ce manque de statistiques[65].
    - Trois autres mesures concernant des PN qui n'ont pas été appliqués des extrapolations (*Comptes haut-débit fixe, Comptes haut-débit mobile et Abonnements mobiles*) mais le nombre de pays recensés étaient amplement majoritaires (entre 160 et 180).

d) Calcul final.

La feuille finale calcule le nombre de locuteurs de chaque langue par un croisement des informations venant de la feuille **« RP »** avec la feuille **« LOC »**, sur la base la formule SOMMEPROD qui *multiplie les valeurs correspondantes des matrices spécifiées et calcule la somme de ces produits*. S'il y a lieu d'opérer une

[63] Le français en premier lieu, mais aussi les autres langues qui usuellement occupent les premières places dans les secteurs étudiées, notamment l'anglais, le chinois, l'espagnol, l'allemand, le portugais, le russe, le japonais, l'arabe, l'italien, le suédois, le hindi, le polonais, le néerlandais, le turc, le perse, etc.
[64] Voir la définition d'extrapolation : http://www.larousse.fr/dictionnaires/francais/extrapolation/32473
[65] Rappelons-les : la Corée du Nord, le Kosovo, Nauru, les Palaos, Saint-Marin, le Sahara occidental, le Soudan du Sud et le Vatican.

extrapolation, ces paramètres sont consignés dans la formule, autrement, elle est appliquée dans l'état.

Structure de la feuille de calcul utilisée

Rappelons que les calculs ont été effectués sur la base d'une feuille de calcul. Un ensemble définitif de 4 classeurs[66] a été nécessaire pour réaliser l'ensemble des calculs, tant d'analyse comme de synthèse. A cela s'ajoute un classeur contenants les nomenclatures utilisées, par langue, par pays et par UM.

Plusieurs autres classeurs ou sous-classeurs de simulations diverses ont été créés en cours de route, bien entendu, aujourd'hui disparus.

Avec cette étude seront rendus les 5 classeurs, mais réduits en nombre des langues à celles qui sont présentes, au moins une fois, dans les calculs finaux et synthèses.

L'ensemble des calculs de base ont été réalisés sur deux classeurs identiques en structure (l'un pour des chiffres de L1 et l'autre pour des chiffres L1+L2) contenant :

- Un feuille de calcul initiale avec des informations démolinguistiques sur près de 7 500 (6 781 à la fin), langues et macrolangues (en «Y») et près de 230 (191 à la fin) pays et territoires (en «X»). Cette feuille de calcul apportait également les informations démographiques par pays.
- Une feuille de calcul contentant initialement près de 450 (332 à la fin) UM mesurant les usages d'Internent (en «Y») par pays (en «X»).
- Une autre feuille de calcul contenant des statistiques sur le contenu linguistique d'une quinzaine (finalement 12) UM et de 24 UM indiquant la langue d'interface de certaines applications (en «X»). Ce fichier est croisé (en «Y») par langue.
- Une feuille de calcul global donnant les chiffres par langue de l'ensemble des 344 UM d'usage et UM linguistiques (UM d'interface exclues).
- Ce classeur contient également 6 feuilles intermédiaires de calcul :
  - 1 feuille de calcul permettant l'obtention de pourcentages de locuteurs contenant le croisement des informations de locuteurs par pays
  - 1 feuille de calcul permettant l'obtention des chiffres sur le nombre d'internautes par pays
  - 3 feuilles permettant des calculs pour les extrapolations, croisant les informations des UM d'usage avec les pays
  - 1 feuille de calcul contenant la nomenclature contrôlée des UM

L'ensemble des calculs de synthèse ont été réalisés sur trois classeurs, dont deux presque identiques (l'un pour L1, l'autre pour L1+L2) contenant, ces deux derniers :

[66] En l'occurrence, il s'agit de classeurs Excel. Les auteurs auraient préféré d'utiliser OpenOffice, mais le manque de pratique et certaines incompatibilités constatées entre Excel (déjà utilisé pour des travaux antérieurs) et Calc (classeur d'OpenOffice) les en ont dissuadés.

- La feuille de calcul finale copiée, en chiffres, du premier classeur
- Une feuille calculant les 10 premières positions pour chaque UM, pour chaque segment et pour chaque catégorie étudiée. Un calcul final d'ensemble complète cette feuille de calcul.
- Une feuille de calcul permettant la création des graphiques et des cadres reproduits en partie dans ce document, soit pour L1, soit pour L1+L2.
- Le troisième classeur (petit) calcule, sur la base des deux derniers des synthèses comparant les chiffres L1 et L1+L2

L'ensemble des classeurs sont préparés de manière à éviter le maximum de manipulations. Une fois saisies les informations décrites ci-dessous de la première feuille de calcul, le résultat est automatiquement généré et une fois copié cette dernière feuille de calcul sur le 2e classeur, la génération des graphiques et cadres se fait automatiquement.

Les seules manipulations nécessaires à la création de nouveaux résultats concernent la saisie des informations suivantes :

- Nombre de locuteurs des langues traitées par pays
- Chiffres concernant les UM d'usage d'internet par pays
- Chiffres concernant les UM linguistiques par langue
- Données concernant l'extrapolation (feuille de calcul « EX »).

Bien entendu, des manipulations supplémentaires seraient nécessaires dans le cas de rajout ou suppression de pays ou territoires d'étude, de rajout ou suppression de langues d'étude, de rajout ou suppression d'UM étudiées (soient-elles linguistiques ou d'usage d'internet) ou de nécessité de production de cadres et graphiques supplémentaires.

Un manuel d'opérations décrivant les options nécessaires pour procéder à ces changements sera fourni une fois réalisé la 2e mesure de cette étude.

Équipement de base

Tous les calculs et génération de cadres et graphiques ont été réalisés sur un ordinateur de bureau ayant les caractéristiques suivantes :

- Intel Core i5-2500 à 3,3 Ghz
- Mémoire 8 Go
- SE : Windows 7 Ultimate (2013)
- Excel Microsoft Office 2010 - 32 bits.

Cette configuration -déjà ancienne- est suffisante, mais c'est une base minimum à cause de la densité des calculs ayant été effectués. Il est recommandable de ne pas utiliser une configuration inférieure, car un autre ordinateur de l'équipe a présenté des problèmes récurrents de mémoire, compliquant la tâche, à cause du nombre important d'informations gérées et des opérations de calcul.

Résultats des mesures

Selon les différentes analyses exposées dans ce document, le français serait de plus en plus présent dans le cyberespace, si l'on compare nos précédentes études (voir ***Évolution de la langue française page 45***), tant d'une manière générale comme par regroupements.

Les auteurs ne sauraient trop mettre en garde sur l'interprétation des résultats obtenus, et leur usage. Ils rappellent que ces résultats doivent être pris comme le calcul final de mesures réalisées se trouvant à l'intérieur d'une fourchette large où la partie basse sera donné par les chiffres données en tenant en compte exclusivement l'usage du français en tant que langue maternelle (L1) et la partie haute sera celui indiqué par les chiffres qui considèrent le français en tant que langue maternelle et langue seconde (L1+L2). Également, le lecteur est invité à prendre note des nombreuses mises en garde sur les limitations et biais des sources consultés et indicateurs produits, ainsi que sur certaines définitions propres à cette étude.

Cette étude produit une première série de mesures pour 2016 [67] sur la place de la langue française dans l'Internet et son usage, permettant de la classer dans chaque application ou espace étudié, ainsi que sous l'angle de différents regroupements, comme indiqué précédemment, par catégories, segments, UM et d'autres que nous verrons.

Les résultats sont le fruit d'une mesure ponctuelle réalisée sur des données antérieures au 15 décembre 2016. En particulier, les données démolinguistiques procèdent des sources consultées entre juin et octobre 2016, notamment Wikipédia, Ethnologue et Josué et les données sur le nombre ou pourcentage d'internautes (InternetWorldStats, UIT, WebIndex) s'arrêtent au mois de juin 2016.

À l'heure où ce rapport est écrit, de nouvelles statistiques sur le nombre d'internautes sont diffusées et apportent une évolution de taille et seront prises en compte pour la 2è mesure de cette étude.

Ce chapitre présente le résultat des analyses effectuées sur un corpus initial de plus de 7500 langues (don finalement près de 7 000 ont été gardées), plus de 500 objets de mesure -ou indicateurs différents sur les usages de l'Internet (dont finalement 344 ont été retenus) et près de 230 pays et territoires (dont 191 ont été finalement retenus[68]). Plusieurs centaines de sources ont été consultées, mais les auteurs n'ont retenu qu'une quinzaine pouvant satisfaire aux critères d'universalité et de sérieux au même temps.

[67] La 3e si on considère celles déjà réalisées en 2010 et 2013, avec des méthodologies comparables, la seconde ayant nourri le chapitre Internet du rapport 2014 Le Français dans le Monde publié chez Nathan.
[68] Comme expliqué auparavant, 190 pays ont été retenus plus un groupement spécifique réunissant l'ensemble des pays et territoires dont l'UIT n'aurait pas de statistiques concernant le nombre d'internautes.

Ainsi, voici une synthèse des mesures sur la place relative du français avec les autres langues de la planète, divisées par catégories, par segments et quelques exemples sur certains espaces pris au choix.

Synthèse des résultats

Selon les différentes calculs et additions, la langue française se placerait entre les 4e et 5e places parmi les langues les plus utilisées de l'Internet Voici un tableau de synthèse des places occupées, détaillées par la suite.

Langue française figure entre la 4e et 5e langue de l'Internet selon les différentes estimations.

Ces classements doivent être considérés tant en matière de potentiel d'utilisateurs francophones de l'Internet, pour la grande majorité des cas étudiés, comme de contenus en langue française, pour une minorité des paramètres pris en compte.

Synthèse	Seulement locuteurs langue maternelle	Locuteurs langue maternelle et langue seconde
Place du français tous espaces confondus (sur 344)	4	4
excek	4	4
Place du français sur la base de 36 sous-catégories (segments)	4	4
Place du français parmi les langues qui prennent les 10 premières places de tous les espaces.	5	4
Nombre de fois où le français prend l'une des 10 premières places		
Le français occupe la 1ère place :	17	17
2e place :	3	5
3e place :	34	58
4e place :	47	64
5e place :	49	65
6e place :	50	38
7e place :	29	27
8e place :	24	13
9e place :	15	17
10e place :	13	11
Le français n'occupe aucune place parmi les 10 premières :	64	30
Total	344	344

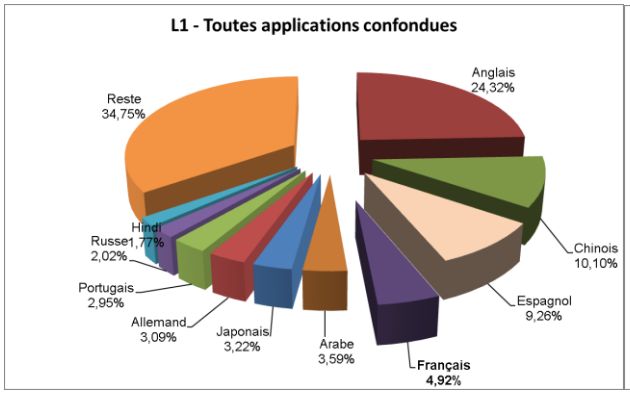
Place de la langue française par « unité de mesure »

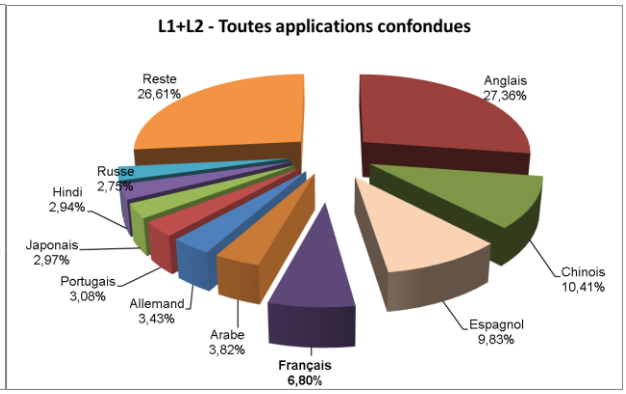
Sur 344 indicateurs différents, la langue française se placerait le plus souvent entre la **3e** et la **7e place** si on ne prenait en compte que les locuteurs de langue maternelle (L1), et entre la **3e** et la **5e place** si les locuteurs de langue seconde (L2) y étaient également considérés.

La somme de l'ensemble des mesures *convertie en pourcentages globaux* donneraient le français **4e langue** de l'Internet avec un **4,92 %** de l'ensemble collecté concernant seulement

les locuteurs de langue maternelle et **6,80 %** si on prend en compte également la langue seconde. Voir ***Annexe – Analyse par UM*** pour des informations détaillées.

Les graphiques suivants montrent une comparaison entre les 10 principales langues autant pour L1 que pour L1+L2. « Reste » correspond à l'ensemble des langues ne figurant pas détaillées.

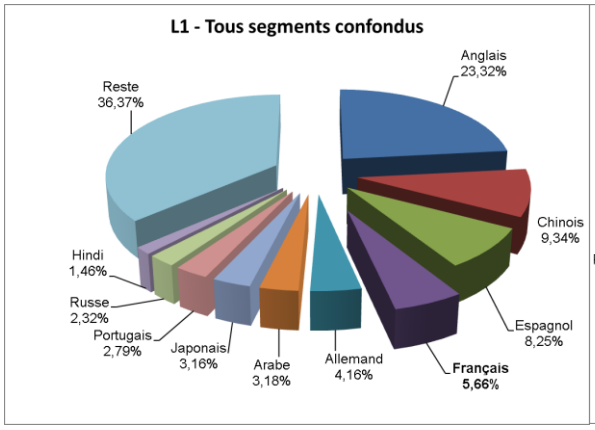


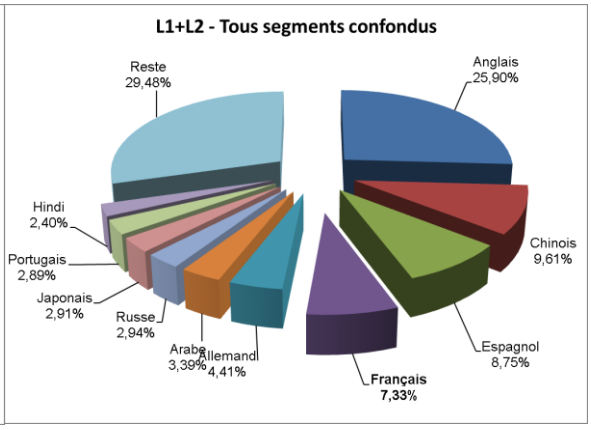


Place de la langue française par segments

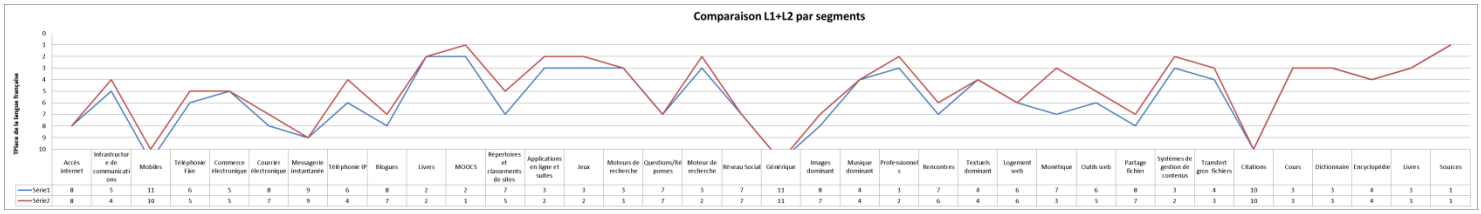
Un premier regroupement entre applications, espaces ou usages ayant certaines affinités ou se trouvant en proche compétition appelés **« segments »** (36 en tout, voir ***page 7***), le français se trouverait entre la **3**e et la **9**e place, mais plus fréquemment entre les places **3**e et **7**e. Par contre, elle ne figurerait pas dans les 10 premières places dans les segments correspondant aux mesures sur le nombre d'accès aux *mobiles*, ainsi que sur l'usage des *« messageries instantanées »* et des *« réseaux sociaux à où le textuel est prédominant »*.

La somme de l'ensemble des **segments** convertie en pourcentages globaux donneraient le français **4**e langue de l'Internet avec un **5,66 %** de l'ensemble collecté concernant seulement la langue maternelle et **7,33 %** si on prend en compte également la langue seconde. Voir graphiques suivants. Voir ***Annexe – Analyse par segments*** pour des informations détaillées.





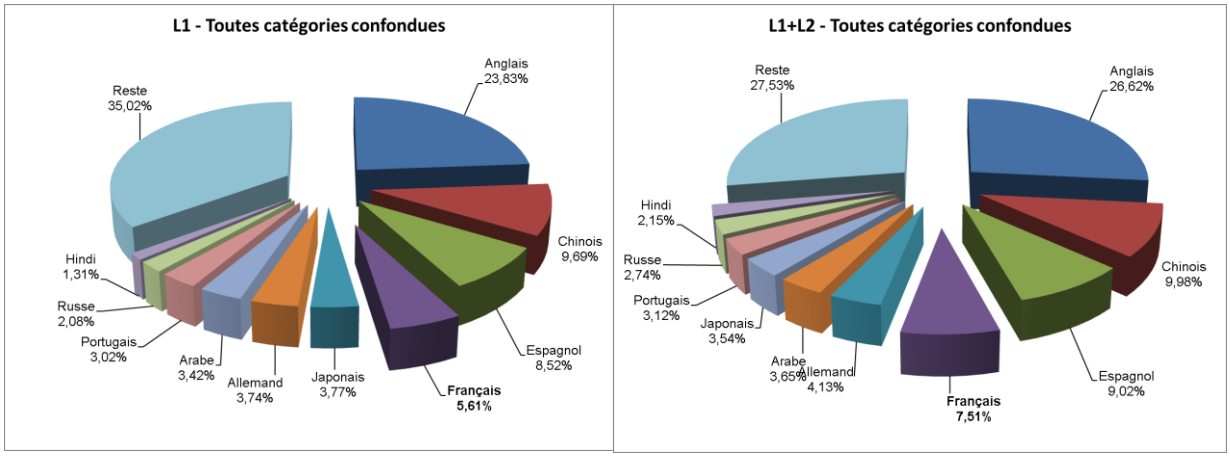
*Le graphique suivant (qui peut être mieux apprécié dans l'**Annexe – Analyse par segments**), montre la comparaison du comportement des principales langues par segments entre L1 et L1+L2.*



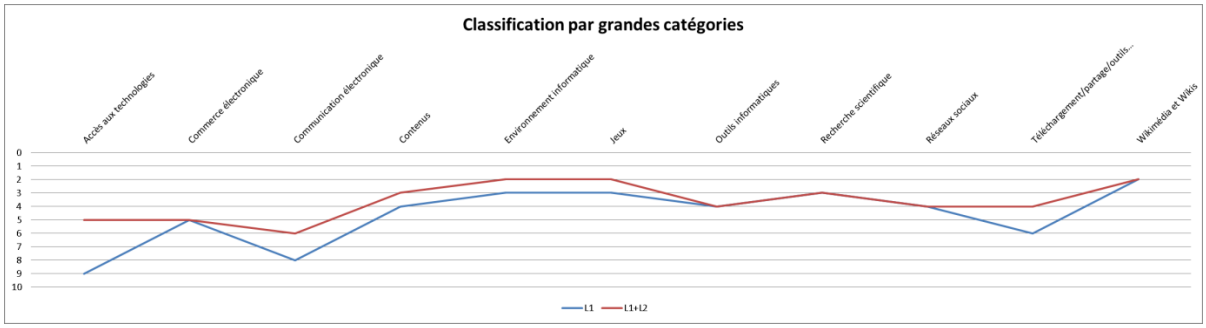
Place de la langue française par catégories

Un regroupement majeur de l'ensemble d'indicateurs, appelé **« Catégories »** (au nombre de 11, voir ***Catégories*** en ***page 8***) dans ce document montre également que le français se classe entre la **2**e et la **9**e, mais plus souvent entre la **3**e et la **6**e place, si on ne comptait que les locuteurs L1, et entre la **2**e et la **7**e place, mais plus souvent entre la **2**e et la **4**e place, si on comptait également les locuteurs L2.

La somme de l'ensemble des **catégories** convertie en pourcentages globaux donnent le français **4**e langue de l'Internet avec un **5,61 %** de l'ensemble collecté concernant seulement la langue maternelle et **7,51 %** si on prend en compte également la langue seconde. Voir ***Annexe – Analyse par Catégories*** pour des informations détaillées.

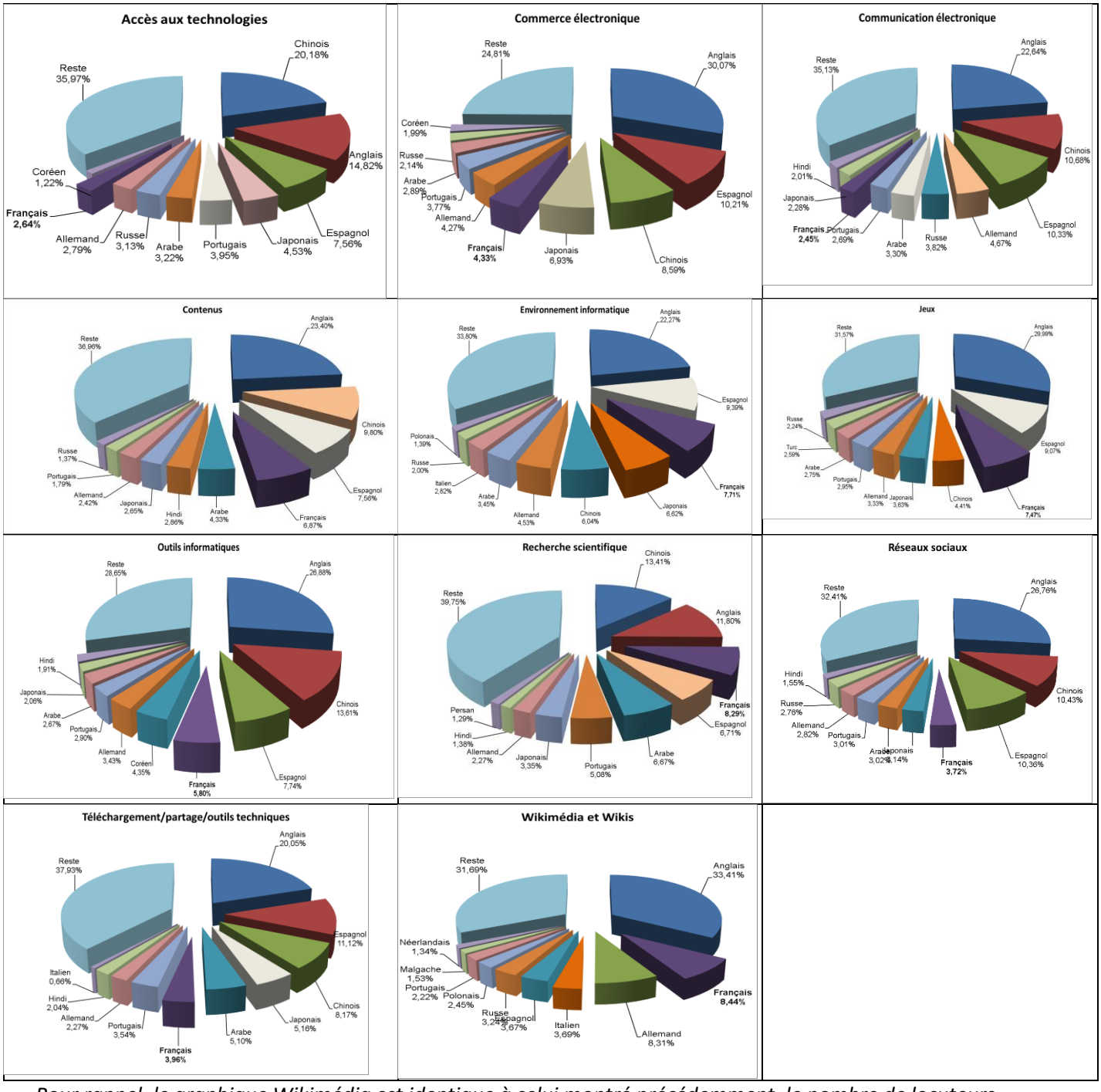


Le graphique suivant, montre la comparaison du comportement des principales langues par catégories entre L1 et L1+L2.



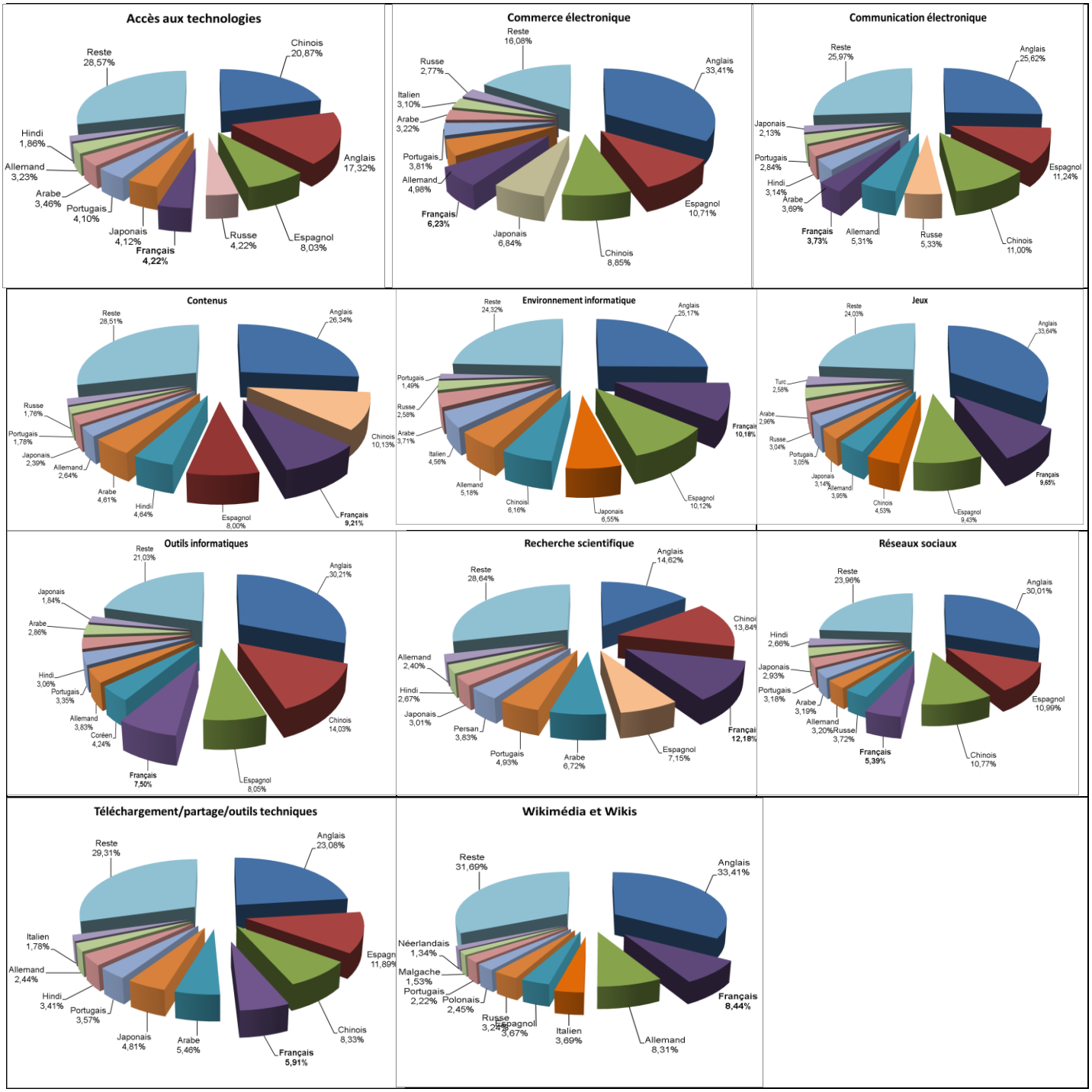
Les graphiques suivants montrent la répartition par langue des espaces occupés, le premier concernant les usages en prenant en compte les locuteurs de langue maternelle (L1), le second concerne l'ensemble des francophones (L1+L2).

Par catégories - Langue maternelle L1)



Pour rappel, le graphique Wikimédia est identique à celui montré précédemment, le nombre de locuteurs pris en compte n'affecte pas les résultats

Langue maternelle et langue seconde (L1+L2)



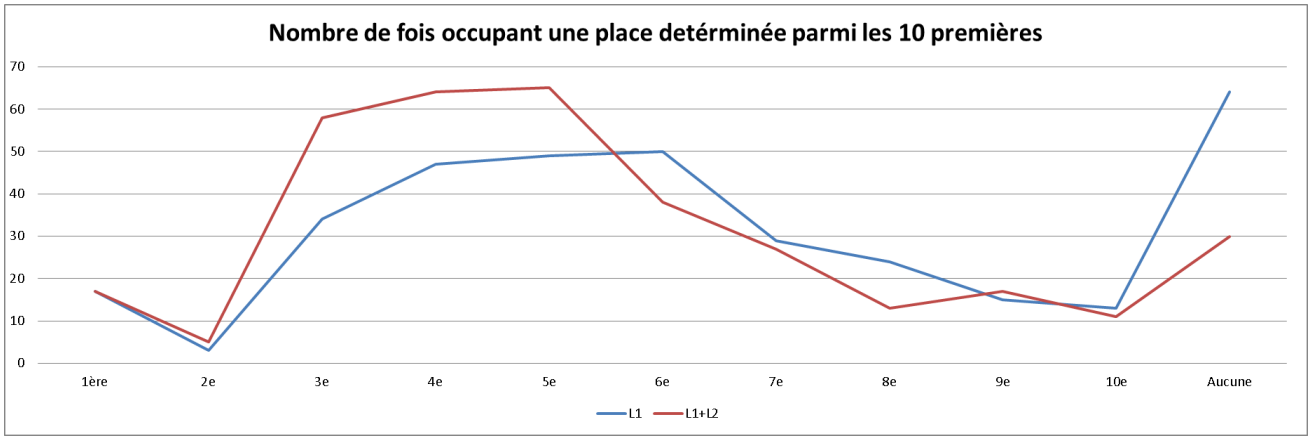
Nombre de fois où la langue française figure parmi les 10e premières langues des UM étudiées

La langue française figure parmi les 10 premières dans **281** espaces étudiés sur **344** en si on ne comptabilise que les **locuteurs de langue maternelle** (par la suite, **L1**) et dans **315** indicateurs si on comptabilise également les **locuteurs de langue seconde** (par la suite **L1+L2**).

Ceci la classerait **5e** parmi les langues trouvant ***au moins une place parmi les 10 premières***, pour **L1**, derrière **l'anglais** (340), **l'espagnol** (334), **le chinois** (326), **l'arabe** (304) et devant **l'allemand** (267), **le portugais** (257), **le japonais** (252), **le hindi** (172) et **le russe** (144).

Et le français occuperait la **4e** place des langues trouvant ***au moins une place parmi les 10 premières***, si on comptait également **L1+L2**, derrière **l'anglais** (344), **l'espagnol** (338), **le chinois** (327) et devant **l'arabe** (311), **l'allemand** (263), **le portugais** (255), **le japonais** (207), **le hindi** (217) et **le russe** (188).

Voir ***Annexe – Résultats complémentaires*** pour plus de détails

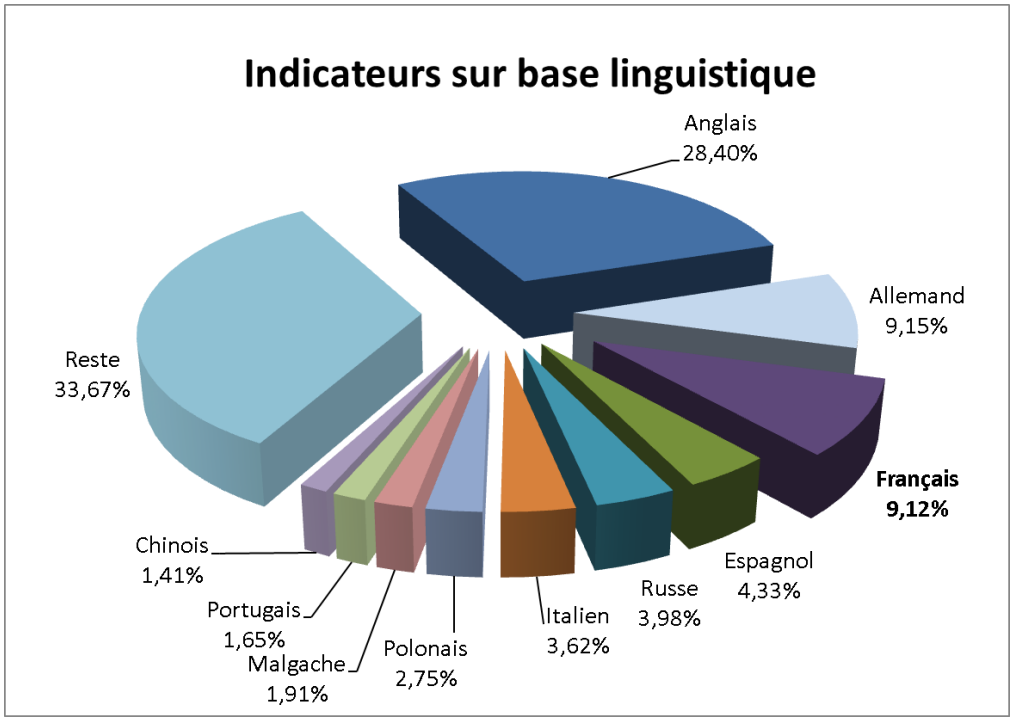


Place de la langue française par UM linguistique

Un groupement spécifique est à retenir, celui lié **exclusivement** aux mesures linguistiques (indicateurs non liés à l'usage, mais aux contenus et leurs éditeurs), et qui montre que la langue française est **troisième**, talonnant de près l'allemand (2e derrière l'anglais) et loin devant l'espagnol, le russe et l'italien, comme le montre le graphique suivant. Sur la base des 8 « unités de mesure » linguistiques étudiées, sont produits en français **9,12 %** des contenus.

Ce regroupement qui, du reste, est celui qui tire le français vers le haut dans nos différentes mesures, est néanmoins celui qui est issu du plus petit nombre d'indicateurs, étant couvert majoritairement par le monde *Wikimédia*, ainsi que par *W3Techs* et *Amazon*.

Bien évidemment, ces indicateurs ne sont pas affectés par le nombre de locuteurs, car ils ne concernent que les contenus et pas leur utilisation. Les calculs sur ces indicateurs n'incluent pas les variables de L1 ou de L2



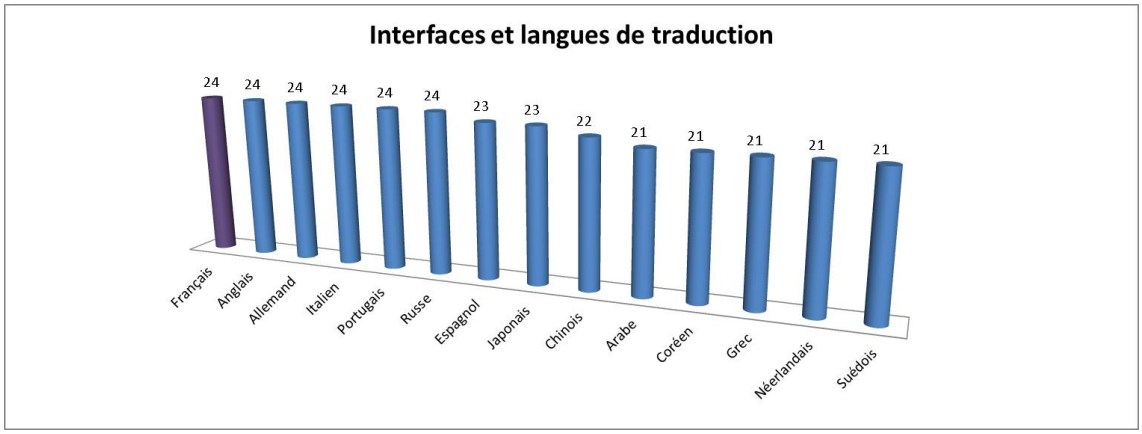
Voir ***Annexe – Résultats complémentaires*** pour plus de détails.

Place de la langue française, langue d'interface et de traduction

Finalement, un recensement des langues d'interface et de traduction sur 24 applications très suivies fait ressortir que **le français** est, avec l'anglais, l'italien, le portugais et le russe, **la langue la mieux servie** car toutes ces applications proposent une **interface en langue française ou une version française en tant que langue traduite**. Suivent le japonais et l'espagnol avec 23 applications, le chinois avec 22 et l'arabe, le coréen, le grec, le néerlandais et le suédois complètent le tableau des langues les mieux servies. Les locuteurs de 273 langues peuvent disposer d'au moins une interface d'applications en ligne dans leur langue, voire servir de canal de traduction. Ces applications sont :

Langues d'interface :
Cortana, DMOZ, DuckDuckGo, Facebook, Google, Google Scholar, Skype, Telegram, Wikilivres, Wikipédia, Wikiquote, Wikisource, Wikiversité

Langues de traduction :
Bing, Dictionnary.com, Duolingo, Free-translator, Google Traduction, IM Translator, On-Line, Reverso, SDL, Systran,.



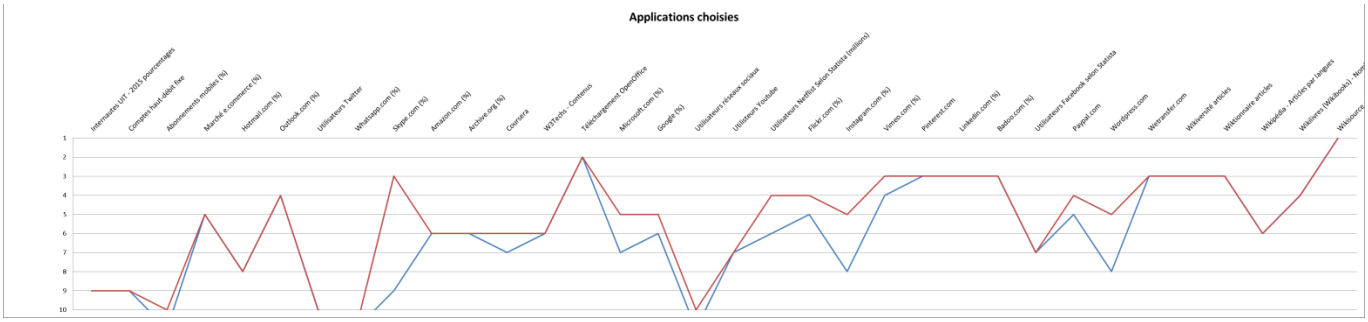
Voir ***Annexe – Résultats complémentaires*** pour plus de détails

Place de la langue française dans certaines UM choisies (L1+L2)

Tous les cadres des UM étudiées sont présentés en annexe (comme toute l'information graphique et statistique qui accompagne ce rapport), mais il est probablement utile de mettre en lumière certaines applications « phares », des espaces d'un intérêt particulier ou d'usages qui sortent de l'usuel.

Ces UM ne sont donc pas prises au hasard, mais leur choix n'est pas moins arbitraire. Elles ne nous donneront pas plus d'indications d'ensemble que les chapitres précédents, mais nous éclaireront sur certaines pratiques des francophones et des langues majoritaires.

Voici un graphe (consultable de manière plus confortable en Annexe également), qui montre la courbe de comparaison entre les deux extrêmes de la fourchette, la courbe rouge correspondant à la partie haute, la bleue à la partie basse (L1 seulement)



Si on se plaçait sur la partie haute de la fourchette (la partie basse n'étant pas évoquée ici mais consultable en ***Annexe – Quelques applications ou espaces au choix***), nous constaterions que la langue française, aurait une présence non négligeable en matière de **nombre d'internautes (9^{e})**, de **comptes mobiles (9^{e})** et de **haut débit mobile (8^{e})** et qu'elle se placerait très bien en matière de **commerce électronique (5^{e}).**

Au niveau des applications, si le français est **absent des premières 10 langues** pour l'utilisation de la messagerie (texte, images et voix) **Whatsapp**[69] et figure en **10e** position en matière d'utilisation de **Twitter**, les francophones semblent beaucoup utiliser des applications comme **Skype, Vimeo, Pinterest et Linkedin (3e place)**, voire **Flickr ou Netflix (4e)**. D'autres applications comme **Instagram (5e)**, **Youtube (6e)** ou **Facebook (7e)** sont très utilisées par les francophones, ainsi que des services comme la messagerie **Outlook (6e)** et l'ensemble de services offerts par **Google (5e)**.

Si la suite **OpenOffice (2e)** semble être très appréciée et téléchargée, les francophones n'en utilisent pas moins les nombreux services de **Microsoft (5e)**.

5e langue d'utilisation de **Wordpress**, le français est la **6e langue d'utilisation d'Amazon** (livres), **Archive** (historique du Web) et **Coursera** (MOOC), et **6e** aussi dans l'évaluation que présente le site de contenus de **W3Techs**.

En dernier, l'univers Wikimédia est très prisé par les francophones car la langue française est la **première** en matière de sources (**Wikisource**), la **2e** en matière de dictionnaires (**Wiktionnaire**) et de cours d'université (**Wikiversité**), **4e** pour le service **Wikilivres** et **6e** pour **Wikipédia** où, rappelons-le, le français a perdu 3 places en quelques années car elle était la 3e langue il n'y a pas si longtemps.

Les cadres consultables en Annexe, donnent un aperçu des langues qui accompagnent le français parmi les 10 premières langues.

Évolution de la langue française

Il aurait été souhaitable de pouvoir comparer l'évolution au fil des ans, mesurant des objets identiques avec les mêmes paramètres. Nos précédentes études, basées sur une même base méthodologique mesurait un nombre important d'objets observés aujourd'hui, mais ce ne sont hélas pas les mêmes paramètres. En effet, d'une part, un bon **nombre d'objets étudiés par le passé n'existent plus**. D'autres que nous observons aujourd'hui **n'existaient pas auparavant**, mais pour un bon nombre de ces objets, **la source d'observation a changé**.

En effet, comme nous l'avons expliqué, en 2010 et en 2013 il était aisé alors de trouver des sources disponibles mesurant convenablement les applications, usages et espaces liés au cyberespace. Le panorama a beaucoup changé depuis et nous avons dû nous rabattre sur un petit nombre de sources, rendant les statistiques peu comparables entre elles.

Ainsi, nos calculs en 2013 affirmait que la langue française serait la *« **4ème** si l'on comptabilisait l'ensemble des locuteurs francophones (L1+L2) et entre la **7ème** et la **8ème** si l'on ne comptait que les locuteurs de langue maternelle »*

[69] Rappelons qu'il y a une corrélation entre le succès de cette application et des tarifs locaux en matière de télécommunication.

Nous constatons aujourd'hui un élan majeur car elle est toujours **4e** en **L1+L2**, mais aussi elle figure entre la **4e** et la **5e** selon le type de mesure que l'on prend et ceci même en ne comptabilisant que la somme des locuteurs natifs (L1).

Or, les **28 UM** (applications, espaces ou usages) que nous pouvons comparer aujourd'hui avec ceux de 2013, ne reflètent en rien cette progression, bien au contraire, car la plupart de comparaisons ponctuelles donnent un résultat inverse à celui reflété par l'ensemble des paramètres.

Ceci est perçu dans les graphiques suivants, mais la comparaison isolée de ces **28** mesures ne peuvent pas nous apporter une grande information, par rapport à l'ensemble de **344** UM analysées aujourd'hui. Par contre, ces graphiques nous apportent des informations considérables sur la progression ou évolution dans l'utilisation de certaines applications par un public francophone.

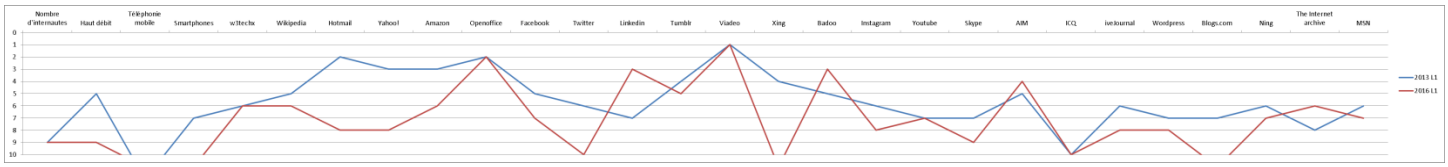
Les deux graphiques suivants (plus explicites en ***Annexe – Évolution de la langue française***), nous montrent une courbe bleue correspondante aux mesures prises en 2013 et une courbe rouge correspondant aux mesures prises en 2016, le premier tableau comparant la progression pour L1, le second, pour L1+L2. Nous voyons ainsi que la plupart d'entre elles montrent des résultats plus intéressants en 2013 qu'en 2016.

En effet, pour « L1 », à peine **4 UM sur 28** ont une valeur **supérieure** en **2016** et **6** one une valeur **identique**, les **18 autres** auraient des **valeurs supérieures en 2013** à celles constatées **aujourd'hui**.

Pour **L1+L2**, **6 UM** ont une valeur **supérieure en 2016** et **5 ont une valeur identique**, les **17 restantes** auraient des valeurs **supérieures en 2013**.

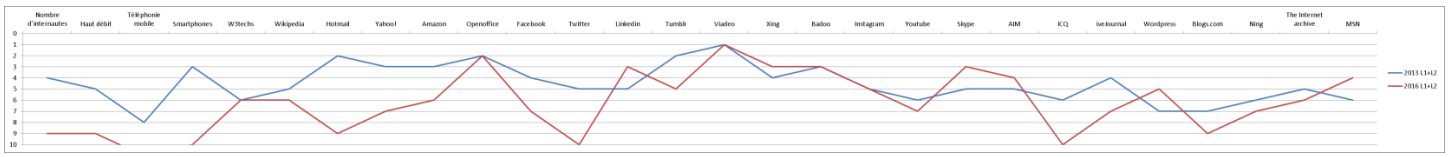
Comparaison 2013-2016 de la place de la langue française pour 28 UM étudiées. Graphique pour L1.

Ce graphique peut mieux être apprécié en Annexe – Évolution de la langue française



Comparaison 2013-2016 de la place de la langue française pour 28 UM étudiées. Graphique pour L1+L2

Ce graphique peut mieux être apprécié en Annexe – Évolution de la langue française



Page Internet

Dans le cadre de ce projet, il a été conçu un système informatique et une page Internet qui devrait être intégrée au site de l'Observatoire de la langue française de l'OIF permettant aussi bien l'édition autonome des données (saisie, importation et exportation des données) que la publication dynamique des résultats en ligne, graphismes inclus.

Des synthèses périodiques élaborées par le système et décrivant de manière graphique les évolutions permettront d'évaluer la mise en place d'instruments d'aménagement linguistique ou d'encouragement à combler des lacunes et inverser le recul –voire absence- de la langue française sur des espaces ou des applications stratégiques.

La remise de cette page est prévue avec les annexes de ce projet, mais elle sera mise en service définitivement à la fin de la 2^{e} mesure que cette étude doit effectuer.

L'ensemble de la page réunit les caractéristiques suivantes :

- Une page principale avec la synthèse des pages pouvant être consultées, accessibles toutes par un simple click, montrant les graphiques et cadres essentiels.
- des pages spécifiques contenant les graphiques, cadres et explications sur :
    - la place qu'occupe la langue française dans son ensemble par rapport aux autres grandes langues de communication avec l'explication pertinente;
    - idem pour les « catégories »
    - idem pour les « segments »
    - Idem pour un large extrait de l'ensemble des espaces étudiés
- un tableau évolutif par rapport à la dernière étude pour chaque type espace étudié;
- des pages extraites de ce rapport ou méritant d'être diffusés
- Une page de conclusions de l'étude (pouvant pointer vers des pages secondaires pour éclairer certains concepts ou accéder aux annexes)
- Une page décrivant la méthodologie utilisée (pouvant pointer vers des pages secondaires pour éclairer certains concepts ou accéder aux annexes)
- Une page de références concernant les sources utilisées pour l'étude.
- Une page spécifique permettant la réaction des utilisateurs concernant la prise en compte de nouvelles données, de nouveaux espaces, propositions d'amélioration, etc.

La structure permet le rajout ou l'élimination d'espaces d'étude et de sources d'information en fonction de l'obsolescence des sources et de l'apparition de sources nouvelles. Tout cela se fait de manière transparente pour le public et peut montrer clairement l'évolution des applications et des usages en relation avec la langue française

Le site est constitué par des bases de données qui servent de structure de conservation organisée pour les données procédant des différentes sources, de manière à rationaliser les moyens, systématiser le traitement de l'observation et faciliter une mise à disposition des informations en temps réel qui deviendrait ainsi une rubrique supplémentaire de l'Observatoire de la langue française. Ces informations adopteront le format de données ouvertes.

Cette page internet est prévue comme plateforme évolutive et ouverte, mais elle nécessite néanmoins des changements informatiques futurs pour une parfaite autonomie.

En effet, elle est pour l'instant dépendante d'un système de classeurs (du type Excel) pour la réalisation des calculs qui sont devenus, au fil de l'avancement du projet, d'une grande complexité. Cette complexité a été résolue par le système de feuilles de calcul, mais elle n'a pas pu être implémentée au niveau des bases de données du système,

Note : Un cahier d'opérateurs sera rendu à l'OIF à la fin de la deuxième mise à jour des données, de manière à prendre en compte toutes les opérations devant être réalisées.

Quelques remarques finales

Si la conclusion finale de ce rapport est la satisfaction de voir l'importante progression de la langue française en termes globaux, notamment si on prend la partie basse de la fourchette considérée dans ce document, il est important de noter, pour la suite, ces considérations finales ponctuelles :

- Cette étude bénéficie des leçons apprises lors des études passées et de l'importance de ce nouvel effort consenti par l'OIF. La volonté d'obtenir un niveau de qualité inégalable dans ce sujet a par ailleurs pu provoquer des retards importants par rapport au calendrier initialement prévu.
- Ce retard implique que certaines obligations seront rendues après les termes prévus du protocole d'accord pour des raisons de calendrier entre la première mesure et la seconde.
- Concernant la page Internet, il est important de noter qu'il est conseillé d'optimiser les aspects d'ingénierie logiciel de manière a la rendre moins dépendante des calculs des classeurs utilisés aujourd'hui
- Concernant les sources d'information d'usage de l'Internet, il serait souhaitable que, dans le futur, le budget d'abonnements soit revu à la hausse, car les auteurs de ce rapport ont sous-estimé la marchandisation des données opérée ces dernières années et les montants sollicités (et concédés) n'étaient pas à la hauteur des besoins.
- En particulier, et malgré nos réserves exprimées sur l'usage d'Ethnologue, il est important d'affirmer que le plus raisonnable en termes de futur est d'acquérir l'accès à cette base des données car elle semble être la plus stable en matière d'informations démolinguistiques, d'autant plus que c'est là sa vocation première.

Finalement, il est particulièrement intéressant de noter que, parmi les à-côtés de cette étude, des enseignements importants sur la présence de l'ensemble des langues de la planète peuvent être tirés. Notamment pour les 620 langues qui occupent au moins une place parmi les 100 premières.

La Francophonie pourrait profiter de cette initiative pour apporter des informations de haut niveau en faveur de la diversité linguistique mondiale.

Notes de fin de document

[I] **Sur la date des sources.** 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. Il n'est pas possible dans le cadre de cette étude de vouloir analyser ni procéder à la mise à niveau des données à l'intérieur des pays. Et il est clair que nous avons dû, assez souvent, croiser des données ayant été collectées à des dates différentes. Ainsi, les données sur le nombre d'internautes par pays sont, par exemple, de 2015 mais les données démographiques et démolinguistiques par pays paraissent être plus récentes.

[II] **Révision des sources. Exceptions.** Il existe cependant quelques cas pour lesquels les auteurs se sont permis d'améliorer les sources premières quand elles préfiguraient des données essentielles dans tout le processus: c'est le cas des données démolinguistiques par pays et des données de pourcentages de personnes connectées par pays. Les deux cas sont décrits en détail dans d'autres chapitres du présent document.

Une autre exception a été, par exemple, dans le cas des données d'abonnement à Netflix offertes par Statista. Nous disposons des données actuelles et d'une projection à 2021. La découverte d'une diminution dans la projection des États-Unis, pays de loin leader dans ce segment, nous a amené à vérifier 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 entrainait l'élimination de ce segment, sauf à obtenir une source fiable par ailleurs sur la croissance aux États-Unis de ce marché, ce qui fut obtenu grâce à l'assistance de Statista. Nous avons donc complémenté l'échantillon de données du segment avec la projection pour 2021 des abonnements aux États-Unis de manière à conserver une donnée exploitable

[III] **Sur les sources démolinguistiques.** Des statistiques fiables existent pour certaines langues parlées dans des territoires bien délimités et connaissant un développement important, à condition qu'elles aient un statut d'officialité ou une protection quelconque et que leur diaspora parlant encore la langue soit bien étudiée. C'est le cas de beaucoup de langues parlées en Europe, comme par exemple, l'italien, l'islandais, le polonais ou le catalan. Mais cette tâche devient compliquée pour des langues qui ne réunissent aucune de ces conditions.

Pour prendre des exemples qui n'affectent pas nécessairement notre étude, mais qui sont parlants pour le public pouvant les consulter, des langues comme l'occitan ou le piémontais, tout en étant parlées dans des régions développées et sur des territoires assez bien délimités, le manque d'institution de tutelle (ou le manque de moyens de celles-ci si elles existent) nuit à la qualité des chiffres concernant le nombre de locuteurs.

Pour les langues occupant des territoires étendus et de conditions socio-économiques diverses, il existe des statistiques assez fiables (bien que se réservant une marge d'erreur) pour certaines et peu crédibles ou vérifiables pour d'autres. La langue française ou la langue espagnole, par exemple, grâce respectivement à l'OIF et à l'Institut Cervantès sont bien répertoriées. Mais pour d'autres langues présentant des caractéristiques similaires, comme les langues anglaise, chinoise ou hindi, nous avons rencontré des divergences trop importantes pour préférer une source à d'autres, notamment sur les locuteurs « L2 ».

À ces complications s'ajoute la divergence des méthodologies dans le comptage des locuteurs des innombrables études démolinguistiques. Pouvoir mettre en comparaison toutes les études existant sur le nombre des locuteurs de l'ensemble des langues de la planète afin de créer une « super-grille » comparative est une tâche titanesque que les auteurs de cette étude ne peuvent pas mettre en œuvre. Et il n'est pas pensable de mettre en parallèle le résultat des différentes études sans une analyse fine car on mettrait sur le même plan des chiffres non comparables car obtenus par des définitions ou des méthodes non homogènes.

Sur la typologie des langues. Un problème supplémentaire est à résoudre en matière de typologie de langues. Pour revenir sur l'exemple proche de l'occitan, est-ce que les locuteurs de gascon doivent être comptabilisés comme des locuteurs de langue occitane ou indépendamment de celle-ci? Certains approuveraient, d'autres moins. Est-ce que la langue allemande doit être prise dans son ensemble ou dans ses différentes formes dialectales, parfois très éloignées les unes des autres? Que faire avec la langue arabe, souffrant aussi d'une grande dispersion dialectale? Faut-il considérer l'arabe littéraire ou classique uniquement ou l'ensemble d'arabes dialectaux ? Les auteurs, en prenant la norme ISO 639-3, tout en étant conscients qu'elle peut présenter de nombreuses anomalies et se prêter à des critiques, et dans l'incapacité de prendre eux-mêmes des décisions linguistiques, adoptent la voie la plus diffusée à l'heure actuelle, et accepté par les autres sources mondiales consultées (notamment Wikipédia et Ethnologue, bien entendu).

[IV] **Caractère religieux du site Projet Josué** Sur cette décision, il est éclairant de la mettre en parallèle avec celle prise Louis-Jean Calvet pour son baromètre proposé aujourd'hui sur la page WikiLF [IV] du Ministère français de la culture[IV], où le dilemme de prendre Ethnologue ou Josué était déjà évoqué. Josué avait été écarté pour son caractère religieux au profit d'Ethnologue, et ce malgré les réserves évoquées sur celui-ci, que nous partageons. Nous nous sommes aussi posé la question de l'opportunité de mettre en lumière un site à visés « évangélisateurs », ce qui pourrait prêter à confusion sur l'intention de notre travail, mais en attendant un accès plus aisé à Ethnologue, la praticité l'a emporté sur la non-neutralité idéologique du Projet Josué.

[V] **Sur la disponibilité de statistiques du Projet Josué.** Josué propose de télécharger des fichiers en format Excel ou Access. Nous avons trouvé, non seulement des différences dans le temps, mais aussi entre ces fichiers téléchargés à une même date. Par exemple, le hindi avait une différence de 300 millions de locuteurs de langue maternelle entre deux versions différentes du fichier Excel.

[VI] **Sur le corpus de langues retenues au départ.** Les possibilités de créer un corpus réduit dès le départ étaient :

1. *Langues ayant un statut d'officialité.* Ceci aurait exclu de facto nombre de langues de la planète qui sont vivantes dans l'Internet. D'après Wikipédia

(https://fr.wikipedia.org/wiki/Liste_des_langues_officielles), 172 langues ont un statut d'officialité ce qui aurait exclu du comptage un bon nombre de créoles, le peul, le corse ou l'occitan non aranais (variante de l'occitan ayant un statut d'officialité au Val d'Aran, Catalogne, Espagne), par exemple

2. *Langues ayant un nombre minimum de locuteurs L1.* Fixer un minimum (10 millions ?, 5 millions ?) aurait exclu un nombre important de langues notamment européennes (islandais, lituanien, letton, etc.) qui ont une vitalité reconnue dans l'Internet.
3. *Langues ayant un nombre minimum de locuteurs L2.* Ça aurait signifié ne garder qu'un trop faible échantillon de quelques dizaines de langues.
4. *Langues de Wikipédia.* Nombre de langues de Wikipédia sont artificielles ou éteintes et certaines n'ont qu'un très faible nombre de locuteurs, alors que d'autres biens plus importants n'y figurent pas. Il y a plusieurs raisons à ne pas intégrer de facto Wikipédia:
    - il y a beaucoup de langues artificielles ou mortes dans Wikipédia, pas compatible avec notre étude où le chiffres des locuteurs compte beaucoup.
    - il y a de toutes petites langues (faible démographie) qui y figurent alors qu'il y a d'autres langues démographiquement importantes qui n'y figurent pas. Si la raison de cette présence était la vitalité de ces langues, ce serait bien de les mentionner, mais souvent (c'est Wikimédia qui le dit), il ne s'agit que des volontés individuelles et souvent partiales (pour véhiculer une idée politique, une religion, etc.).
    - Il y a plusieurs dialectes et non pas de langues, ce qui est incompatible avec notre taxonomie de départ.
5. Enfin, une règle combinée aurait pu être aussi définie, mais la méconnaissance des réalités de terrain des continents asiatique et africain des auteurs aurait probablement causé des distorsions d'analyse.

[VII] **Pays et territoires traités.** Ont été fusionnés avec leur « métropole de référence » certains territoires ou domaines d'outre-mer de la France (Guadeloupe, Guyane, Saint-Martin, Martinique, Mayotte, Nouvelle-Calédonie, Polynésie française, La Réunion, Saint-Pierre-et-Miquelon et Wallis-et-Futuna) mais aussi des territoires, régions autonomes ou déléguant une certaines souveraineté à l'Australie (Îles Cocos, Christmas et Norfolk), la Chine (Hong-Kong et Macao), le Danemark (Iles Féroé et Groenland), les États-Unis (Guam, Marianne du Nord, Porto Rico, îles Vierges américaines, Samoa) , la Nouvelle-Zélande (Iles Cook, Niue et Tokelau), les Pays-Bas (Aruba, Bonaire, Saint-Eustache et Saba, Curaçao, Saint-Martin), la Norvège (Svalbard et île Jan Mayen), le Royaume Uni (Anguilla, Bermudes, îles Vierges britanniques, îles Caïmans, Gibraltar, Guernesey, Jersey, îles Malouines, île de Man, Montserrat, île Pitcairn, Sainte-Hélène, Ascension et Tristan da Cunha, Saint-Christophe-et-Niévès, Territoires britanniques de l'océan Indien et les îles Turques-et-Caïques).

[VIII] **Type de données Index.** Les calculs tels qu'ils avaient été appliqués sur les index les rendaient inutilisables ou sujets à une mauvaise interprétation, car ils aboutissent à une valeur qui représenterait une sorte de pondération d'un service ou note attribuée et ne se correspond pas avec la nature « démographique et démolinguistique » de l'approche. La méthodologie appliquée pour

toutes les autres UM d'usage, à savoir le croisement avec les informations démolinguistiques, donnait des résultats incongrus alors que cela avait toute logique pour des indicateurs où interviennent des ensembles d'individus.

Voici un exemple éclairant : Si la France a une valeur 100 pour l' « UM e-gouvernement » , ceci est vrai pour l'ensemble de la population française, peu importe la langue utilisée. Si parmi les autres pays francophones l'on trouve des valeurs inférieures, en croisant avec les informations démolinguistiques, ceci ferait descendre la note pour la langue française. Mais, la langue bretonne, par exemple, héritant de la note de la France (100) et n'ayant pas de locuteurs dans un autre pays, se serait trouvé avec une note 100, soit supérieure à celle de la langue française. Une simple lecture pourrait faire croire que les locuteurs de langue bretonne sont mieux servis en e-gouvernement que les locuteurs de langue française.

[IX] **Extrapolations.** L'observation des résultats des extrapolations effectuées ont montré une cohérence constante, sauf dans le cas exceptionnel des projections de Netflix pour 2021 où le pays ayant la majorité des abonnements n'était pas renseigné (!) et que l'extrapolation a pu compenser.

[X] **Extrapolations retenues**. La méthode d'extrapolation retenue (« au prorata des pourcentages de personnes connectées par pays ») s'applique quand il est raisonnable de considérer que les valeurs de l'UM considérée sont naturellement proportionnelles au pourcentage mondial de personnes connectées, par exemple, dans le cas, majoritaire, des mesures de trafic des sites et des applications. Ainsi, 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 l'UM en question. Le reste extrapolé des quantités non renseignées ou le pourcentage restant, selon le cas, sera réparti au prorata des poids respectifs en termes d'internautes des pays restants.

La méthode dite des quartiles a été écartée de ce rapport, mais voici sa description initiale : *« Nous avons attribué aux pays non renseignés un des quartiles des données des pays renseignés en fonction du pourcentage de personnes connectées. Après plusieurs essais, il nous est apparu opportun de déterminer ainsi l'attribution des quartiles:*

- *Si le pays a moins de 15% de sa population connectée : premier quartile (note la plus basse)*
- *Si le pays a plus de 35% mais moins de 65% de sa population connectée : deuxième quartile*
- *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 : dernier quartile (note la plus haute). »*