Une histoire brève de l'observation des langues dans l'Internet

Daniel Pimienta, Observatoire de la Diversité Linguistique et Culturelle dans l'Internet, mai 2022

La mesure de l'espace de représentation des langues dans l'Internet[1] ne passionne pas les foules et pourtant les enjeux, sur les plans linguistique, culturel, socio-économique et même géopolitique, sont loin d'être neutres. Beaucoup de langues sont menacées ou simplement en déclin et l'intensité de leur présence dans l'Internet est un indicateur déterminant de leur futur. Le commerce électronique brasse, en 2020, 20% du total des ventes mondiales du commerce de détail et les plateformes doivent parler la langue de leurs clients pour rentrer dans cette compétition.

Une légende parcourt l'Internet depuis ses origines à propos d'une exclusivité étatsunienne qui s'est également traduite en la croyance d'une domination stable et pérenne, qui ferait de l'anglais à jamais la prétendue *lingua franca* dans le cyberespace.

Ce survol historique prétend débusquer la **mésinformation** à ce sujet, source de renoncements regrettables quant à un objectif que tous les acteurs du développement reconnaissent essentiel : l'importance de la **création de contenus en langue locale et de politiques publiques visant à en favoriser les conditions (lutte contre la facture numérique, accompagnée de programmes de littératie numérique)**.

Cet itinéraire dans le temps empruntera le chemin particulier d'une organisation pionnière en la matière et qui, à travers de nombreux aléas, en est restée, jusqu'à aujourd'hui, un acteur notable : *l'Observatoire de la diversité linguistique et culturelle dans l'Internet.*

Cet observatoire a débuté en 1995, comme un projet d'une ONG (organisation non gouvernementale) de recherche-action et de terrain, dont le nom exprime l'essence de la vision : Association Réseaux et Développement (FUNREDES pour son sigle en espagnol)[2]. Ce projet est devenu ensuite un véritable programme[3], de par la pérennité de ses actions, entre 1996 et 2017, date laquelle cette ONG s'est dissoute ; l'Observatoire a repris par la suite une forme associative indépendante, tout en revendiquant et maintenant son héritage historique[4].

FUNREDES s'est constituée, en 1993, à partir d'une matrice originelle créée au sein de l'Union Latine, après que l'auteur a réussi à convaincre, en 1987, son visionnaire secrétaire général, Philippe Rossillon, inquiet de l'association étroite entre anglophonie et réseaux informatiques, que les réseaux allaient connaitre un destin fulgurant et qu'il importait de défendre la diversité linguistique et culturelle de l'intérieur plutôt que de s'opposer à un envol irrépressible.

Entre 1988 et 1993, quand ce programme de l'Union latine est devenu indépendant, se transformant en FUNREDES, tout en maintenant des liens harmonieux et actifs avec l'Union latine, une activité importante s'est développée, avec, pour certaines, le soutien de l'Union Européenne et la collaboration de l'UNESCO. Trois réseaux nationaux ont été créés (Pérou,

[1] Les deux indicateurs de base sont, pour chaque langue, le pourcentage mondial de locuteurs connectés et le pourcentage de contenus sur la Toile.
[2] Fundación Redes y Desarrollo (http://funredes.org)
[3] http://funredes.org/lc
[4] http://funredes.org/lc/JO-OBDILCI.pdf

République Dominicaine, Haïti) ainsi que le premier logiciel multilingue d'interface aux réseaux depuis un ordinateur personnel (MULBRI[5]), actions pilotes d'un grand projet de réseau latinoaméricain (REDALC[6]), avec, comme originalité, une vision centrée sur l'utilisateur, sa culture et sa langue et l'importance de la littératie numérique ainsi que l'intégration des professionnels de l'information ; tout cela en contraste avec un contexte où la vision technologique était très largement majoritaire et transversale.

FUNREDES prenait donc le relais, en 1993, et résistait à une vision de la fracture numérique simpliste, qui consisterait seulement en une question d'accès technologique, et poussait les concepts de **fracture de contenus et fracture linguistique**, en partant du principe d'un besoin naturel de naviguer dans le cyberespace dans sa langue maternelle et de l'évitement du piège de l'acculturation, ainsi que de l'importance d'associer la **littératie numérique** à toute politique de lutte contre la fracture numérique (Pimienta, 2007).

C'est l'expression du Président Chirac, lors du sommet de la Francophonie à Cotonou, en 1995, qui voyait dans l'Internet naissante une entité entièrement anglophone et étasunienne qui provoquait le désir de confronter ce préjugé avec la réalité, en tentant d'y mesurer la place des langues et des cultures. Rapidement, l'Union latine, à travers son Directeur pour la terminologie et les industries de la langue, Daniel Prado, s'associait à la démarche et une longue et fructueuse collaboration commençait.

À cette époque, les moteurs de recherche rapportaient fidèlement le nombre d'occurrences d'un mot ou d'une expression dans l'ensemble des pages Web indexées, lesquelles représentaient une proportion des pages existantes supérieure à 80%. C'était donc un outil formidable pour de telles études.

Pour les langues latines[7], l'anglais et l'allemand, un échantillon de mots fut constitué, dans chaque langue, soigneusement sélectionnés pour représenter un ensemble sémantiquement et syntaxiquement équivalent, une tâche plus facile à dire qu'à réaliser. Pour la culture, une liste d'un vaste ensemble de personnages était constituée dans une série de catégories (lettres, science, musique, cinéma…). Les mesures d'occurrences permettaient, avec des outils statistiques traditionnels, d'obtenir des résultats dans les deux champs, langue et culture[8].

La proportion de l'anglais dans la Toile était mesurée entre 75% en 1997 et 52%, en 2001. Les personnalités latines étaient bien représentées dans les secteurs culturels où la **séparation entre commerce et culture** étaient la plus marquée ; par contre, dans les secteurs régis par les lois du marché, la culture des États-Unis s'imposait nettement. Les résultats sur la culture se stabilisaient mettant en exergue les personnages les plus « mondialisés » et l'étude s'interrompaient après la troisième campagne, en 2001[9].

Quant aux résultats sur les langues, ils devaient s'interrompre en 2007 car les moteurs de recherche, Google en premier, avaient évolué de manière incompatible avec la méthode utilisée : les retours du nombre d'occurrences perdaient en crédibilité, la couverture de l'index se rétrécissait énormément, pouvant descende en dessous des 5% de l'ensemble des pages de la Toile, et le jeu publicitaire commençait à pervertir les résultats des moteurs de recherche qui

[5] https://funredes.org/gopher/b/6/6.4/6.4.3/6.4.3.2/6.4.3.2.2/lb.html
[6] https://funredes.org/gopher/b/6/6.1/6.1.1/6.1.1.1/lg.html
[7] Espagnol, français, italien, portugais et roumain.
[8] Voir https://funredes.org/lc2005/francais/index.html. Pour une synthèse consulter (Pimienta, 2001).
[9] https://funredes.org/lc2005/francais/index.html

devenaient différents pour chaque utilisateur, en fonction de leurs recherches passées et d'autres facteurs liés aux informations personnelles récoltées par Google. Cela faisait de ce projet un des premiers témoins des effets de l'évolution toxique des moteurs de recherche et de l'inversion en cours où « *tel est recherché qui croyait chercher* », marque du *capitalisme de surveillance* naissant, lequel allait s'imposer et changer la trajectoire de l'Internet, de l'utopie de démocratie participative initiale (Pimienta, 2005) vers la situation actuelle de menace contre les démocraties (Pimienta, 2020).

La préhistoire : jusqu'en 1997

Il est utile de comprendre le contexte de cette période en ce qui concerne les réseaux. Le Web est né en 1992, au moment où les différents réseaux de la recherche (comme BITNET/EARN ou HP-Net) et libertaire (comme Usenet ou Fidonet) convergeaient vers le protocole Internet, pendant que le Vidéotex français rassemblait pour quelques années encore, autour du Minitel, plus d'utilisateurs que l'ensemble des réseaux mondiaux… En 1993, année considérée comme celle de l'apogée du Minitel, 6,5 millions sont installés, servant 14,5 millions d'utilisateurs, alors que l'Internet atteint un peu plus de 10 millions d'utilisateurs, le monde de la recherche devenant minoritaire.

L'Internet nait avec un système de codage des caractères à 7 bits (ASCII[10]), qui permet sans écueil de coder la langue anglaise qui n'a pas de signes diacritiques, mais handicape la plupart des autres langues qui doivent coder plus de caractères différents que les 128 permis. Il faudra attendre quelques années, avec la création du protocole MIME[11], en 1997, pour dépasser progressivement cette limite dans les courriels et dans les pages Web, jusqu'au succès du standard UNICODE[12] qui accommode les alphabets de différentes langues, en constante évolution pour localiser plus de langues.

L'ancêtre du Web, Gopher, un système simple de menus en arborescence, aurait permis une mesure facile de la place des langues mais apparemment personne n'a eu cette idée et il est permis de penser qu'à la naissance du Web, en 1992, plus de 80% des « sites Gopher », en général des universités, étaient en anglais ; de la même manière il est raisonnable de placer la part initiale de l'anglais dans le Web, à sa naissance en 1992, à 100%, même si un belge francophone, Robert Cailliau, beaucoup moins connu que Tim Berners Lee, en était un second contributeur notable (Jardon, 2019).

Les premières tentatives de mesures de la place des langues dans la Toile remontent à la période 1997-2000 et sont évoquées, avec celles de la période consécutive, 2000-2005, dans (Pimienta et al., 2009).

L'effervescence initiale : 1997-2007

Un dizaine d'acteurs se sont manifestés dans la période, certains du monde de la recherche, d'autres animés par des considérations de marketing. Si l'on se concentre sur les éléments les plus sérieux sur le plan méthodologique[13], l'idée d'une évolution de la présence de l'anglais

[10] https://fr.wikipedia.org/wiki/American_Standard_Code_for_Information_Interchange
[11] https://fr.wikipedia.org/wiki/Multipurpose_Internet_Mail_Extensions
[12] https://fr.wikipedia.org/wiki/Unicode
[13] Outre les études de l'Observatoire déjà mentionnées (1998-2007), celle de Xerox (2000), celles du projet japonais Language Observatory Project (LOP), celle du projet catalan de l'IDESCAT et les mesures indirectes

dans le Web pour la période, passant de 80% à 50%, fait sens. Mais les éléments les plus influents ne sont pas forcément les plus sérieux et trois études ponctuelles nord-américaines (en 1997, 2000 et 2003), partageant une méthodologie invalide[14], ainsi qu'une opération de marketing d'un des moteur de recherche de l'époque, Inktomi, comportant une erreur grossière, confortent dans les médias l'idée d'une présence **stable** de l'anglais autour de 80%, dans la décennie 1997-2007.

Pourtant, la stabilité est vraiment la dernière caractéristiques crédible pour un domaine en croissance à la fois exponentielle et géographique comme l'Internet ! Il faudra la publication par l'UNESCO de deux textes sur le sujet (UNESCO, 2006) et (Pimienta et al., 2009), pour que les médias projettent enfin une valeur plus réaliste et proche de 50% pour la présence de l'anglais sur la Toile. Dans la même période une entreprise, GlobalReach[15], produit, depuis 2000, des données crédibles sur la répartition des utilisateurs de l'Internet par langue.

La naissance du projet académique LOP, coordonné par Yoshiki Mikami (2005) de l'Université de Nagasaki au Japon, sous la forme d'un consortium mondial d'universités, utilisant une technique basée sur des algorithmes de reconnaissance des langues et des méthodes puissantes d'exploration de la Toile[16], donne espoir en la professionnalisation académique de ce sujet, d'autant plus que rapidement des collaborations sont entreprises entre le LOP, FUNREDES et l'Union latine, trois membres actifs du Réseau mondial pour la diversité linguistique, MAAYA[17], né en 2006, sous l'impulsion d'Adama Samassekou, en marge du Sommet mondial pour la société de l'information, et fédérateur d'actions significatives dans ce champ[18].

Cette légère effervescence sur le thème va cependant se calmer dans la période suivante 2007-2017

La traversée du désert : 2007- 2017

Le projet UPC/IDESCAT s'arrête en 2006 ; GlobalReach cesse de produire ses données en 2007 ; et en 2011, le LOP disparait, emporté par le tsunami qui affecte le Japon. Quant à l'Observatoire, dans le cadre de FUNREDES, et sous le chapeau de MAAYA, il propose, entre 2010 et 2013, un grand projet européen de recherche sur le thème et parvient, avec le soutien conjugué de l'Union Latine, de l'OIF[19] et de l'UNESCO, à créer un puissant consortium européen de recherche[20] qui répond à deux appels du programme européen de recherche[21]. Mais

---

réalisées avec Google avec la technique « du complément de l'ensemble vide ». Voir (Pimienta et al., 2009) pour plus de détails et sources.

[14] La méthode commune était de sélectionner au hasard, à partir des numéros IP, 3000 sites et de leur appliquer un algorithme de reconnaissance de langues. Pour valider cette approche il aurait fallu répéter cela plusieurs fois et traiter statistiquement les multiples résultats comme une variable aléatoire en étudiant, sa distribution (moyenne, variance, etc.). Envoyé la fléchette une seule fois au centre de la cible ne prouve pas l'habilité du tireur…

[15] https://web.archive.org/web/20000412001030/http://www.glreach.com/globstats/index.php3

[16] Le LOP ne cherche pas à explorer l'univers entier mais se concentre sur des espaces géographiques plus limités rendant possible cette exploration systématique.

[17] https://web.archive.org/web/20150704174747/http://www.maaya.org/?lang=fr

[18] MAAYA organisait quatre Symposium Internationaux sur le Multilinguisme dans le Cyberespace, en 2009, 2011, 2012 (voir https://web.archive.org/web/20150704174747/http://www.maaya.org/?lang=fr) et, en 2019, sous l'impulsion de Claudio Menezes (voir https://doity.com.br/iv-simc). À l'initiative de Daniel Prado, MAAYA réunissait plusieurs auteurs dans un ouvrage qui reste la référence sur le thème (MAAYA, 2012). Pour l'histoire de MAAYA, voir (Pimienta et Prado, 2016).

[19] Organisation Internationale de la Francophonie : https://francophonie.org

[20] Voir https://web.archive.org/web/20180831105048/http://dilinet.org/mod/resource/view.php?id=105

[21] Voir https://funredes.org/lc/dilinet/fr/.

la priorité ne semble pas fondamentale pour l'Union Européenne et l'effort reste vain, malgré une première tentative qui passe à un demi-point d'évaluation du seuil requis et un notable investissement humain et financier dans la période 2010-2012.

Il faut donc se résoudre à rester dans l'approche *artisanale* : dans le cadre de MAAYA, et avec le soutien de l'OIF, des collaborations se poursuivent pour des études ponctuelles centrées sur le français dans l'Internet, lesquelles nourrissent les travaux de l'Observatoire de la langue française de l'OIF et l'ouvrage *La langue française dans le monde* (OIF, 2014) et (OIF, 2019), ou sur l'espagnol (Pimienta et Prado, 2016) ; cependant la production systématique d'indicateurs pour plusieurs langues n'est plus possible.

En 2012, l'Union Latine suspend ses activités ; en 2017, FUNREDES cesse ses activités ; vers la fin de la période, MAAYA rencontre des difficultés à maintenir ses activités et le flambeau du thème du plurilinguisme dans le cyberespace est repris par le secteur IFAP de l'UNESCO[22] et son dynamique antenne russe, dirigée par le charismatique Evgeny Kuzmin, qui réunit régulièrement les acteurs sans but lucratif autour du thème de la diversité linguistique et culturelle dans le cyberespace, entre 2008 et 2019[23], les premières en coordination avec MAAYA, et toujours dans le cadre de l'UNESCO, qui reste l'entité des Nations Unies qui se préoccupe formellement de ce sujet[24] (UNESCO, 2015).

Pendant cette période, deux acteurs commerciaux deviennent incontournables, parce qu'ils sont les seuls à produire des données et qu'ils parviennent à maintenir leurs activités jusqu'à aujourd'hui :

- InternetWorldStats, une entreprise de marketing en Colombie, produit, depuis 2002, des données sur l'Internet dans le monde, y inclus, depuis 2004, son classement des 10 langues les plus utilisées dans l'Internet, en termes d'utilisateurs[25].
- W3Techs, une entreprise qui produit des données autour des technologies du Web, et inclut, depuis 2011, dans sa liste à forte couleur technologique, un très apprécié classement des langues sur la Toile, qu'elle met à jour au quotidien[26] et dont elle maintient l'historique[27].

Le nombre de théories ou élaborations linguistiques bâties sur l'édifice de ces deux sources est impressionnant ; pourtant, l'expérience acquise par l'Observatoire et l'analyse des nombreux biais de la méthode W3Techs lui permettent d'estimer que les données produites exagèrent la place de l'anglais dans des proportions très importantes mais, jusqu'à 2017, il ne lui est pas possible d'opposer d'autres chiffres.

Naissance et maturation d'une alternative : depuis 2017

Le plus fidèle soutien à ce projet depuis son origine, l'OIF, permet en 2017, à travers son soutien à MAAYA, de faire éclore une nouvelle approche qui permet à l'Observatoire de produire de nouveau des indicateurs (Pimienta, 2017). Le modèle établi part de l'idée initiale de Daniel Prado qui a guidé les travaux entre 2012 et 2017: multiplier les sources quantitatives les plus

[22] https://en.unesco.org/programme/ifap
[23] http://www.ifapcom.ru/en/722/
[24] Voir https://www.unesco.org/en/communication-information/multilingualism-cyberspace
[25] https://www.internetworldstats.com/stats7.htm
[26] https://w3techs.com/technologies/overview/content_language
[27] https://w3techs.com/technologies/history_overview/content_language/ms/y

diverses sur la présence des langues dans l'Internet et, pour pallier à l'évidente rareté de ces sources, compléter par des sources par pays, plus fréquentes, en les transformant, à partir des données démolinguistiques, en sources par langue.

En 2017, l'Observatoire parvient à donner une cohérence mathématique à cette approche indirecte, à la structurer et à la généraliser pour en tirer des résultats valables pour un grand nombre de langues[28]. La méthode permet une estimation directe du pourcentage de personnes connectées par langue et une estimation indirecte du pourcentage des contenus par langue sur la Toile, ainsi que d'autres indicateurs utiles et elle marque un point d'inflexion important dans cette histoire de la mesure. Le modèle établi s'appuie sur 3 type de sources :

1. Les données démolinguistiques : nombre de locuteurs de chaque langue dans chaque pays, en différentiant les locuteurs première (L1) et seconde langue (L2).
2. Le pourcentage de personnes connectées à l'Internet par pays, donnée mise à jour chaque année par l'UIT[29], qui joue un rôle essentiel dans les calculs.
3. Un nombre aussi important que possible de sources quantitatives à propos des langues ou des pays, en relation avec des éléments concernant directement (par exemple, nombre d'abonnés aux réseaux sociaux) ou indirectement (par exemple, nombre de mobile par habitant) l'Internet, classifiées entre *trafic*, *usages, contenus*, *index*[30], *interfaces ou programmes de traduction en ligne*.

À partir de ces 3 piliers et en mettant en marche une série de calculs impliquant les données en entrée (extrapolation des données incomplètes, pondérations démolinguistiques, pondérations par les pourcentages de personnes connectées, moyennes, moyennes réduites, méthode des quartiles…) le modèle produit les indicateurs en sortie[31].

Sans surprise, les résultats contredisent les données de W3Techs qui font l'objet de biais très importants[32] et le pourcentage de l'anglais se retrouve en 2017 au niveau attendu par l'extrapolation des courbes préexistantes : 30% des sites seraient en anglais et le français est la quatrième langue en termes de contenus, derrière le chinois et l'espagnol., avec une avance confortable sur les suivantes : russe, allemand, portugais et arabe.

La méthode est complexe, sa mise en œuvre fort consommatrice de temps, étant donné la quantité de sources à trouver, évaluer puis utiliser et l'accent, depuis le début, est mis sur **l'analyse complète des biais** induits par la méthode, par ses hypothèses de travail[33] et par les nombreuses sources mises en œuvre.

---

[28] Dans un premier temps, pour limiter les biais provoqué par les hypothèses simplificatrices requises pour faire fonctionner le modèle, la limite est fixée aux 149 langues dont le nombre de locuteurs première langue est supérieur à 5 millions.

[29] Union Internationale des Télécommunications : http://itu.int

[30] Cet élément fait référence à des classements des pays dans leur progrès vers la société de l'information (gouvernement électronique, données ouvertes, etc.).

[31] Voir (Pimienta, 2017) pour les détails de la méthode et son substrat théorique.

[32] Pour comprendre les raisons de ces biais, dont le principal est l'absence de prise en compte du multilinguisme dans la Toile, voir (OIF, 2022) ou (Pimienta, 2022).

[33] Une hypothèse de travail nécessaire au modèle est que les locuteurs des différentes langues dans un pays partagent le même pourcentage de connexion à l'Internet (le taux moyen national fournit par l'UIT). Cette hypothèse interdit de comparer les langues au sein d'un même pays, elle est difficilement applicable aux langues à faible nombre de locuteurs, et tend à donner un biais positif pour les langues d'immigration dans les pays en développement (qui peuvent être moins connectées que la moyenne) et, à l'inverse, un biais négatif pour les langues européennes dans les pays en développement (qui ont tendance à être mieux connectées que la moyenne).

Entre 2017 et 2022, un effort important est donc consacré à la chasse au biais, avec un progrès notable en 2021[34], qui permet d'utiliser la meilleure source démolinguistique existante[35] et d'étendre les résultats aux 329 langues de plus d'un million de locuteurs L1, à partir desquels sont produits des résultats intéressants sur la cyber-géographie des langues (Pimienta, 2021).

	Langues africaines	Langues américaines	L'arabe comme macro-langue	Langues asiatiques	Langues européennes	Reste	TOTAL
Internautes %	**29,8%**	**56,7%**	**64,0%**	**49,3%**	**82,6%**	**47,06%**	**56,91%**
Contenus	**2,89%**	**0,22%**	**3,09%**	**44,77%**	**45,39%**	**3,64%**	**100%**
Présence virtuelle	**0,31**	**0,71**	**0,88**	**0,93**	**1,47**	**0,47**	**1**
Productivité des contenus	**0,56**	**0,69**	**0,79**	**1,00**	**1,15**	**0,57**	**1**
Locuteurs L1+L2	**9,21%**	**0,31%**	**3,53%**	**48,24%**	**30,91%**	**7,81%**	**100%**
Population connectée	**5,21%**	**0,32%**	**3,89%**	**44,63%**	**39,51%**	**6,36%**	**100%**
Langues avec L1>1M	**138**	**8**	**1**	**135**	**47**		**329**

En mars 2022, la méthode atteint sa maturité et tous les biais sont maitrisés, au prix d'une redéfinition de certains indicateurs[36], et d'une reformulation de la vision méthodologique : il s'agit d'une approximation indirecte des contenus, basée sur l'observation expérimentale que le rapport entre le pourcentage mondial de contenus et le pourcentage mondial de locuteurs connectés est toujours resté compris entre 0,5 et 1,5 (pour les langues à existence numérique complète).

Cela suggère l'existence d'une sorte de loi économique naturelle, qui lierait, pour chaque langue, *l'offre* (contenus et applications web) à *la demande* (locuteurs connectés à l'Internet). Lorsque le nombre de personnes connectées augmente, le nombre de pages web augmente rapidement, plus ou moins dans la même proportion. Il en est ainsi parce que les gouvernements, les entreprises, les institutions éducatives, etc., et une partie des nouveaux utilisateurs créent des contenus pour répondre à cette demande et/ou pour augmenter l'offre.

En ce qui concerne, les langues de France, une autre approche a donc été utilisée (Pimienta et Prado, 2014) et une base de données mise en ligne : http://baseldf.fr.

[34] Grâce au soutien du Ministère des affaires étrangères du Brésil, à travers l'Institut international de la langue portugaise (https://iilp.cplp.org), sous la coordination et l'appui linguistique de la Chaire UNESCO pour les politiques linguistiques pour le multilinguisme (https://www.unescochairlpm.org), en la personne de son responsable, Gilvan Müller de Oliveira.

[35] Ethnologue (https://www.ethnologue.com)

[36] Les statistiques linguistiques de Wikimedia, par ailleurs l'application de l'Internet ayant la plus grande diversité linguistique, d'une qualité et extension rares dans ce contexte, nourrissaient un indicateur de contenus en entrée du modèle. Un effort très important est développé dans la version de 2021 pour compenser les biais naturellement occidentaux de Wikipedia et autres éléments de la galaxie Wikimedia, en établissant des formules qui pénalisent les versions linguistiques basées sur des copies d'autres langues et faiblement mise à jour, et en intégrant toutes les encyclopédies en ligne existantes dans toutes les langues. Le résultat est décevant et une conclusion inéluctable s'impose : les encyclopédies ne sont pas un reflet fidèle de la réalité des contenus par langue. L'indicateur est supprimé et se retrouve en sortie du modèle, entrainant une redéfinition, laquelle finalement apporte plus de clarté aux concepts utilisés. Se passer des meilleures statistiques existantes sur les langues dans l'Internet, celle de Wikimedia, est une frustration, mais cette décision imposée par la réalité a permis d'obtenir un modèle où l'ensemble des biais est maitrisé et la logique mathématique, assise sur des opérations de pondération, souvent avec la répartition des personnes connectées par pays (et indirectement par langue), a fini par imposer sa cohérence au modèle.

Il est important de noter que des enquêtes et des études ont constamment rapporté que les internautes préfèrent utiliser leur langue maternelle dans l'Internet et profitent également de l'occasion pour utiliser, comme deuxième option, leur(s) deuxième(s) langue(s)[37].

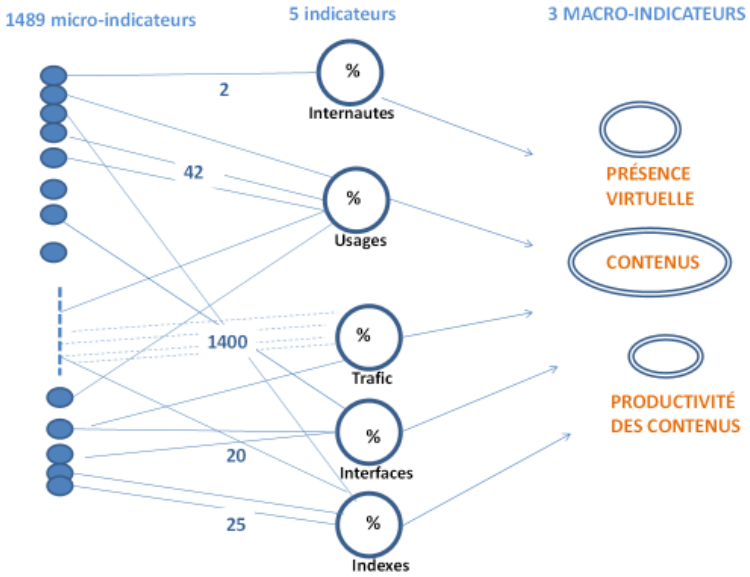
Ainsi, en fonction de chaque langue, il y a une sorte de modulation du rapport mentionné (% de contenus / % de connectés), pour le rendre plus ou moins supérieur ou inférieur à un, certaines langues ayant une meilleure productivité de contenus que d'autres, en fonction d'un ensemble de facteurs les concernant, dans chacun de leurs contextes nationaux, tels que:

- Évidemment, le nombre correspondant de locuteurs L2, puisque certaines personnes produisent, par exemple pour des raisons économiques, des contenus dans une langue différente de leur langue maternelle.

Mais aussi:

- La proportion du trafic Internet, en fonction du contexte tarifaire, culturel ou éducatif du pays.
- Le nombre d'abonnements aux réseaux sociaux et autres applications de l'Internet.
- Le support technologique numérique de la langue et sa présence dans les interfaces d'application et les programmes de traduction, qui faciliteraient ou non la production de contenus.
- Le niveau de submersion du pays où vit le locuteur en termes de manifestation de la société de l'information (commerce électronique, applications du gouvernement pour payer les impôts, etc.).

Ainsi, s'il était possible de collecter différents indicateurs sur chacune des caractéristiques évoquées, on pourrait mesurer les modulations de cet indicateur autour de la valeur un et en déduire la proportion des contenus. C'est exactement ce que réalise le modèle établi en utilisant près de 1500 sources (micro-indicateurs) pour calculer 5 indicateurs qui permettent de produire les sorties du modèle (macro-indicateurs), comme le présente le schéma suivant.



[37]Voir par exemple le rapport d'enquête de l'Union européenne https://ec.europa.eu/commission/presscorner/detail/en/IP_11_556 ou, pour le cas difficile de l'Inde, ce rapport : https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

Et cela permet d'obtenir des résultats suivants pour les 30 langues avec les plus forts pourcentages de contenus[38]:

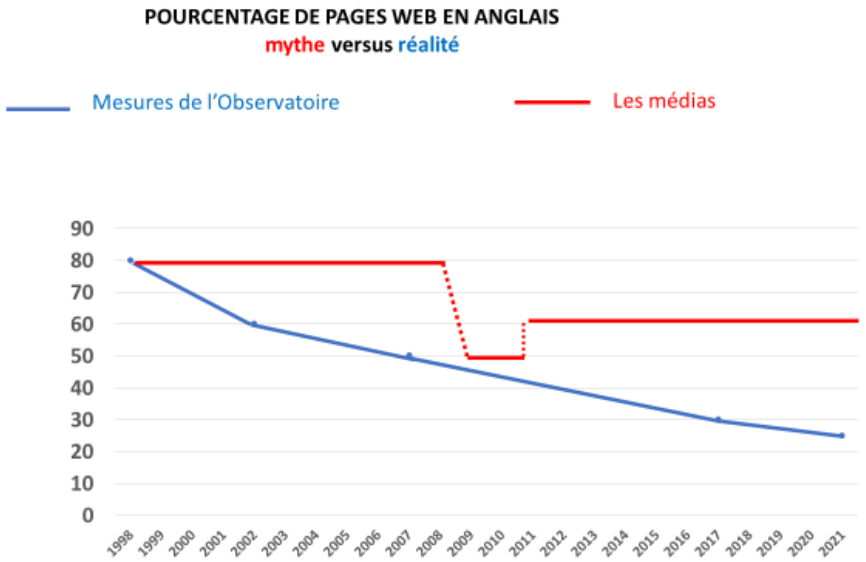
Rang Contenus			Internautes	Population Mondiale	Locuteurs connectés	Contenus	Présence Virtuelle	Productivité Contenus
L1+L2	ISO	LANGUES	L1+L2	L1+L2	L1+L2	L1+L2	L1+L2	L1+L2
1	zho	*chinois*	18,46%	14,72%	71,38%	**21,60%**	1,47	1,17
2	eng	anglais	14,83%	13,01%	64,86%	**19,60%**	1,51	1,32
3	spa	espagnol	6,79%	5,24%	73,72%	**7,85%**	1,50	1,16
4	hin	hindi	4,19%	5,80%	41,16%	**3,76%**	0,65	0,90
4	rus	russe	3,51%	2,49%	80,32%	**3,76%**	1,51	1,07
4	fra	français	2,98%	2,58%	65,80%	**3,33%**	1,29	1,12
4	por	portugais	2,99%	2,49%	68,43%	**3,13%**	1,26	1,05
4	ara	*arabe*	3,97%	3,53%	63,99%	**3,09%**	0,87	0,78
9	jpn	japonais	1,99%	1,22%	92,63%	**2,66%**	2,18	1,34
9	deu	allemand	2,04%	1,30%	89,17%	**2,37%**	1,82	1,16
11	msa	*malais*	2,36%	2,36%	56,93%	**1,96%**	0,83	0,83
12	tur	turc	1,17%	0,85%	78,05%	**1,14%**	1,35	0,98
12	ita	italien	0,87%	0,66%	75,83%	**1,00%**	1,53	1,14
12	kor	coréen	0,90%	0,79%	65,16%	**0,98%**	1,24	1,09
15	fas	*persan*	1,08%	0,81%	75,91%	**0,88%**	1,09	0,82
15	ben	bengali	1,11%	2,58%	24,55%	**0,88%**	0,34	0,79
15	vie	vietnamien	0,92%	0,74%	70,96%	**0,85%**	1,15	0,92
18	urd	ourdou	0,95%	2,22%	24,38%	**0,66%**	0,30	0,70
18	tha	thaïlandais	0,80%	0,59%	77,95%	**0,65%**	1,12	0,82
18	pol	polonais	0,60%	0,39%	87,09%	**0,63%**	1,59	1,04
18	mar	marathe	0,69%	0,96%	41,06%	**0,58%**	0,60	0,83
18	tel	télougou	0,68%	0,92%	41,69%	**0,56%**	0,60	0,82
18	tam	tamil	0,61%	0,82%	42,15%	**0,51%**	0,62	0,83
24	jav	javanais	0,62%	0,66%	53,76%	**0,44%**	0,66	0,70
24	nld	néerlandais	0,38%	0,24%	91,14%	**0,41%**	1,73	1,08
26	guj	gujarati	0,44%	0,60%	41,47%	**0,36%**	0,61	0,83
26	ukr	ukrainien	0,40%	0,32%	71,02%	**0,35%**	1,09	0,88
26	kan	kannada	0,41%	0,57%	41,11%	**0,33%**	0,59	0,82
29	ron	roumain	0,32%	0,23%	79,57%	**0,30%**	1,29	0,93
29	aze	*azerbaïdjanais*	0,33%	0,23%	81,54%	**0,28%**	1,21	0,85
		RESTE	**22,60%**	**30,10%**		**15,13%**		
		TOTAL	**100,00%**	**100,00%**		**100,00%**		

L'intervalle de confiance dans les résultats est de l'ordre de +-20% et c'est pour cela que les langues marquées de la même couleur doivent être considérées comme ayant un pourcentage de contenus identique. Ainsi, en 2022, le français est la quatrième langue de l'Internet en termes de contenus, en compagnie de l'hindi, du russe, du portugais et de l'arabe, chacune de ces langues représentant entre 3 et 4% de l'ensemble des contenus alors que l'anglais et le chinois représentent chacun entre 16 et 24% des contenus.

[38] Les pourcentages sont exprimés par rapport à la population mondiale L1+L2. Selon la source Ethnologue, la population mondiale (L1) est de 7 231 699 136 de personnes tandis que la population mondiale de locuteurs de langue L1 ou L2 est de 10 361 716 756, c'est-à-dire que plus de 43% de la population mondiale serait au moins bilingue.

L'ensemble des résultats est laissé en accès libre (CC-BY-SA 4.0) sur la page http://funredes.org/lc2022.

Reste le difficile chemin à parcourir pour faire perdre les (mauvaises) habitudes acquises depuis plus de 10 ans d'utiliser, sans précaution, des sources sérieusement biaisées qui laissent penser à tort que l'anglais est resté stable entre 2011 et 2022 au-dessus de 60% des contenus de la Toile… alors qu'en réalité il a été rejoint par le chinois et que sa place est désormais autour de 20% ! L'histoire se répète et les deux courbes suivantes résument cette histoire de mésinformation.



Le futur des langues dans l'Internet

À long terme, le moment viendra quand les locuteurs de toutes les langues seront des utilisateurs avec des taux de connectivité supérieur à 90%, comme c'est le cas aujourd'hui du norvégien, danois, suédois, catalan, japonais et finlandais, pour citer les champions. Mais, pour les contenus, il restera probablement des écarts notables entre les représentations respectives des langues dans le monde et dans le cyberespace, certaines en sur-représentation (présence virtuelle supérieure à 1) tandis que d'autres seront moins avantagées. Un des indicateurs produits, le *degré de cyber-mondialisation* d'une langue :

$$CGI(L) = (L1+L2) / L1(L) \times S(L) \times C(L)$$

Où:

L1+L2/L1(L) est le rapport du multilinguisme de la langue L
S(L) est le pourcentage de pays du monde qui détiennent des locuteurs de la langue L
C(L) est le % de locuteurs de la langue L connectés à l'Internet.

renseigne sur les **atouts stratégiques** d'une langue dans le cyberespace. Cet indicateur montre que l'avantage de l'anglais va continuer et que sa présence virtuelle continuera d'être parmi les plus hautes. Le français se place ensuite avec un écart notable sur les suivants (allemand, russe

et espagnol). La démographie reste le facteur essentielle, associée à la capacité des langues à attirer des apprenants seconde langue. Les langues africaines, qui restent aujourd'hui les moins présentes dans le cyberespace, pourront prendre leur revanche, si la fracture numérique y est résolue, vers 2050, quand la population de l'Afrique pourrait avoir doublé. Cette perspective pourrait aussi bénéficier aux langues européennes les plus présentes sur ce continent : l'anglais et le français en premier lieu.

Quant au futur de l'observation des langues, on ne peut qu'espérer que ce domaine attire de nouvelles volontés et connaisse une plus grande diversité d'approches dont tout le monde pourrait bénéficier. Même si le secteur se professionnalise enfin et que l'utilisation des outils algorithmique de la reconnaissance des langues soit accompagnée du sérieux méthodologique requis[39], les approches artisanales comme celle de l'Observatoire garderont leur espace.

BIBLIOGRAPHIE

- Pimienta D. (2001). « Quel espace dans l'internet en dehors de la langue anglaise et de la culture « made in USA » ? », *Les Cahiers du numérique*, V2. https://www.cairn.info/revue-les-cahiers-du-numerique-2001-3-page-205.htm

- Pimienta D. (2005). "At the Boundaries of Ethics and Cultures: Virtual Communities as an Open-Ended Process Carrying the Will for Social Change (the "MISTICA" experience)" in *Localizing the Internet. Ethical Issues in Intercultural Perspective*., Capurro, R. & al. (Eds.). Schriftenreihe des ICIE Bd. 4, München: Fink Verlag, 2005 https://funredes.org/mistica/english/cyberlibrary/thematic/icie/

- Mikami Y. et al. (2005). "The language observatory project (LOP)", *Proceedings of the 14th international conference on World Wide Web, WWW 2005*, Chiba, Japan https://www.researchgate.net/publication/221022705_The_language_observatory_project_LOP

- UNESCO. (2006). « Mesurer la diversité linguistique dans l'Internet », CI.2005/WS/06 https://unesdoc.unesco.org/ark:/48223/pf0000142186_fre

- Pimienta D. (2007). « Fracture numérique, fracture sociale, fracture paradigmatique », dans *Fractures, mutations, fragmentations : de la diversité des cultures numériques*, sous la direction de Kiyindou A. Paris : Hermès Science Publications : Lavoisier. ISBN 978-2-7462-2220-5 https://funredes.org/mistica/francais/cyberotheque/thematique/fracture_paradigmatique.pdf

- Pimienta D., Prado D., Blanco A. (2009). « Douze années de mesure de la diversité linguistique sur l'Internet: bilan et perspectives » UNESCO CI-2009/WS/1. https://unesdoc.unesco.org/ark:/48223/pf0000187016_fre

- Maaya. (2012). « Net.lang. Réussir le cyberespace multilingue », Coordonné par Vannini L., Le Crosnier H., C&F Éditions, ISBN 978-2-915825-08-4, 2012 - https://cfeditions.com/NetlangFR/ (téléchargeable en français, anglais ou russe).

- OIF. (2014). « Le français dans l'Internet », *Rapport 2014 "La langue française dans le monde"*, pp. 501-541, Nathan, 2014 -http://www.francophonie.org/Rapports-Publications.html

- Pimienta D., Prado D. (2014). « Étude sur la place des langues de France dans l'Internet », *Langue & Recherche*, Délégation générale à la langue française et aux langues de France, 2014.

[39] À ce propos, des suggestions sont faites à W3Techs dans (Pimienta, 2021) pour permettre à son algorithme de surmonter les biais existants, à un coût modéré, et offrir l'observation crédible que tout le monde, l'Observatoire en premier, souhaite.

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

- UNESCO. (2015). « Une décennie de promotion du multilinguisme dans le cyberespace », CI-2015/WS/5 https://unesdoc.unesco.org/ark:/48223/pf0000232743_fre

- Pimienta D., Prado D. (2016). "Medición de la presencia de la lengua española en la Internet: métodos y resultados", en *Revista Española de Documentación Científica 39(3)*, e141. ISSN-L:0210-0614. - http://dx.doi.org/10.3989/redc.2016.3.1328

- Pimienta D., Prado D. (2016) « Un milliard de Latins… dans l'Internet ? », in *Hermès, La Revue, 2016/2 (n° 75) Langues romanes : un milliard de locuteurs*. http://www.cairn.info/revue-hermes-la-revue-2016-2.htm

- Prado D. (2016) « Les langues romanes minoritaires et l'Internet », *in Hermès, La Revue, 2016/2 (n° 75) Langues romanes : un milliard de locuteurs.* http://www.cairn.info/revue-hermes-la-revue-2016-2.htm

- Pimienta D., Prado D. (2016) "Ten Years of MAAYA, the World Network for Linguistic Diversity: Time for Balance and Perspectives", in Proc. of *Multilingualism in Cyberspace*, IFAP/UNESCO – P184. http://www.ifapcom.ru/files/2016/UGRA_ENGL_BLOK_WEB.pdf

- Pimienta D. (2017). « Une approche alternative pour produire des indicateurs des langues dans l'Internet. », *Observatoire de la diversité linguistique et Culturelle dans l'Internet*, https://funredes.org/lc2017/Langue%20Internet%20Alternative.docx

- OIF. (2019). « La présence de la langue française dans le cyberespace (synthèse)", *Rapport 2019 "La langue française dans le monde"*, pp. 337-341, OIF, Gallimard. https://www.francophonie.org/sites/default/files/2021-04/LFDM-20Edition-2019-La-langue-française-dans-le-monde.pdf
Étude complète accessible à http://observatoire.francophonie.org/2018/Place-francais-sur-Internet-D-Pimienta.pdf

- Jardon Q. (2019). « Alexandria », Gallimard, ISBN 978-2-0728-5287-9

- Pimienta D., Rodríguez Leal LG., (2020), - "Rock the Internet Blues; Une vision critique de l'évolution de l'Internet depuis la perspective de la société civile", https://funredes.org/RockInternetBlues

- Pimienta D. (2021) « Internet et diversité linguistique : la cyber-géographies des langues avec le plus grand nombre de locuteurs», traduction et actualisation de l'article paru dans *LinguaPax Review 2021, Language Technologies and Language Diversity*, en anglais, catalan et espagnol.
https://funredes.org/lc2022/CyberGFR.pdf

- OIF. (2022). « La présence de la langue française dans le cyberespace », dans *"La langue française dans le monde 2019-2022"*, Gallimard/OIF- ISBN : 9782072976865
https://gallimard.fr/Catalogue/GALLIMARD/Hors-serie-Connaissance/La-langue-francaise-dans-le-monde
Synthèse accessible en ligne : https://www.francophonie.org/sites/default/files/2022-03/Synthèse_La_langue_française_dans_le_monde_2022.pdf

- Pimienta D. (2022). « Ressource : Indicateurs de présence des langues sur Internet », traduction de l'article en anglais présenté à SIGUL2022, Marseille. http://funredes.org/lc2022/Res.Ind.lang.Internet.fr.pdf