Explorer de manière fiable la présence des langues dans l'Internet

L'Internet est une ressource précieuse pour les linguistes car il offre un vaste espace facilement accessible où ils peuvent observer l'évolution des langues au fil du temps. Daniel Pimienta, directeur de l'Observatoire de la diversité linguistique et culturelle dans l'Internet (OBDILCI), a développé une méthode pour mesurer la présence des langues sur le Web, laquelle a été considérablement améliorée ces dernières années. Cette méthode a permis le développement d'une base de données complète qui pourrait soutenir la recherche linguistique, les politiques publiques liées aux langues et les stratégies de commerce électronique.

L'avènement de l'Internet et son utilisation généralisée ont ouvert de nouvelles voies intéressantes pour l'étude des langues. Mesurer la présence en ligne de différentes langues offre des indications précieuses sur leur utilisation et leur développement futurs.

Des estimations fiables de l'usage des langues dans l'Internet pourraient à terme guider l'élaboration de politiques publiques visant à influencer leur présence dans le cyberespace. L'Observatoire de la diversité linguistique et culturelle dans l'Internet, institut de recherche fondé en 1996, est spécialisé dans le développement de méthodes efficaces pour mesurer la présence et l'évolution des langues dans l'Internet.

Même si les algorithmes de reconnaissance des langues – des outils informatiques capables d'identifier les langues écrites – semblent être des moyens adéquats pour déterminer la prévalence des langues en ligne, le Web est désormais devenu si vaste qu'il est très peu pratique d'appliquer ces outils à l'ensemble du contenu en ligne. Certaines études ont ainsi utilisé ces algorithmes pour analyser des sous-ensembles de la Toile, mais cette approche expérimentale s'est révélée inefficace, conduisant à des résultats biaisés et souvent peu fiables.

Jusqu'à récemment, la source de statistiques sur l'usage des langues en ligne la plus consultée (W3Techs) s'appuyait sur des algorithmes pour analyser les sites web classés comme les plus visités. Bien que ces statistiques offrent des informations intéressantes, elles pourraient ne pas refléter avec précision la présence des langues en ligne en raison du manque de prise en compte de la nature souvent multilingue des sites Web, lacune qui entraine d'importants biais.

En 2017, l'Observatoire de la diversité linguistique et culturelle dans l'Internet a conçu une nouvelle approche qui pourrait permettre de mieux suivre les progrès et la prévalence des langues en ligne. Grâce à cette approche, Daniel Pimienta et ses collègues ont pu identifier des indicateurs significatifs concernant la présence de 343 langues dans l'Internet.

INDICATEURS DE PRESENCE LINGUISTIQUE EN LIGNE

Dès 1998, les chercheurs de l'observatoire avaient introduit une approche pour étudier la présence de sept langues en ligne, qui s'appuyait sur les données collectées par les moteurs de recherche tels qu'AltaVista et Google. Cependant, en 2007, ils ont commencé à se rendre compte que les résultats des moteurs de recherche devenaient peu fiables et ont donc cherché des méthodes alternatives.

La nouvelle approche introduite en 2017 s'attaque aux forts biais associés aux efforts antérieurs dans ce domaine de recherche. Initialement, les chercheurs ont appliqué cette approche à 138 langues, soit celles parlées par plus de 5 millions de locuteurs natifs, mais ils ont récemment pu l'étendre à 343 langues, celles parlées nativement par plus d'un million de personnes.

En utilisant la méthode proposée, Pimienta et ses collègues ont compilé un ensemble d'indicateurs de la présence de ces 343 langues en ligne. Ces indicateurs ont été divisés en trois grandes catégories, à savoir les indicateurs intermédiaires, macro et avancés.

Les indicateurs intermédiaires, tous exprimés en pourcentage, incluent les internautes (c'est-à-dire les locuteurs d'une langue donnée connectés à l'Internet), l'utilisation de services ou d'applications spécifiques de l'Internet et le trafic rapporté sur des sites ou applications. Ils comprennent également une approximation du niveau de support numérique des langues et ce que l'on appelle des indices, qui sont des évaluations de pays basées sur des paramètres de la société de l'information, transformés par pondération en évaluations des langues.

Le deuxième ensemble d'indicateurs, appelé macro-indicateurs ou résultats du modèle, comprend les locuteurs connectés (c'est-à-dire le pourcentage de locuteurs de langue première et seconde dans le monde qui sont connectés à l'Internet), le pourcentage de contenu Web dans chaque langue, la productivité des contenus (le rapport entre contenus Web et internautes) et la présence virtuelle (le rapport entre contenus Web et locuteurs).

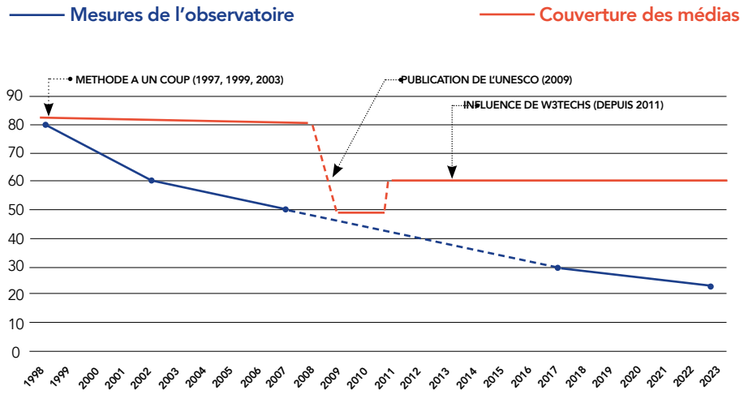
Enfin, les indicateurs les plus avancés identifiés par les chercheurs incluent la cyber-géographie des familles de langues ou, en d'autres termes, la répartition des langues sur le Web en groupes géographiques (c'est-à-dire européennes, asiatiques, arabes, américaines et africaines) et un indicateur de cyber-mondialisation. La valeur de ce dernier indicateur, calculée à partir de certains autres indicateurs, exprime essentiellement les « avantages stratégiques » d'une langue donnée dans l'Internet.

UNE NOUVELLE METHODE POUR ETUDIER LA PRESENCE DES LANGUES

La nouvelle approche de l'observatoire consiste à estimer indirectement la quantité relative de contenus Web par langue. Pour le faire, elle prend en compte des facteurs cruciaux qui sont souvent ignorés lors de la description de la présence d'une langue dans l'Internet, mais qui devraient être pris en compte pour éviter les erreurs ou les biais.

Premièrement, l'équipe considère l'existence probable d'une « loi économique » relative à l'Internet, qui lie l'offre (c'est-à-dire le contenu Web disponible dans une langue) à la demande (c'est-à-dire le nombre de locuteurs de cette langue connectés au réseau). Les études antérieures avaient montré que plus les locuteurs d'une langue donnée sont connectés à l'Internet, plus il existe de pages Web dans cette langue.

POSITION DE L'OBSERVATOIRE AU SUJET DE L'EVOLUTION DES POURCENTAGES DE PAGES WEB EN ANGLAIS
mythe vs réalité



Le pourcentage de pages Web en langue anglaise a considérablement diminué au fil du temps, selon les mesures de l'observatoire.

Entre 2011 et 2022, W3Techs a présenté des statistiques suggérant que l'anglais est présent en ligne de manière stable avec plus de 50 % des contenus, mais l'analyse de l'observatoire suggère que ce n'est pas le cas.

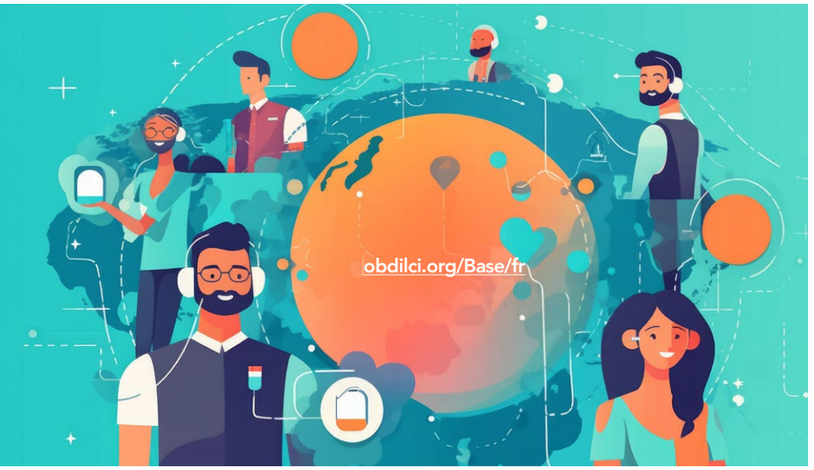
En outre, des recherches antérieures suggèrent que les internautes préfèrent souvent communiquer dans leur langue maternelle lorsque le contenu recherché est disponible dans cette langue, mais qu'ils sont bien disposés à utiliser leur ou leurs langues secondes dans le cas contraire. Dans certains cas, les internautes peuvent également créer du contenu dans leur seconde langue pour des raisons économiques et recourir à des services de traduction pour ce faire.

La présence d'une langue en ligne est également liée à la quantité de trafic Internet vers différents sites, au nombre d'abonnements aux réseaux sociaux et aux progrès des différents pays en termes de services en ligne pour les citoyens. Les indicateurs de présence dans l'Internet créés par les chercheurs prennent collectivement en compte tous ces facteurs, dressant ainsi un tableau plus détaillé de la quantité et de la manière dont les différentes langues existent sur la Toile.

UNE BASE DE DONNEES FACILEMENT ACCESSIBLE

En utilisant leur approche, Pimienta et ses collègues ont entrepris de calculer des indicateurs de présence en ligne pour les langues parlées nativement par plus d'un million de personnes dans le monde. Cela leur a permis de constituer une base de données complète résumant la présence de ces langues en ligne, que l'observatoire prévoit de mettre à jour chaque année.

Les valeurs qu'ils ont obtenues sont très intéressantes, car elles ne correspondent souvent pas aux valeurs produites par d'autres efforts de mesure linguistique. Par exemple, entre 2011 et 2022, le site



La base de données de l'observatoire est librement disponible aux linguistes explorant la présence de différentes langues en ligne.

Les personnes de langue japonaise semblent être les plus virtuellement présentes, générant proportionnellement plus de contenus que les personnes connectées d'autres langues.

W3Techs a présenté des statistiques suggérant que l'anglais représente de manière stable et de loin la majorité des contenus en ligne (plus de 50 %), alors que l'analyse de l'observatoire indique que ce n'est pas le cas (environ 20 % aujourd'hui).

Pimienta et ses collègues ont découvert que les langues proposant le plus de contenu Web sont l'anglais et le chinois. On estime que chacune de ces langues représente entre 16 et 26 % de tout le contenu Web, suivies par l'espagnol (entre 7 et 9 %), l'arabe, l'hindi, le russe, le français et le portugais (3 à 4 %), le japonais, l'allemand, et le malais (2 à 3 %), suivis par bengali (1,5 à 2,3 %) et turc, vietnamien, italien, coréen et persan (0,8 à 1,2 %). Dans l'ensemble, les résultats de l'équipe suggèrent que les contenus de la Toile sont de plus en plus multilingues, la prévalence de l'anglais diminuant progressivement de 80 % en 1998 à 20 % aujourd'hui. Il reste clair que cela ne signifie pas que la quantité de contenus en anglais a diminué au fil du temps, mais plutôt que les contenus en ligne dans de nombreuses autres langues ont augmenté, réduisant ainsi le pourcentage des contenus en anglais.

Collectivement, les 215 langues restantes représentent 18 à 26 % des contenus sur la Toile. En examinant plusieurs indicateurs de présence linguistique en ligne, la base de données des chercheurs offre plusieurs autres informations qui pourraient éclairer la recherche linguistique, les politiques publiques et les stratégies de commerce électronique.

Par exemple, Pimienta et ses collègues ont découvert que les personnes parlant norvégien sont les plus connectées à l'Internet, avec un nombre impressionnant de 98,8 % de locuteurs connectés, suivis par les locuteurs du danois (98,7 %), du suisse allemand (94,1 %), du catalan (94,5 %), et du finnois (92,8 %). Par ailleurs, les personnes de langue japonaise apparaissent comme les plus virtuellement présents, générant, proportionnellement à leur nombre, plus de contenus que les locuteurs d'autres langues.

Les indicateurs liés à la cyber-géographie des langues ont également fourni des informations précieuses. Par exemple, ils ont montré que même si le nombre de langues parlées par plus d'un million de personnes en Afrique est supérieur à celui de chacune des autres régions géographiques, les locuteurs de langues africaines sont moins connectés à l'Internet ; cependant la tendance récente est enfin à la croissance pour les langues africaines.

Enfin, l'indicateur de cyber-mondialisation met l'accent sur l'avantage stratégique des langues anglaises et françaises. En d'autres termes, parler une de ces deux langues ouvrira davantage d'opportunités futures aux internautes.

EXPLORER DE MANIERE FIABLE LA PREVALENCE DES LANGUES DANS L'INTERNET

Les travaux récents de l'Observatoire de la diversité linguistique et culturelle dans l'Internet ont permis de produire de nouvelles données fiables expriment la présence de différentes langues en ligne. Alors que l'utilisation de l'Internet continue de croître à l'échelle mondiale, ces données pourraient éclairer la façon dont les langues utilisées par les internautes progressent sur le Web.

De nombreuses statistiques existantes sur la présence des langues dans l'Internet se sont révélées trompeuses, ne parvenant pas à représenter de manière adéquate la mesure dans laquelle les langues existent sur la Toile. Pourtant, certaines de ces statistiques ont été largement utilisées dans les études linguistiques et diffusées dans les médias, entraînant davantage de confusion et de désinformation.

La base de données en ligne de l'observatoire est accessible au public et pourrait être utilisée par les linguistes qui explorent la présence de différentes langues dans l'Internet. Parallèlement, elle pourrait également servir de référence aux décideurs de politiques publiques, aux stratèges du commerce électronique et aux fournisseurs de services dans l'Internet, en les aidant à mieux comprendre, depuis leur propre perspective, l'utilisation des différentes langues dans l'Internet.

En coulisse

Dr Daniel Pimienta

E: pimienta@funredes.org **W:** funredes.org/pimienta **W:** obdilci.org
www.linkedin.com/in/pimienta

Détail

Biographie
Daniel Pimienta a étudié les mathématiques appliquées et l'informatique à Nice (France). D'abord architecte de systèmes de télécommunications chez IBM, il a ensuite créé FUNREDES, une ONG de recherche-action pionnière en TICpD et a été secrétaire général du Réseau Mondial pour la Diversité Linguistique (MAAYA). Il est actuellement responsable de OBDILCI, produisant des indicateurs sur la présence des langues dans l'Internet.

Financement
Organisation Internationale de la Francophonie, gouvernement brésilien, Instituto Internacional da lingua portuguesa, Direction du français et des langues de France du ministère de la Culture de France

Collaborateurs
Chaire UNESCO sur les politiques linguistiques pour le multilinguisme

Références

- Pimienta, D, (2022) Ressource : Indicateurs sur la présence des langues dans l'Internet. Actes du SIGUL2022 @LREC2022, 89-91. aclanthology.org/2022.sigul-1.11 - Version française
- Pimienta, D, Blanco, A, Müller de Oliveira, G, (2023) La méthode derrière la production sans précédent d'indicateurs de la présence des langues dans l'Internet. Frontiers in Research Metrics and Analytics, 8. doi.org/10.3389/frma.2023.1149347 - Version française
- Pimienta, D, Prado, D, Blanco, A, (2009) Douze années de mesure de la diversité linguistique dans l'Internet : bilan et perspectives. Publications de l'UNESCO pour le Sommet mondial sur la société de l'information. CI-2009/WS/1. https://unesdoc.unesco.org/ark:/48223/pf0000187016_fre

Objectifs de recherche

Daniel Pimienta étudie la présence des langues dans l'Internet.

Réponse personnelle

Qu'est-ce qui vous a inspiré pour mener cette recherche?

« En 1995, lors du Sommet de la Francophonie à Cotonou (Bénin), le président français Chirac a présenté l'Internet comme un espace 100 % anglophone. À cette époque, j'étais une sorte d'évangéliste de l'Internet dans l'Association Réseaux et Développement (FUNREDES) et je pensais que cette affirmation était fausse et ne reposait pas sur des données prouvées. En réaction, j'ai décidé de lancer un effort de recherche visant à mesurer la prévalence des langues dans l'Internet. Ce projet a mûri en 1998 avec l'aide de l'Union Latine, et produit une série de résultats jusqu'en 2007. Il a traversé un chemin difficile après 2007 et jusqu'en 2017, lorsqu'il a pu se relancer dans une approche nouvelle et prometteuse permettant d'étendre le nombre des langues traitées. »

Comment la base de données d'indicateurs linguistiques que vous avez compilée pourrait-elle éclairer les politiques publiques visant à renforcer la présence des langues en ligne?

« Aujourd'hui, les stratégies visant à renforcer les langues doivent se concentrer principalement sur le cyberespace en raison de son puissant impact mondial. Quelle que soit la politique que vous développez, dans quelque domaine que ce soit, vous avez besoin d'indicateurs signifiants, fiables et pérennes pour définir votre stratégie et pouvoir évaluer fréquemment les impacts de vos actions, afin de les adapter en conséquence. Les indicateurs de ces politiques linguistiques dans le cyberespace se caractérisent depuis trop longtemps par des données biaisées, surestimant largement la réalité de la présence de l'anglais et sous-estimant la part notable du multilinguisme dans l'Internet, démotivant ainsi les efforts de production de contenu local. Cette désinformation doit cesser et nous sommes heureux de pouvoir contribuer de cette manière et renforcer toutes les actions en faveur du multilinguisme sur le réseau. »