Organisation
des Nations Unies
pour l'éducation,
la science et la culture

Douze années de mesure de la diversité linguistique sur l'Internet : bilan et perspectives

par Daniel Pimienta, Daniel Prado et Álvaro Blanco

Publications de l'UNESCO pour le
Sommet mondial sur la societé de l'information

Secteur de la Communication et de l'Information

Douze années de mesure de la diversité linguistique sur l'Internet : bilan et perspectives

Organisation
des Nations Unies
pour l'éducation,
la science et la culture

UNIÓ LLATINA
UNIÓN LATINA
UNION LATINE
UNIONE LATINA
UNIÃO LATINA
UNIUNEA LATINĂ

Douze années de mesure de la diversité linguistique sur l'Internet : bilan et perspectives

par
Daniel Pimienta, Daniel Prado et Álvaro Blanco

Responsabilité

Les idées et opinions exprimées dans cette publication sont celles des auteurs et ne reflètent pas nécessairement les vues de l'UNESCO. Les appellations employées dans cette publication et la présentation des données qui y figurent n'impliquent de la part de l'UNESCO aucune prise de position quant au statut juridique des pays, territoires, villes ou zones, ou de leurs autorités, ni quant à leurs frontières ou limites.

Note bibliographique recommandée :

Douze années de mesure de la diversité linguistique sur l'Internet : bilan et perspectives

Edité par la Division de la société de l'information, Secteur de la communication et de l'information, UNESCO, 76 p., 21 cm.

Coéditeur : Anneflore Lemoulinier, Union latine, responsable de la traduction française

I – Douze années de mesure de la diversité linguistique sur l'Internet : bilan et perspectives

II – Daniel Pimienta, Daniel Prado et Álvaro Blanco

III – Langues; Internet; toile; diversité linguistique; politiques linguistiques; indicateur

Publié en 2009

Par l'Organisation des Nations Unies pour l'éducation, la science et la culture

7, place de Fontenoy, 75352 Paris 07 SP, Paris, France

CL-2009/WS/1

Table des matières

Préface

L'Internet constitue un atout majeur pour améliorer la libre circulation de l'information et des idées dans le monde. Attachée à la construction des sociétés du savoir, l'UNESCO s'engage activement dans les actions visant à améliorer la diversité culturelle et linguistique sur l'Internet et à permettre l'accès à l'information pour tous.

La célébration en 2008 de l'Année internationale des langues par l'UNESCO a attiré l'attention des décideurs politiques du monde entier et d'une plus grande partie de l'opinion publique sur l'intérêt stratégique des langues et des politiques linguistiques pour le développement. Avec l'introduction des noms de domaine internationalisés (IDN) dans les adresses Internet, la question de l'accès à l'Internet dans des écritures et des langues locales est au centre des récents débats sur la gouvernance de l'Internet.

Aujourd'hui, la communauté internationale vise de plus en plus à permettre au plus grand nombre de personnes d'accéder et d'utiliser l'Internet dans leurs propres écritures et langues. La relation entre les langues sur l'Internet et la diversité linguistique dans un pays indique que les pays ont un rôle important à jouer dans l'adoption d'une politique linguistique appropriée pour l'Internet. Une telle politique linguistique nécessite un volet spécifique pour traiter la diversité linguistique dans le monde virtuel, ainsi que des chiffres pertinents basés sur des indicateurs fiables afin d'évaluer la situation.

À cette fin, l'UNESCO a demandé à des experts de Funredes et de l'Union latine d'actualiser l'étude intitulée « Mesurer la diversité linguistique sur l'Internet » publiée dans le cadre du Sommet mondial sur la société de l'information en 2005. L'UNESCO s'est engagée pour une approche de statistiques et de mesures qui va au-delà d'une vision strictement technocentrée afin de prendre en compte l'importance des contenus et d'un contexte susceptible de favoriser et d'encourager la diversité, tout en reconnaissant les

limites dans la mesure des cultures et des contenus représentés sur l'Internet.

Cette étude présente diverses méthodes employées ces douze dernières années pour mesurer la diversité linguistique sur l'Internet. Les résultats mettent en évidence certains des mythes entourant les chiffres existants, notamment en ce qui concerne la présence dominante de l'anglais sur la Toile.

J'espère qu'elle sera largement utilisée, notamment par les décideurs politiques et les personnes responsables de l'introduction, l'application et l'évaluation des politiques linguistiques, des stratégies, des programmes et des projets sur l'Internet. Elle devrait aussi être utile aux chercheurs universitaires intéressés par la mesure de la diversité linguistique sur l'Internet. Je leur recommande cette étude. J'espère également que cette publication, en accord avec la Recommandation de l'UNESCO sur la promotion et l'usage du multilinguisme et l'accès universel au cyberespace adoptée en 2003, facilitera l'élaboration de politiques linguistiques favorables à la diversité culturelle et linguistique sur l'Internet.

Abdul Waheed Khan
Sous-Directeur général pour la
communication et l'information
UNESCO

Remerciements

Les auteurs de cette publication sont les suivants : Daniel Pimienta (Directeur de FUNREDES, membre du Comité exécutif du Réseau MAAYA, chercheur à l'Université Antilles-Guyane en Martinique), Daniel Prado (Directeur de la Direction terminologie et industries de la langue de l'Union Latine) et Alvaro Blanco (Directeur de FUNREDES en Espagne). Plusieurs autres personnes de l'Union latine ou de FUNREDES ont participé à ce projet : Marcel Sztrum a géré l'équipe de linguistes pour la sélection de l'échantillon de mots ; Benoit Lamey a écrit le programme permettant de limiter l'intervention humaine dans le processus de mesure ; Roger Price a fourni des éléments essentiels à la partie statistique de la recherche.

RÉSUMÉ

Funredes et l'Union latine ont conçu une méthode originale, qui utilise les moteurs de recherche et un échantillon de mots concepts ayant la meilleure équivalence dans différentes langues latines, ainsi qu'en anglais et en allemand, afin de mesurer leur présence proportionnelle dans le cyberespace, et plus particulièrement sur la Toile. La méthode, appliquée de 1996 à 2008, a permis d'élaborer des indicateurs intéressants pour mesurer la diversité linguistique. En outre, une étude simple a été réalisée afin d'apprécier la place de la culture associée aux langues latines et à l'anglais.

Ce document décrit la méthodologie et ses résultats, ainsi que ses avantages et ses limites, et offre un panorama des méthodes alternatives et de leurs résultats. En conclusion, le document présente l'examen des perspectives d'un domaine considéré, à l'époque, comme manquant de rigueur scientifique. Cela a conduit à une certaine désinformation sur la présence dominante de l'anglais sur la Toile, et attire maintenant peu à peu l'attention des organisations internationales et du monde universitaire.

Toutes les données détaillées pertinentes sur la méthodologie et ses résultats sont librement disponibles sur la Toile. Plusieurs publications présentant les résultats de la recherche de manière chronologique ont été réalisées, mais elles ne permettent pas aux lecteurs d'obtenir un aperçu complet du projet. Cet article tente de résoudre ce problème en fournissant une description complète du projet et des résultats dans un document unique. Il fait référence aux précédents rapports, afin que les chercheurs et décideurs politiques qui le souhaiteraient puissent en savoir plus sur la méthode ou les résultats. Le but de cet article est également de sensibiliser le grand public au thème de la diversité linguistique dans le cyberespace.

1. INTRODUCTION

1.1 NAISSANCE D'UN PROJET

En décembre 1995, lors du Sommet de la Francophonie à Cotonou, il a été publiquement affirmé que la présence de l'anglais sur la récente Toile d'araignée mondiale dépassait les 90 %. Ce chiffre a été le moteur de déclarations critiquant l'Internet pour son parti pris linguistique inhérent, déclenchant une réaction de Funredes pour la défense de l'Internet. L'équipe de Funredes a d'abord essayé, en vain, de localiser la source de ce chiffre qui connote une dominance de l'anglais sur l'Internet, puis a cherché un moyen de produire par elle-même quelques premiers chiffres bruts. L'idée est alors venue naturellement d'exploiter la puissance des moteurs de recherche (un monde dominé à l'époque par Altavista), ce qui a permis d'établir une première estimation pour obtenir une idée très approximative de la répartition de l'anglais, du français et de l'espagnol sur la Toile[1]. Par ailleurs, une autre estimation approximative de la représentation des cultures associées à ces langues a été réalisée en mesurant la présence en ligne des noms de personnes célèbres dans différents domaines et en les comparant[2].

Ces deux processus — et en particulier celui de la mesure linguistique — manquaient manifestement de valeur scientifique à ce stade, mais ont permis de :

1) faire une estimation très approximative de la présence de l'anglais sur la Toile autour de 80 % ;
2) évaluer les obstacles linguistiques à surmonter pour obtenir une méthode fiable de mesure de la diversité linguistique sur la Toile basée sur le nombre d'occurrences d'un mot donné figurant dans les pages de la Toile indexées par les moteurs de recherche ;

1 http://funredes.org/lc2005/francais/L1.html/

2 http://funredes.org/lc2005/francais/C1.html/

3) montrer que la nature mondiale de la Toile permettait une représentation équitable de la culture française (du moins quand elle était clairement dissociée des enjeux commerciaux), mais pas de la culture espagnole (la situation a depuis beaucoup évolué).

En outre, cette étude sur la diversité linguistique en ligne a probablement engagé un processus de récolte d'informations intéressantes pour la documentation et les archives sur l'Internet. Elle a également contribué à l'analyse des comportements historiques des moteurs de recherche.

En tout état de cause, ces résultats ont préparé le terrain pour ce qui est devenu la seule série de mesures répétées et cohérentes de la présence d'un sous-ensemble de langues sur la Toile et dans d'autres espaces en ligne[3]. De 1996 à 2008, la méthodologie et les résultats obtenus ont été présentés de manière transparente. Au cours de ces 12 années d'étude, la présence de l'anglais n'a pas cessé d'être médiatisée à 80 %, malgré la formidable vitesse d'évolution de la démographie de l'Internet qui montre des chiffres passant de 80 % à 40 % en termes d'internautes anglophones[4]. Cette désinformation est devenue un défi majeur à surmonter pour l'équipe de recherche.

En entreprenant cette étude, il n'était pas tant question d'une lutte pour la défense du français, mais plutôt, en cohérence avec le rôle de Funredes en tant qu'ONG impliquée dans le domaine des TIC pour le développement[5], d'un plaidoyer pour la création de contenus locaux. À cette époque, l'idée largement répandue était celle d'une dominance massive, envahissante et stable de l'anglais sur la Toile, dans le contexte d'un monde virtuel qui était supposé refléter la diversité linguistique et culturelle du monde « réel ». Cette idée semblait conspirer contre l'évident besoin

3 Dans ce document, les mots et expressions tels qu'« en ligne », « monde virtuel », « Internet » ou cyberespace sont synonymes ; le mot « Toile » est utilisé pour faire référence à un sous-ensemble de l'Internet (comme les groupes de discussion, les blogues ou Wikipédia). http://www.oclc.org/

4 Voir la référence de GlobalStat ci-dessous.

5 Les ONG travaillant dans ce domaine tentent d'utiliser les TIC dans le but de permettre aux personnes, communautés et pays de changer positivement leur condition socioéconomique. La diversité linguistique sur l'Internet est un enjeu important pour lutter contre la fracture numérique.

d'une politique cohérente de création de sites Internet en « langues maternelles » et de développement de contenus locaux.

En 2005, l'UNESCO publiait un rapport (voir Mesurer la diversité linguistique sur Internet dans les références) qui cherchait à fournir une vue d'ensemble des différentes perspectives sur la présence de la diversité linguistique sur l'Internet. Le rapport inclut également des chiffres relatifs à la présence en ligne de l'anglais. À titre d'exemple, dans son article, Paollillo soutient le chiffre de 80 % — principale référence des études de l'OCLC – comme étant une bonne estimation de la présence de l'anglais sur l'Internet. Pimienta, qui a coordonné une série d'articles écrits par des chercheurs à travers le monde[6], plaide quant à lui pour une présence de l'anglais située aux alentours de 50 %. La publication de ce rapport a probablement marqué un tournant historique, conduisant à une plus grande ouverture sur la thématique de la diversité linguistique dans l'Internet et déclenchant l'intérêt de plusieurs chercheurs dans ce vaste, mais négligé, domaine d'étude.

1.2 OBJECTIFS DE L'ARTICLE

Comme il l'a été mentionné, le but de cet article est d'abord de fournir un panorama complet et une analyse de l'ensemble des études menées par Funredes et l'Union latine de 1996 à 2008. Bien que la méthode et les résultats aient toujours été publiés avec une totale transparence tout au long du projet, ils étaient présentés comme des séries d'évènements décrits de manière séquentielle, forçant le lecteur à suivre la chronologie pour comprendre le travail. Cela ne donnait pas une vision claire, cohérente et pédagogique du travail.

Cet article rectifie ce défaut. Il s'agit de la première tentative de synthèse et d'analyse des résultats produits par les différentes études. Y sont décrites les valeurs et les limites de la méthodologie et des résultats. D'autres projets de recherche sont analysés et leurs limites exposées. Cet article expliquera également pourquoi

6 Certains bons articles n'ayant pas été sélectionnés dans le rapport final, faute de place, peuvent être lus ici : http://funredes.org/lc/francais/unesco/index.htm.

et comment la méthodologie utilisée n'est plus viable, en raison de la récente évolution des moteurs de recherche. Cela a conduit les chercheurs à considérer le besoin d'outils plus ambitieux visant à mieux refléter la réalité de l'intégralité de la Toile. Ainsi, le second objectif de l'article est d'examiner l'état d'une nouvelle discipline qui fait partie de la ***cybermétrie***.

Cet article permettra de fournir aux décideurs du monde entier une vision de l'évolution et des tendances des langues sur l'Internet. Il permettra également d'offrir un matériel complet aux chercheurs ou aux décideurs politiques intéressés par le domaine de la diversité linguistique sur l'Internet, à une époque où les questions en jeu reçoivent enfin l'attention qu'elles méritent[7]. Plus particulièrement, cet article veut en finir définitivement avec la désinformation sur l'ampleur de la dominance de l'anglais sur la Toile. En conclusion, les auteurs expliqueront leurs plans pour continuer à mesurer la diversité linguistique et culturelle sur l'Internet – plans pour lesquels ils sont publiquement à la recherche de collaborations et de soutiens.

Bien que certaines parties de cet article nécessitent quelques connaissances techniques (en linguistique, en statistique ou sur l'Internet) pour être bien comprises, il offre également une vue d'ensemble qui peut intéresser les non-spécialistes. Le dernier objectif est par conséquent de sensibiliser les citoyens numériques[8], afin que la Société civile comprenne mieux l'importance de la diversité linguistique dans le cyberespace.

7 Comme en témoigne le Forum sur la gouvernance d'Internet qui a eu lieu à Rio de Janeiro (2007), où une table ronde coordonnée par le ministre brésilien de la Culture, Gilberto Gil et à laquelle le Président du Réseau MAAYA, Adama Samassékou, a participé, a été consacré à ce thème. Voir : http://www.intgovforum.org et plus particulièrement : http://www.intgovforum.org/Rio_Meeting/IGF2-Diversity-13NOV07.txt.

8 « netizen » en anglais.

2. CONTEXTE DU PROJET

Généralement, les spécialistes divergent lorsqu'il s'agit de donner des chiffres démographiques sur les langues. Les définitions et les frontières sont complexes et atteindre un consensus n'est pas facile. Les informations ci-dessous s'appuient principalement sur les chiffres de David Crystal (voir ***Language and the Internet*** dans les références). En ce qui concerne plus spécifiquement les langues latines, l'Union latine sera la référence pour cet article.

Le nombre de langues créées par les êtres humains est estimé à environ 40 000, parmi lesquelles entre 6 000 et 9 000 sont encore en usage (chiffres variables selon les statistiques). Certaines sources estiment qu'une langue disparaît tous les deux mois en moyenne.

Dans ce contexte, la préservation de la diversité linguistique devient un enjeu de taille. La question de savoir si l'Internet constitue une aubaine ou une menace pour la diversité linguistique se pose alors naturellement.

La réponse n'est pas simple et dépend de nombreux paramètres sur la langue en question : est-ce une langue locale, nationale, internationale ou une ***lingua franca*** ? Provient-elle d'un pays développé ? Existe-t-il une politique linguistique ? Existe-t-il une politique linguistique pensée pour le cyberespace ?

Pour résumer la situation, une classification simple des langues est proposée ci-après. Elle n'a pas la prétention d'offrir des réponses définitives, mais au contraire a la volonté d'ouvrir la réflexion sur un sujet sur lequel la société civile n'a pas été autant sensibilisée qu'elle l'a été pour la diversité biologique, bien qu'il y ait des corrélations évidentes entre ces deux points[9]. Étant donné la situation actuelle de notre planète, l'absence de politique de protection contre la réduction de la diversité biologique pourrait nuire à l'avenir de la société. La même question peut être posée pour la diversité culturelle, et justifie notre attention. La classification explique les

9 La corrélation entre les régions du monde ayant une diversité biologique faible/élevée et le nombre de langues parlées est forte.

implications de l'exposition et de la présence sur l'Internet pour chaque type de langue.

Tabeau 1 : Catégorisation des langues pour les besoins de politiques pour le cyberespace

TYPE	L'INTERNET CONSTITUE-T-IL UNE AUBAINE ?
Principales langues parlées[1]	L'Internet pourrait augmenter la présence en ligne de ces langues, notamment lors d'une période de transition où la répartition des internautes par langue n'est pas équilibrée comme conséquence de la fracture numérique. Note : la thèse ici est que l'anglais a fini de traverser cette période transitoire il y a quelques années.
Langues officielles couvrant plus d'un pays développé (comme l'italien ou le néerlandais)	Il y a une occasion à saisir dans le monde virtuel. Le statut « international » de ces langues peut favoriser la confiance entre les locuteurs pour créer facilement des relations transfrontalières.
Langues officielles parlées dans un seul pays développé (comme le norvégien, le grec, le danois ou le japonais)	Une vigoureuse politique linguistique virtuelle est nécessaire pour obtenir une présence dans le monde virtuel comparable ou supérieure à celle dans le monde réel. Malgré une certaine longévité en ce qui concerne la place de ces langues dans le monde, leurs locuteurs peuvent ressentir un obstacle pour les relations internationales.
Langues locales de pays développés (sarde, gallois, galicien, frison, etc.)	Ces langues sont menacées par une forte pression venant à la fois de l'anglais et de leurs langues nationales respectives. Le diagnostic est incertain sans une politique linguistique virtuelle. Chaque cas varie et dépend de divers critères, même si le cas du catalan est à suivre comme un exemple de réussite tant au niveau virtuel que non virtuel.
Lingua franca de locuteurs de certains pays en développement (haoussa, quechua, pulaar, swahili, etc.)	Un avenir positif devrait être possible là où la fracture numérique est vraiment surmontée et que des politiques linguistiques virtuelles sont définies.
Langues de pays en développement couvrant plus d'un pays, mais seulement utilisées par des locuteurs natifs (aymara, guarani, créoles, etc.)	Un avenir positif devrait théoriquement être possible là où la fracture numérique est vraiment surmontée. Cependant, il y a actuellement une corrélation entre le manque d'accès à l'informatique et le fait d'appartenir à une communauté autochtone qui ne montre pour le moment aucun signe de changement. Le cas du Paraguay où le guarani se dote d'instruments conformément à son statut d'officialité doit être suivi avec intérêt.

TYPE	L'INTERNET CONSTITUE-T-IL UNE AUBAINE ?
Langues officielles d'un seul État en développement (slovène, albanais, etc.)	Elles sont soumises à une forte pression venant à la fois de l'anglais et des langues voisines puissantes, ce qui pourrait engendrer des perspectives négatives en l'absence de politiques virtuelles.
Langues locales de pays en développement (chabacano, maya, mapuche, etc.)	Si la langue dispose des outils linguistiques nécessaires (et notamment, en premier lieu, d'un système d'écriture et d'une grammaire stables et normalisés), une politique linguistique basée sur la création de contenus locaux pourrait se révéler utile. Mais il y a peu d'exemples aujourd'hui de ce type de situation favorable.
Langues menacées de disparition (comme l'aïnou)	L'Internet pourrait, au pire, se révéler comme un formidable outil de conservation du patrimoine écrit et oral, au mieux, comme un accélérateur de politiques de réhabilitation de la langue.
Langues très sérieusement menacées de disparition (comme le yagan)	L'Internet permettrait au moins la conservation du patrimoine de cette langue, si des mesures de numérisation sont prises suffisamment tôt.

Le principal message qui émane de ce tableau est le besoin d'établir des politiques linguistiques, tant dans le monde réel (et le catalan est une bonne référence à étudier), que dans le monde virtuel (où l'analyse des actions de l'Organisation internationale de la Francophonie pourrait se révéler intéressante, les études montrant des résultats positifs obtenus à partir d'une politique volontariste de production de contenus en français).

Afin de créer une politique linguistique significative, la première étape consiste, d'une part, à obtenir des chiffres quantifiant la situation, et d'autre part, à évaluer et à suivre les effets de la politique, en se basant sur des indicateurs fiables. De nos jours, une politique linguistique complète doit inclure un volet spécifique pour le monde virtuel, impliquant différentes dynamiques, logiques et règles, par comparaison avec le monde réel. De là, se pose naturellement la nécessité de disposer de données fiables sur la présence des langues sur l'Internet.

Dans ce domaine, la situation a été paradoxale et frustrante. L'histoire de l'Internet est étroitement liée à l'histoire de la recherche et au monde universitaire. Cependant, avec l'avènement de la Toile et la croissance de la partie commerciale de l'Internet, le

secteur universitaire a partiellement laissé la création de données démographiques sur l'Internet au secteur privé et peut-être plus dangereusement au secteur mercatique. Cela a eu pour conséquence la création de données privées et non publiquement disponibles. Cela a également conduit au fréquent manque de transparence des méthodologies de recherche. Les chiffres publiés n'ont pas toujours été produits de manière scientifique. En outre, le manque d'objectivité de ces données conduit à penser qu'elles ont pu être guidées par des intérêts commerciaux ou autres qui peuvent influencer les résultats obtenus ou chercher à révéler des informations particulières.

Dans un domaine dans lequel la démographie a évolué à une vitesse sans précédent dans l'histoire de l'Humanité, cela a permis la création de mythes, comme celui selon lequel l'anglais aurait une présence sur la Toile incroyablement dominante et stable aux alentours de 80 %. Ce mythe a, en règle générale, bénéficié de l'absence d'une réponse critique de la part du monde universitaire.

Cette époque semble aujourd'hui révolue, comme en témoignent les faits suivants :

1) des séries cohérentes d'actions organisées et suivies par l'UNESCO[10] ;

2) l'émergence de Maaya[11], à la suite du Sommet mondial sur la Société de l'information ;

3) le ***Language Observatory Project***, lancé par un réseau d'universités[12].

Il y a par conséquent un intérêt croissant des décideurs et des universitaires à reprendre le contrôle sur ce domaine et à contribuer de façon significative à l'émergence de politiques linguistiques pour le monde virtuel, basées sur l'utilisation d'indicateurs fiables.

10 http://www.unesco.org/cgi-bin/webworld/portal_observatory/cgi/page.cgi?d=1&g=Cultural_Diversity_and_Multilingualism/index.shtml

11 Réseau mondial pour la diversité linguistique - http://maaya.org

12 Dirigé par le professeur Mikami, de la Nagaoka University of Technology, le Language Observatory (http://www.language-observatory.org/) est constitué d'un consortium international de partenaires.

Dans ce contexte historique, ce projet peut être considéré comme une tentative unique, pionnière et engagée de recherche-action de la part de la Société civile, de résistance à l'influence de la désinformation à propos de l'Internet. En effet, la diversité linguistique sur l'Internet est stratégique, puisque directement liée à des enjeux fondamentaux comme la fracture numérique et, indirectement, à la question générale de la gouvernance de l'Internet[13].

13 Le Forum de la gouvernance sur l'Internet s'intéresse actuellement de près à la question des noms de domaines internationaux alors qu'il ne s'agit que de la partie visible de l'iceberg de la diversité linguistique sur l'Internet. Voir Comment assurer la présence d'une langue dans le cyberespace ? dans les références.

3. HISTORIQUE DU PROJET

L'histoire de ce projet de recherche est documentée dans deux sites Internet. Pour les personnes intéressées par le suivi de l'évolution du projet de 1996 à 2005, le site Internet[14] original historique du projet fournit toutes les informations nécessaires classées par ordre chronologique des campagnes de mesures. Le second site[15] présente les résultats après 2005.

Le tableau suivant résume les étapes du projet à travers ces périodes.

Tableau 2 : Campagnes de mesures et étapes du projet

DATES	ÉTUDE LINGUISTIQUE	PRÉSENCE DE L'ANGLAIS SUR LA TOILE / Moteur de recherche	ÉTUDE CULTURELLE
06/96	L1 : résultats très approximatifs - anglais, français et espagnol - la Toile	~80 % Altavista	C1 : première série de résultats sur la culture
03/97	L2 : répétition de L1	~80 % Altavista	
03/98	L3 : répétition avec un échantillon plus important. - méthode du « complément de l'ensemble vide » - analyse de la méthode Alis - décision de consolider la méthode en collaboration avec l'Union latine	~80 % Altavista	

14 http://funredes.org/lc2005/francais/index.html

15 http://funredes.org/lc/

DATES	ÉTUDE LINGUISTIQUE	PRÉSENCE DE L'ANGLAIS SUR LA TOILE / Moteur de recherche	ÉTUDE CULTURELLE
09/98	L4 : première étude faite avec une méthodologie fiable, en partenariat avec l'Union latine et avec le soutien financier de l'Agence de la Francophonie - ajout de l'italien, du portugais et du roumain - ajout de Usenet - début de la création d'indicateurs linguistiques	75 % Hotbot Dejanews	C2 : deuxième série de résultats sur la culture avec plusieurs améliorations de l'échantillon et du modèle de classement - progression notable de la présence de personnalités françaises et espagnoles
08/00	L5 : deuxième étude faite avec une méthodologie fiable, en partenariat avec l'Union latine - création d'un programme visant à générer automatiquement l'ensemble du processus, de l'interrogation des moteurs de recherche aux résultats statistiques - ajout de l'allemand	60 % Google + Fast	
01/01 06/01 08/01 10/01 02/02 02/03 02/04 05/04 03/05	L64 : - campagnes de mesures sans modification de la méthode - nouveaux indicateurs par pays et par langue pour le français (2002) - nouveaux indicateurs par pays et par langue pour le portugais (2003) - nouveaux indicateurs par pays et par langue pour l'anglais (2004)	de 55 % à 47 % Fast Yahoo Google	C3, sept. 2001
10/05	- campagnes de mesures sans modification de la méthode	45 % Google	Nouveaux résultats sur la culture
03/06	- campagnes de mesures sans modification de la méthode	45 % Google	
12/07	- ajout du catalan	45 % Yahoo	
05/08	- campagnes de mesures sans modification de la méthode	Yahoo	Nouveaux résultats sur la culture

Septembre 1998 marque le début de l'utilisation de méthodes et de résultats fiables dans l'étude. Avec la mise en place d'un programme en PHP[16] pour l'automatisation de l'ensemble du processus, y compris l'entretien d'une base de données des résultats, septembre 2000 marque le début d'une gestion de projet professionnelle et systématique.

16 PHP est un langage de programmation utilisé pour produire des pages dynamiques.

4. MÉTHODOLOGIE

La méthodologie établie repose sur la combinaison des éléments suivants :

- l'utilisation du nombre d'occurrences de chaque mot concept par langue, tel que mesuré par les moteurs de recherche[17] ;
- un échantillon de mots concepts dans une sélection de langues données ;
- un ensemble d'outils statistiques.

Les moteurs de recherche ont été sélectionnés sur la base d'un minimum de critères spécifiques établis pour l'étude, tels que :

- offrir des chiffres fiables pour le comptage ;
- permettre un traitement équitable des diacritiques[18] ;
- couvrir la plus grande partie possible de l'espace de l'Internet analysé.

Les échantillons de mots concepts destinés à être comptés par les moteurs de recherche sélectionnés ont été choisis pour leur congruence conceptuelle parmi les langues de l'étude, à savoir :

- une parfaite équivalence syntaxique ;
- la meilleure équivalence sémantique ;
- la plus grande neutralité culturelle possible.

La compilation du décompte de pages sur la Toile pour chaque mot concept (la somme des résultats des différents mots associés

17 Le nombre de pages de la Toile contenant un mot ou une expression donnés tel qu'il est calculé par les moteurs de recherche.

18 Les diacritiques présents dans la plupart des langues utilisant l'alphabet latin, mais absents de l'anglais, permettent souvent, comme on le sait, d'identifier des sens différents (caña en espagnol a un sens différent de cana, côte en français est différent de cote ou encore de côté). Au début, l'Internet, avait tendance à ne pas permettre la codification des diacritiques, puisqu'il utilisait une norme de codage de caractères basé sur l'alphabet anglais (ASCII - American Standard Code for Information Interchange) de seulement 7 bits et donc n'autorisant pas plus de 128 caractères.

à chaque concept[19]) est considérée comme une variable aléatoire dont la distribution est traitée statistiquement (moyenne, variance, et intervalle de confiance, à l'aide de la Loi de Fisher).

L'objectif est de produire une estimation du poids relatif de la langue en question par rapport à l'anglais tel qu'il est mesuré dans l'index[20] du moteur de recherche sélectionné. Dans certaines circonstances (la taille de l'index étant le facteur clé), il est raisonnable d'extrapoler le résultat comme une représentation juste de la répartition des langues dans la Toile (visible)[21].

Afin d'obtenir un pourcentage absolu pour les langues étudiées, tel que mesuré dans l'espace sélectionné de l'Internet, le poids absolu de l'anglais doit d'abord être déterminé pour servir de point de référence et de comparaison. Malheureusement, la méthode de recherche utilisée ne permet pas cela. Ce poids doit donc être déterminé à l'aide d'une étape manuelle supplémentaire qui consiste à combiner des informations extraites de différentes sources avec une estimation du poids relatif du reste des langues non incluses dans l'étude. La répétition périodique de l'ensemble de ces procédures permet aux chercheurs d'obtenir une vision de l'évolution de la présence des langues au fil du temps.

Bien que la Toile ait constitué le principal objet de l'étude, d'autres parties du cyberespace ont également été étudiées, telles que les groupes de discussion ou, plus récemment, les blogues ou Wikipédia.

La méthode de recherche implique également la considération des éléments suivants :

- les critères précis pour la validation et l'utilisation des moteurs de recherche ;
- les critères linguistiques utilisés pour construire l'échantillon de termes (et les corrections nécessaires dans certains cas) ;

19 Avec les corrections correspondantes, dans certains cas.

20 Par « index du moteur de recherche », il faut entendre ici l'ensemble des pages de la Toile indexées par le moteur.

21 La Toile invisible (ou Toile profonde) est la somme des pages dynamiques produites par des bases de données ou d'autres mécanismes de programmation des pages dynamiques. Certains auteurs estiment qu'elle pourrait être entre 100 et 500 fois plus importante que la Toile visible (voir White Paper: The Deep Web dans les références).

- les outils statistiques utilisés pour atteindre les résultats finaux ;
- la conception d'indicateurs à partir de ces résultats ;
- la nature, la signification et les limites des résultats obtenus.

4.1 MÉTHODOLOGIE LINGUISTIQUE

La cadre ci-après contient la liste des concepts (en français) qui ont été utilisés pour la comparaison linguistique.

Tableau 3 : Liste des mots concepts

ambiguïté, causalité, fromage, compatibilité, contiguïté, dangereux, décembre, densité, disparité, divisibilité, élasticité, électricité, février, féminité, fertilité, fidélité, fraternité, vendredi, hétérosexualité, homosexualité, cheval, humidité, maladie, immortalité, immunité, incompatibilité, infaillibilité, infériorité, infidélité, instabilité, inviolabilité, irrégularité, irresponsabilité, juin, genou, couteau, poumon, masculinité, lundi, octobre, parité, probabilité, productivité, puberté, responsabilité, sexualité, singularité, supériorité, jeudi, aujourd'hui, vérité, mardi, uniformité, universalité, université, mercredi, jaune

Deux exemples de l'ensemble des mots associés à chaque concept sont présentés ci-après. Ces exemples sont donnés dans les langues utilisées dans le projet de recherche. Par convention, les mots apparaissant en ***italique*** sont ceux qui ne sont pas correctement orthographiés, mais qui seront tout de même mesurés (comme c'est le cas des mots français après l'élimination des signes diacritiques) ; les mots apparaissant en MAJUSCULES sont ceux qui souffrent de problèmes d'homographie interlinguistique ou autres (et qui nécessitent un traitement particulier).

Tableau 4 : Exemples de mots concepts

anglais	espagnol		italien	portugais	roumain	allemand	catalan
fidelity fidelities faithfulness faithful- nesses	fidelidad FIDEL- IDADES	fidélité fidelite fidél- ités fidel- ites	fedeltà fedelta	fidelidade FIDELIDADES	fidelitate fidelitatea fidelității fidelitatii fidelități fidelitati fidelitățile fidelitatile fidelităților fidelitatilor	TREUE TREUEN	fidelitat FIDELITATS
Monday Mondays	lunes	lundi lundis	lunedì lunedi	segunda-feira segundas- feiras	luni lunea	montag MONTAGES montags MONTAGE MONTA- GEN	Dilluns

Le tableau complet, qui contient un peu plus de 1 700 mots, peut être consulté en ligne à l'adresse : http://funredes.org/lc/francais/historia/listapa.htm.

Comment cette liste finale de mots concepts a-t-elle été obtenue ? D'abord, en établissant un ensemble de critères pour obtenir les meilleurs mots concepts avec des équivalents multilingues. À partir de là, un nombre considérable de mots concepts potentiels a été testé et filtré[22]. Cependant, il était impossible d'obtenir des résultats parfaits et des traitements postérieurs ont été nécessaires pour éviter les biais statistiques (comme le partage du nombre de citations du terme ***fidelidades*** entre l'espagnol et le portugais ou encore le décompte de ***montage*** et ***montages*** qui a été déduit en séparant le nombre d'occurrences en français de celles en allemand).

Certains problèmes reconnus ont été considérés comme acceptables, car n'ayant qu'un impact marginal sur le processus statistique (par exemple le fait que ***Treue*** et ***TREUEN*** sont

22 En réalité, plusieurs centaines de mots ont été étudiés avant de parvenir au tableau final et le processus est le résultat d'un travail d'équipe intense qui a duré plusieurs mois et qui a été marqué par une forte collaboration entre l'Union latine et Funredes et à l'intérieur de chaque institution.

également des formes de l'adjectif ***fidèle***, ce qui donne plus de portée sémantique aux mots allemands).

La liste ci-après décrit l'ensemble des critères utilisés pour créer l'échantillon de mots concepts.

Critère 1 : Neutralité culturelle

Définition : propriété d'un mot en relation avec sa fréquence d'apparition dans une langue donnée en fonction de sa valeur culturelle et de son sens.

Exemples : ***vin, parfum, gastronomie*** ne sont pas culturellement neutres en français.

Règle : écarter les termes non neutres culturellement.

Critère 2 : Homographie interlinguistique

Définition : l'orthographe d'un terme dans une langue est identique à celle d'un terme dans une autre langue, que le sens soit le même ou non. L'homographie est forte lorsque l'orthographe est exactement la même (diacritiques compris). Elle est faible, lorsque la seule différence entre les termes est due aux diacritiques.

Exemples : ***casa*** a le même sens en espagnol et en portugais ; ***red*** signifie ***réseau*** ou ***filet*** en espagnol et ***rouge*** en anglais ; ***gestión*** en espagnol et ***gestion*** en français ont des sens similaires. L'homographie peut aussi se limiter à une partie d'un mot composé comme dans l'usage anglais de ***mardi-gras.***

Règles :

- éviter les concepts qui incluent des mots de moins de quatre lettres pour réduire les probabilités d'homographies avec les langues n'entrant pas dans le cadre de l'étude ;
- quand un mot avec homographie est présent dans l'échantillon (ce qui est courant en espagnol et en portugais), il faut diviser le nombre de pages proportionnellement à la présence relative de chaque langue (les mots marqués en majuscules dans le tableau sont sujets à cette règle) ;

- chaque fois que c'est possible, corriger les résultats numériques calculés en supprimant le décompte du mot ou de l'expression (par exemple, le décompte de ***mardi*** en français est obtenu après soustraction du décompte de ***mardi-gras*** en anglais).

Critère 3 : Homographie par emprunt

Définition : quand un mot dans une langue donnée est également utilisé tel quel dans d'autres langues.

Exemples : les mots anglais ***business, sandwich*** ou ***software*** sont utilisés dans de nombreuses autres langues sous la forme anglaise. Le mot français ***déjà vu*** se dit tel quel en anglais.

Règle : rejeter les concepts incluant de tels mots.

Critère 4 : Homographie avec abréviations

Définition : quand un mot dans une langue donnée a la même orthographe qu'une abréviation fréquemment utilisée dans d'autres langues.

Exemples : le nombre ***sept*** se confond avec l'abréviation de ***September*** en anglais.

Règle : rejeter les concepts incluant de tels mots. La règle d'éviter les mots de moins de quatre lettres réduit la probabilité de telles occurrences.

Critère 5 : Homographie avec des noms propres fréquents

Définition : quand un mot d'une langue donnée a la même orthographe qu'un nom propre fréquent dans une autre langue.

Exemples : ***Julio*** signifie ***juillet*** en espagnol, mais c'est également un prénom très répandu. ***Windows*** (***fenêtre*** en français) est aussi le nom d'une marque de logiciel fréquemment citée sur l'Internet.

Règle : rejeter les concepts incluant de tels mots.

Critère 6 : Homographie avec possibles erreurs de frappe ou d'orthographe

Définition : quand l'orthographe d'un mot avec une erreur commune d'orthographe correspond à un mot existant.

Exemples : ***embassador*** en anglais, s'il est écrit avec un seul « s » correspond au même concept en roumain.

Règle : rejeter les concepts incluant de tels mots seulement si la langue cible est l'anglais (seul cas où les statistiques pourraient être affectées de manière significative).

Critère 7 : Polysémie ou champs sémantiques différents

Définition : quand un même mot a différents sens exprimés par différents mots dans d'autres langues.

Exemples : ***prix*** en français signifie à la fois ***price*** et ***prime*** en anglais, ***premio*** et ***precio*** en espagnol.

Règle : rejeter les concepts incluant de tels mots ou veiller à inclure tous les sens correspondants dans toutes les langues.

Critère 8 : Caractéristiques grammaticales discordantes (verbe, nom)

Définition : quand le même mot possède différentes fonctions grammaticales (c'est-à-dire quand il correspond à la fois à un verbe et à un nom), qui sont exprimées par différents mots dans d'autres langues.

Exemples : ***Love*** en anglais (qui est à la fois un nom et un verbe) correspond à la fois au nom ***amour*** et au verbe ***aimer*** dans différentes conjugaisons (***aime, aimes, aimons, aimez, aiment***…) en français.

Règle : éviter de tels mots. C'est la raison pour laquelle il n'y a pas un seul verbe dans l'échantillon.

Critère 9 : Caractéristiques grammaticales discordantes (adjectif, nom)

Définition : quand le même mot possède différentes fonctions grammaticales (adjectifs ou noms), qui sont exprimées par différentes formes de mots dans d'autres langues.

Exemples : ***yellow*** correspond en espagnol à ***amarillo, amarilla, amarillos, amarillas.*** Le couple ***instability/instabilities*** correspond aux variantes roumaines suivantes : ***instabilitate, instabilitatea, instabilităţii, instabilităţi, instabilităţile, instabilităţilor.***

Règle : ces mots sont acceptables pourvu que l'on veille à multiplier les variantes en genre, nombre et cas dans les autres langues, quand le besoin d'équivalence l'exige. Cela explique pourquoi le nombre de mots dans l'échantillon peut varier en fonction de la langue.

Critère 10 : Synonymie

Définition : quand un même concept dans une même langue est exprimé par différents mots selon le pays dans lequel il est utilisé.

Exemples : selon le pays hispanophone dans lequel il est utilisé, le mot français ***essence*** se dit ***nafta, gasolina*** ou ***carburante*** en espagnol.

Règle : ces mots sont acceptables pourvu que l'on veille à multiplier les variantes synonymiques nationales ou régionales.

Critère 11 : Variation orthographique

Définition : quand un même mot a diverses orthographes selon le pays dans lequel il est utilisé.

Exemples : ***Theater*** en anglais américain et ***theatre*** en anglais britannique. ***Electricidade*** au Portugal et ***eletricidade*** au Brésil.

Règle : ces mots sont acceptables pourvu que l'on veille à multiplier les différentes variantes orthographiques.

L'application de ces onze critères a permis d'établir la liste actuelle de mots concepts qui demeure inchangée depuis le début du processus. Des langues supplémentaires ont été mesurées en rajoutant au tableau une nouvelle colonne, les mots correspondant à la langue en question.

Traitements postérieurs

Comme le montrent les deux exemples cités plus haut, le filtrage complet des mots concepts pour éliminer tous les problèmes linguistiques n'a pas été possible (la probabilité d'homographies interlinguistiques étant assez élevée). Certains traitements étaient parfois nécessaires pour réduire les biais statistiques indésirables. La décision de corriger ou non les chiffres obtenus était basée sur des considérations statistiques pragmatiques. Une correction était généralement effectuée à l'aide d'une simple règle de proportionnalité linguistique, suivant une estimation de l'apparente prépondérance de la langue étudiée sur l'Internet. La situation la plus fréquente était avec le suffixe pluriel *-idades* qui est assez courant en espagnol et en portugais. Il a donc été décidé de diviser le nombre de telles occurrences entre les langues en question, en fonction de leur présence proportionnée en ligne. Les autres situations ayant nécessité un traitement particulier sont toutes décrites dans les rapports précédents[23].

Tous ces traitements ont été intégrés dans le programme informatique et sont effectués sans intervention humaine. Toutefois, dans chaque campagne, un examen manuel complet des résultats des moteurs de recherche a été systématiquement mené (à l'aide du programme informatique qui met en évidence les anomalies statistiques). Cela a permis de détecter les possibles situations conflictuelles pouvant survenir en raison d'abréviations homographiques ou de noms propres devenant fortement présents sur l'Internet.

23 http://funredes.org/lc2005/english/L4index.html

Tentative de modification de l'échantillon

À un moment donné, une expérience intensive a été entreprise en utilisant des expressions[24] composées de quelques mots, au lieu de mots concepts uniques en vue de surmonter la probabilité d'homographies. Le résultat a été extrêmement frustrant compte tenu du temps considérable investi et de l'irrégularité des résultats obtenus. Dans de nombreux cas, le décompte de l'occurrence des expressions anglaises était significativement bien en deçà de celui des autres langues étudiées. Une analyse du phénomène a conduit à attribuer ces comportements chaotiques à la perte de linéarité de la fonction mathématique inhérente. Les exemples suivants illustrent la situation :

En anglais	Nombre d'occurrences	En français	Nombre d'occurrences
« networks »	3 834 260	« réseaux »	326 250
« development »	21 258 510	« développement »	909 790
« networks » and development »	201	« réseaux et développement »	61
« note bank » = 150 000 « billet de banque » = 128 000 «billete de banco » = 18 700			

Pour la suite, l'équipe a donc décidé de conserver l'échantillon. L'échantillon avait démontré sa validité en fournissant des résultats cohérents au fil du temps et démontré des changements de tendances compréhensibles conformes à l'évolution des moteurs de recherche.

4.2 MÉTHODOLOGIE DES MOTEURS DE RECHERCHE

L'ensemble du processus de cette étude a été caractérisé par une adaptation permanente aux comportements des moteurs de recherche. L'activité principale de chaque campagne de mesure a donc consisté à vérifier si les moteurs de recherche pouvaient répondre aux objectifs de la méthodologie, et dans de trop

24 Un nouvel échantillon de plus de 200 expressions a été créé.

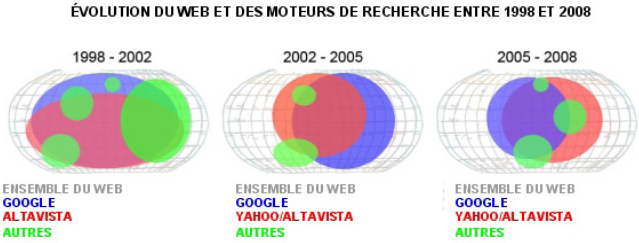
nombreux cas, à comprendre la raison pour laquelle des résultats non valides sont apparus. L'établissement de la méthodologie linguistique a demandé un investissement initial important dans ce projet. Malheureusement, la nature changeante des index des moteurs de recherche et de leurs fonctionnalités a rendu cette partie du travail encore plus longue et imprévisible !

La méthode a impliqué de vérifier attentivement le comportement de chaque moteur disponible et indépendant, en termes de diacritiques et de décompte des pages. Dans un contexte de manque de transparence (et de fréquents changements) de la part des fournisseurs de moteurs de recherche, les tests ont été multipliés maintes fois afin de comprendre l'irrégularité de certains résultats produits. À plusieurs reprises, les résultats obtenus ont conduit à une perte élevée de confiance en la méthode, mais à force d'efforts intenses et de collaboration, les chercheurs ont finalement découvert la raison de ces étranges données produites. Cela permit de démontrer que les résultats fournis par les moteurs de recherche n'étaient pas complètement fiables[25]. La partie correspondante de la méthodologie du projet a donc été réparée en fonction des mécanismes découverts et en utilisant un programme informatique.

Lors des premières séries de mesures entreprises avec la méthodologie entièrement redéfinie, les résultats se sont révélés très satisfaisants. Tous les moteurs de recherche alors utilisés produisirent des résultats non seulement statistiquement valides, mais également très proches les uns des autres. Cela ne fit que renforcer la confiance en la méthode employée. Mais la situation changea avec le temps. Le schéma suivant explique pourquoi la situation est aujourd'hui semée d'embûches.

25 C'est précisément la situation rencontrée avec Google depuis 2005 qui a obligé l'équipe, après des mois d'essais, à abandonner ce moteur et à utiliser Yahoo (Altavista) à la place. Le nombre d'occurrences fourni par Google pour un mot donné avec n'importe quel paramètre de langue et de domaine étant, contre toute logique, beaucoup plus faible que la somme du nombre d'occurrences par langue ou par domaine…

Figure 1 : Couverture des moteurs de recherche au fil du temps



Trois phénomènes expliquent aujourd'hui l'absurdité de continuer à utiliser la méthodologie originale basée sur l'utilisation des moteurs de recherche :

1) Les index des moteurs de recherche représentent à l'heure actuelle moins de 30 % de tout l'univers du cyberespace (contre plus de 80 % dans le passé) et sont de plus en plus souvent réceptifs à des critères commerciaux cachés qui augmentent considérablement le biais linguistique en faveur d'une prédominance de la langue anglaise[26].
2) Les moteurs de recherche sont de plus en plus « intelligents » (par exemple, ils recherchent des concepts dans différentes langues), ce qui perturbe la méthode du décompte de pages.
3) L'augmentation de la publicité ancrée dans les pages de la Toile apporte de nouveaux biais aux résultats de la recherche[27].

On pourra toujours faire valoir que la méthode utilisée permet de comparer les biais linguistiques des différents moteurs de recherche. Cependant, il ne sera plus possible de prétendre que les résultats pour un moteur de recherche donné fournissent une représentation exacte de la proportionnalité de diverses langues dans l'ensemble de la Toile.

En réalité, la représentation graphique de l'évolution de la place des langues au fil des campagnes (***cf.*** chapitre V - Résultats) a

26 Cela devrait conduire à des pertes durables de niches commerciales pour Google comme dans le cas d'utiliser Exalead pour faire des recherches dans la Toile francophone.

27 Il est de plus en plus fréquent que des pages qui ne sont pas en anglais contiennent des publicités dynamiques en anglais.

déjà été altérée dans le passé par d'importants changements dans le comportement des moteurs de recherche. En 2001, la présence en ligne de toutes les langues mesurées a reculé proportionnellement par comparaison avec l'anglais et en contradiction avec la tendance observée. L'anglais faisait-il soudain un retour en force sur la Toile ? La présence asiatique croissante en ligne avait-elle renforcé la prolifération de l'anglais ? Une analyse attentive et patiente[28] a permis de conclure que cette situation était principalement le reflet de la refonte de l'index de Google qui, dans une phase de transition, avait augmenté le biais vers l'anglais (un biais qui a de toute façon toujours existé dans une certaine mesure).

Entre 2003 et 2004, Google et Yahoo étaient les deux meilleures options de moteur de recherche en termes de réponse aux exigences de la mesure linguistique en ligne. La taille considérable de l'index de Google (3 milliards de pages) et la clarté de sa gestion des diacritiques à cette époque ont conduit à choisir Google comme principal moteur de recherche. MSN a été écarté pour avoir une forte tendance à favoriser l'anglais et Exalead parce qu'il favorisait le français. La plupart des autres moteurs de recherche étaient liés à ceux mentionnés ci-dessus ou disposaient d'un trop petit index pour être utilisés.

En 2006, l'étude a été confrontée à une période prolongée de résultats incohérents pendant quatre campagnes de mesures et il était impossible de trouver la raison expliquant le comportement aberrant des moteurs de recherche. Il était même question de mettre fin au projet. Une explication fut finalement trouvée dans ce que l'on a baptisé l'opération Big Daddy[29] de Google qui a consisté en une refonte totale de son index et des serveurs hébergeant la base de données. La redéfinition et la reconstruction de l'index ont nécessité une longue période de transition. Il est apparu évident que la reconstruction avait une nette tendance à commencer par la Toile anglaise avant les autres langues. Cela a totalement faussé les résultats. Puis, ils sont progressivement

28 Plusieurs mois ont été nécessaires pour vérifier et rejeter les différentes hypothèses pour expliquer une telle situation, par exemple, savoir si la montée des pays asiatiques a déclenché une montée de l'anglais.

29 http://www.webworkshop.net/googles-big-daddy-update.html ou http://www.mattcutts.com/blog/bigdaddy

redevenus cohérents avec ceux des précédentes campagnes[30]. La confiance était revenue, jusqu'à ce que, quelques mois plus tard, Google entreprît de nouvelles modifications rendant le moteur définitivement inutilisable pour le projet. Ce dernier fait est survenu après plusieurs mois de travail en 2007 et a obligé l'équipe à revenir à Yahoo (qui utilise le moteur d'Altavista) jusqu'à ce qu'il soit finalement décidé de trouver une autre manière de poursuivre l'étude (ce qui est expliqué plus loin dans l'article).

La partie suivante fournit plus de détails sur les spécificités de la méthodologie utilisée, y compris une analyse statistique, le processus et les résultats.

4.3 MÉTHODOLOGIE STATISTIQUE

L'ensemble des 57 valeurs du total du nombre de pages de la Toile qui comporte la présence de chaque mot concept dans chaque langue, divisé par la valeur correspondante pour le même mot concept en anglais (ce qui représente le pourcentage d'une langue donnée par rapport à l'anglais), a été traité comme une ***variable statistique aléatoire*** sur laquelle sont appliqués les outils traditionnels pour une ***fonction de Gauss*** (ou ***distribution normale***). Le ***coefficient de variance***[31] est ensuite calculé. Une valeur de 0 indiquerait un résultat constant (ce qui est absolument impossible) alors qu'une valeur de 1 indiquerait une fonction exponentielle, représentative d'une situation aléatoire ***normale***. Entre 0 et 1, le coefficient de variance indiquerait un bon résultat (avec peu d'écart). Une valeur supérieure à 1 remettrait en question la validité de la méthode, indiquant une fonction hyperexponentielle représentant une dispersion excessive. Cette valeur était utilisée pour contrôler les mesures, les résultats supérieurs à 1 étant généralement le signe d'une anomalie dans le processus. De manière générale, les mesures ont toujours offert des résultats

30 La meilleure garantie de la méthode réside dans le fait que les nouveaux résultats des mesures ont toujours montré une sorte de continuité avec les résultats historiques. En outre, les changements de tendances dans les résultats ont toujours été accompagnés d'arguments valables sur ce qui s'est passé dans le domaine.

31 Le coefficient de variance est la racine carrée de l'écart-type au carré divisé par la moyenne au carré.

statistiques crédibles basés sur cet indicateur. Ensuite, l'intervalle de confiance à 90 % et 99 % était calculé à l'aide de la Loi de Student-Fisher, permettant de situer la validité des résultats dans une fenêtre.

4.4 CONCEPTION DES INDICATEURS

Le premier indicateur conçu concernait la présence d'une langue donnée dans l'Internet par rapport à sa présence dans le monde réel (ou ***présence pondérée***). Un quotient égal à 1 exprime la normalité, un quotient inférieur à 1 exprime une faible présence virtuelle (comme constaté pour l'espagnol et le portugais dans les premières éditions) et un quotient supérieur à 1 indique une forte présence virtuelle (comme c'est le cas pour l'anglais et, dans une moindre mesure, pour le français, l'italien et l'allemand). L'évolution de cet indicateur pour une langue donnée démontre la manière dont elle pourrait améliorer sa présence virtuelle et parvenir à une présence (dans le monde réel) normale ou même supérieure. Dans le cas de l'anglais, dont le quotient est encore largement au-dessus de 1, les douze années de mesure ont montré une baisse (de 7 à 4), puis une stabilisation de sa position (avec la limitation précédemment expliquée sur le fait que les résultats après 2005 ne sont pas représentatifs de l'ensemble de la Toile d'araignée mondiale). Cet indicateur pourrait être utile pour mesurer l'efficacité d'une politique linguistique virtuelle.

En utilisant les estimations du nombre d'internautes pour une langue donnée (fournies pendant de nombreuses années par GlobalStat[32] et depuis 2005 par Internet Worldstats[33]), il est possible d'établir un indicateur de la ***productivité linguistique*** (le nombre de pages produites par les internautes, normalisé à la valeur de 1). Il convient cependant de signaler les limites de fiabilité des chiffres de ces organisations. L'équipe de recherche considérait l'exactitude des chiffres fournis avec une marge de plus ou moins 20 %, puisque la méthodologie est basée sur des données fournies par de multiples sources nationales n'ayant pas forcément des approches normalisées. Ceci affecte évidemment

32 http://global-reach.biz/globstats/index.php3

33 http://www.internetworldstats.com/

les chiffres produits par Funredes et l'Union latine dans les mêmes proportions.

L'un des premiers résultats intéressants des mesures a été de découvrir que l'écart entre les grands producteurs et les faibles producteurs par langue n'était pas si important. La plupart des langues mesurées avaient un résultat proche de 1. Ceci implique une sorte de règle naturelle entre la proportion de producteurs de contenus et le nombre total d'internautes. En termes de politique linguistique, cela signifie qu'une politique qui veut renforcer la production de contenus dans une langue donnée, doit préalablement augmenter le nombre d'internautes. Une autre leçon à tirer de cet indicateur était que l'apparente loi naturelle de proportionnalité avait tendance à perdre de l'influence ces dernières années. Ceci pourrait être interprété par le fait que les internautes ayant adopté tardivement les technologies de l'Internet ont tendance à être plus consommateurs que producteurs de contenus (malgré l'essor des blogs). Ceci tend à conforter l'idée de création de nouvelles politiques orientées davantage sur l'alphabétisation numérique et la maîtrise de l'information que sur le simple accès à l'Internet.

D'autres indicateurs pour chaque langue sont indiqués dans le tableau ci-après (chiffres de 2007) et permettent d'approfondir l'analyse :

Tableau 5 : Indicateurs pour les langues dans l'Internet

	EN	ES	FR	IT	PO	RU	AL	CAT	
LOCUTEURS (MILLIONS)[5]	670	400	130	60	205	30	120	9	6 607[6]
LOCUTEURS EN % DE POPULATION MONDIALE	10,1 %	6,1 %	2,0 %	0,9 %	3,1 %	0,5 %	1,8 %	0,1 %	130 %[7]
INTERNAUTES PAR LANGUE (MILLIONS)[8]	366	102	58	31	47	5	59	2	1 154[9]
INTERNAUTES EN % DE LOCUTEURS	54,6 %	25,4 %	44,9 %	52,3 %	23,1 %	16,5 %	49,1 %	23,1 %	
INTERNAUTES EN % DE POPULATION MONDIALE	5,5 %	1,5 %	0,9 %	0,5 %	0,7 %	0,1 %	0,9 %	0,0 %	17,5 %
% D'INTERNAUTES PAR LANGUE	32 %	9 %	5 %	3 %	4 %	0 %	5 %	0,2 %	130 %

% PAGES DE LA TOILE PAR LANGUE[11]	45,0 %	3,8 %	4,4 %	2,7 %	1,4 %	0,3 %	5,9 %	0,1 %	100 %
PRODUCTIVITÉ LINGUISTIQUE PAR LANGUE[12]	1,42	0,43	0,87	0,98	0,34	0,66	1,16	0,74	1
PAGES DE LA TOILE PAR INTERNAUTES DANS UNE LANGUE DONNÉE	4,44	0,63	2,24	2,93	0,45	0,62	3,25	0,96	

Le rapport entre le nombre de locuteurs et d'internautes dans une langue donnée (*internautes en % de locuteurs*), est un indicateur de la pénétration de la langue sur l'Internet. Il informe sur l'évolution future de sa potentielle croissance. Par exemple, quand une langue atteint 50 % de pénétration, cela signifie que la courbe de croissance a certainement atteint son point d'inflexion et qu'elle va commencer à être asymptotique. La principale raison de la diminution relative de la langue anglaise sur l'Internet est donc tout simplement due au fait qu'elle a déjà culminé en atteignant une considérable présence initiale précoce et transitoire[34].

Le *pourcentage d'internautes par langue* proportionnellement à la population totale d'utilisateurs est un autre indicateur de la fracture numérique et linguistique. On peut par exemple l'interpréter ainsi : « 17,5 % de la population mondiale était connectée à l'Internet en 2007, dont 5,5 % étaient des internautes parlant anglais ». Le *pourcentage d'internautes par langue* est un indicateur également important. Il permet de révéler la diversité linguistique sur l'Internet et a montré une forte et constante évolution depuis les débuts de la Toile d'araignée mondiale.

D'autres indicateurs orientés sur une mesure qui se concentre plus sur les pays individuels (et exige de nouvelles astuces méthodologiques) sont produits depuis 2001. Ils présentent une grande richesse d'informations sur la dynamique de la production de contenus par langue et par pays. Ces indicateurs ont été progressivement étendus dans l'étude pour inclure le français, l'espagnol, l'anglais et le portugais (quatre langues utilisées dans

34 L'expérience du Minitel en France montre que lorsqu'on atteint 60 % de pénétration, la croissance devient très lente en dépit de l'absence de coûts directs.

de nombreux pays et pour lesquelles il est intéressant d'observer et de comparer la contribution par pays).

Ces indicateurs ont pu être obtenus grâce à la possibilité offerte par les moteurs de recherche de mesurer le nombre d'occurrences de pages mentionnées pour une recherche donnée par pays. Le programme a été exécuté à diverses reprises pour différents pays afin d'obtenir les résultats présentés.

Ici, la difficulté méthodologique réside dans le fait qu'il est insuffisant de mesurer l'échantillon par noms de domaines nationaux de premier niveau (***ccTLD*** en anglais), puisque de nombreux serveurs pour un pays donné utilisent les noms de domaines génériques de premier niveau (***gTLD*** en anglais)[35]. Il convient donc de répartir les contenus des ***gTLD*** entre les pays. Pour ce faire, il est nécessaire d'estimer le pourcentage des noms de domaines utilisant les ***ccTLD***. Ces chiffres sont obtenus de la part de collègues travaillant dans les Centres d'information des réseaux ou par la littérature sur le domaine.

Les résultats sont extrêmement riches en enseignements et constituent d'excellents outils pour les décideurs politiques, car ils peuvent être compilés par région, offrant des indications de mesure sur la fracture numérique entre le Nord et le Sud. En outre, ils peuvent donner une idée sur la production de contenus par pays dans une langue étrangère donnée. La méthode ne permettant pas de produire des données très précises, il convient de considérer les résultats avec prudence[36]. Pour plus de détails, voir le chapitre 5.3.

35 Principalement les domaines « .com » et « .org ».

36 Notamment en ce qui concerne le pourcentage des sites Internet américains qui utilisent le nom de domaine .us - chiffre difficile à connaître et qui a été calculé par essais et erreurs pour atteindre un total de 100 %.

5. RÉSULTATS

5.1 PRINCIPAUX RÉSULTATS

Les principaux résultats des campagnes de mesures sont présentés ci-dessous sous forme de graphiques et de tableaux.

POURCENTAGE DE PRÉSENCE SUR LA TOILE PAR RAPPORT À L'ANGLAIS

Le tableau suivant compare la présence des langues dans le cyberespace. Exprimé en pourcentages par rapport à l'anglais (ou l'anglais est de 100 %), il faut lire le tableau comme suit : en septembre 1998, pour 100 pages en anglais, il y avait 3 pages en espagnol, 4 pages en français, 2 pages en italien et 1 page en portugais. Pour avoir une page en roumain, 500 pages en anglais étaient nécessaires.

Tableau 6 : Présence des langues étudiées par rapport à l'anglais sur la Toile

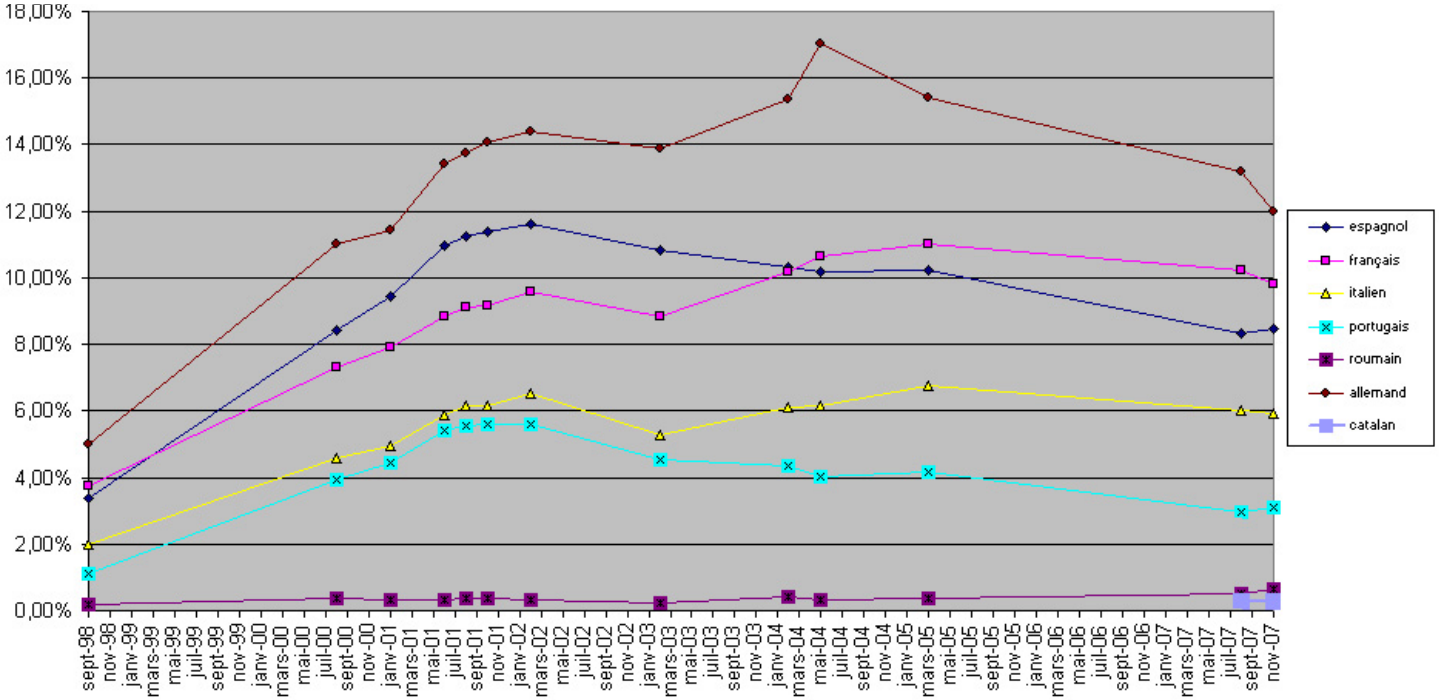
	ES	FR	IT	PT	RO	AL	CAT
09/98	3,37 %	3,75 %	2,00 %	1,09 %	0,20 %		
08/00	8,41 %	7,33 %	4,60 %	3,95 %	0,37 %	11,00 %	
01/01	9,46 %	7,89 %	4,93 %	4,44 %	0,33 %	11,43 %	
10/01	11,36 %	9,14 %	6,15 %	5,61 %	0,36 %	14,08 %	
02/02	11,60 %	9,60 %	6,51 %	5,62 %	0,33 %	14,41 %	
02/03	10,83 %	8,82 %	5,28 %	4,55 %	0,23 %	13,87 %	
05/04	10,19 %	10,64 %	6,15 %	4,02 %	0,31 %	17,02 %	
03/05	10,23 %	11,00 %	6,77 %	4,15 %	0,37 %	15,42 %	
08/07	8,33 %	10,21 %	6,01 %	2,97 %	0,50 %	13,17 %	0,27 %
11/07	8,45 %	9,80 %	5,92 %	3,09 %	0,63 %	13,12 %	0,30 %

Tableau 7 : Intervalles de confiance pour les résultats de 2008

	99 %	90 %	90 %	99 %
espagnol	6,56	7,48	10,74	11,66
français	8,19	8,98	11,77	12,56
portugais	4,01	4,65	6,9	7,53
italien	1,82	2,24	3,73	4,15
roumain	0,52	0,63	0,99	1,09
allemand	7,78	8,53	11,18	11,93
catalan	0,25	0,29	0,44	0,49

Ce tableau s'interprète comme suit : il y a 99 % de probabilités que le pourcentage de pages de la Toile en français par rapport à l'anglais soit compris entre 8,19 % et 12,56 % ; il y a 90 % de probabilités que le pourcentage de pages de la Toile en italien par rapport à l'anglais soit compris entre 2,24 % et 3,73 %.

Figure 2 : Graphique de l'évolution des pourcentages des langues étudiées par rapport à l'anglais



L'analyse du graphique ci-dessus montre deux phénomènes qui ont été expliqués dans la partie méthodologie de cet article :

- En 2003, la présence en ligne de toutes les langues mesurées a diminué dans la même proportion par rapport à l'anglais. Ceci a finalement été interprété comme le résultat d'une situation transitoire liée au changement de mode d'indexation de Google et non pas comme une diminution de la présence de ces langues sur la Toile. Une extrapolation des résultats de 2002 à 2004 reflèterait plus fidèlement la réalité, comme le graphique tend à l'indiquer.
- À partir de 2005, on constate une chute de la présence de toutes les langues mesurées. Au-delà de cette date, il est malheureusement impossible d'extrapoler les résultats de l'indexation des moteurs de recherche comme étant une représentation fidèle de la réalité sur la Toile. Ce qui est mesuré ne peut être considéré que comme une réalité des pages de la Toile indexées par un moteur de recherche spécifique. Cela indique un nouveau biais croissant en faveur de l'anglais pour les moteurs de recherche les plus communs.

Concernant les langues étudiées, on remarque un premier élan entre 1998 et 2002 de l'espagnol et du portugais entraîné par les efforts de l'Amérique latine pour augmenter l'accès à l'Internet, suivi par un affaiblissement relatif par rapport au français, à l'allemand et à l'italien. Un renforcement tardif de la présence du roumain sur la Toile a commencé en 2007 et son développement mérite d'être suivi afin de confirmer s'il va également se stabiliser.

Le tableau suivant constitue une estimation de la présence absolue des langues sur la Toile. Il a été obtenu à partir d'une estimation de l'anglais, puis de l'application des pourcentages comparatifs pour les autres langues de l'étude. L'estimation de l'anglais est établie par itération, en jouant sur les valeurs des autres langues. Il est de plus en plus difficile d'établir avec précision cette estimation, en raison de l'explosion du nombre d'usagers en Asie, mais également du biais des moteurs de recherche (en faveur de l'anglais).

Tableau 8 : Pourcentage absolu des langues étudiées sur la Toile

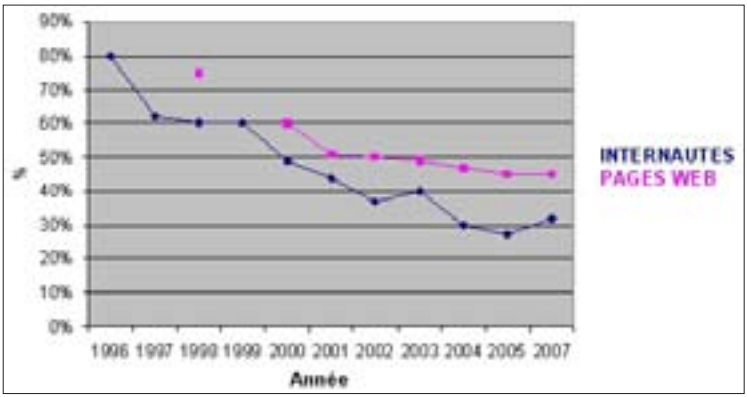
	EN	ES	FR	IT	PT	RO	AL	CAT	Somme[13]	Reste du monde[14]
09/98	75,0 %	2,53 %	2,81 %	1,50 %	0,82 %	0,15 %	3,75 %		11,56 %	13,44 %
08/00	60,0 %	5,05 %	4,40 %	2,76 %	2,37 %	0,22 %	3,00 %		17,80 %	22,20 %
01/01	55,0 %	5,20 %	4,34 %	2,71 %	2,44 %	0,18 %	6,29 %		21,16 %	23,84 %
06/01	52,0 %	5,69 %	4,61 %	3,06 %	2,81 %	0,17 %	6,98 %		23,31 %	24,69 %
08/01	51,0 %	5,73 %	4,66 %	3,14 %	2,84 %	0,18 %	7,01 %		23,55 %	25,45 %
10/01	50,7 %	5,76 %	4,63 %	3,12 %	2,84 %	0,18 %	7,14 %		23,68 %	25,62 %
02/02	50,0 %	5,80 %	4,80 %	3,26 %	2,81 %	0,17 %	7,21 %		24,04 %	25,97 %
02/03	49,0 %	5,31 %	4,32 %	2,59 %	2,23 %	0,11 %	6,80 %		21,35 %	29,65 %
02/04	47,0 %	4,84 %	4,78 %	2,86 %	2,05 %	0,19 %	7,21 %		21,94 %	31,06 %
05/04	46,3 %	4,72 %	4,93 %	2,85 %	1,86 %	0,14 %	7,88 %		22,38 %	31,32 %
03/05	45,0 %	4,60 %	4,95 %	3,05 %	1,87 %	0,17 %	6,94 %		21,57 %	33,43 %
08/07	45,0 %	3,75 %	4,59 %	2,70 %	1,34 %	0,23 %	5,93 %	0,12 %	18,53 %	36,47 %
11/07	45,0 %	3,80 %	4,41 %	2,66 %	1,39 %	0,28 %	5,90 %	0,14 %	18,46 %	36,54 %

L'asymptote apparente de l'anglais à 45 % (voir Figure 2) est due au nouveau biais des moteurs de recherche plutôt qu'à un véritable phénomène sur la topologie linguistique de la Toile. Si la courbe des utilisateurs anglophones est un indicateur fiable des tendances, comme il est raisonnable de le penser (voir ci-dessous), la présence de l'anglais sur la Toile (par opposition à l'indexation des moteurs de recherche) est probablement au-dessous de 40 % ; les valeurs de la dernière colonne pour 2007 souffrent du même problème. La réalité est probablement au-dessus de 40 % pour le reste des langues, principalement en raison de la présence massive du chinois en ligne.

Tableau 9 : Évolution du pourcentage d'internautes anglophones et de pages de la Toile en anglais (tableau)

	96	97	98	99	00	01	02	03	04	05	07
INTERNAUTES[15]	40	72	91	148	192	231	234	288	280	300	366
%INTERNAUTES	80 %	62 %	60 %	60 %	49 %	44 %	37 %	40 %	30 %	27 %	32 %
% PAGES			75,0 %		60,0 %	51,0 %	50,0 %	49,0 %	47,0 %	45,0 %	45,0 %

Figure 3 : Évolution du pourcentage d'internautes anglophones et de pages de la Toile en anglais (graphique)



La forte augmentation entre 2005 et 2007 du pourcentage d'internautes anglophones indiquée dans le graphique ci-dessus est le résultat du changement de source de données de GlobalStat (qui a arrêté de fournir ce genre de données statistiques) vers InternetWorldStats. C'est une manifestation indirecte des limites de ces chiffres comme indiqué dans le chapitre 4.4.

5.2 ANALYSE PAR PAYS

Les résultats les plus intéressants et novateurs de l'étude ont été obtenus grâce à l'application de la méthode utilisant les noms de domaine pour les pays anglophones, francophones, hispanophones et lusophones. Cette approche permet d'obtenir des données frappantes.

L'ensemble des résultats pour chaque langue peut être consulté sur les sites suivants :

- http://funredes.org/lc/francais/medidas/sintesis.htm (pour les résultats de 2005)
- http://dtil.unilat.org/LI/2007/index_fr.htm (pour les résultats de 2007)

Le nombre de résultats intéressants étant trop important pour être décrit en détail dans cet article, une synthèse est proposée ci-après.

Tableau 10 : Principaux pays producteurs de pages de la Toile en français : pourcentage total de pages, suivi de la productivité[37]

	11/2007	5/2005	3/2003
FRANCE	60 % -1,09	60 % - 0,82	54 % - 0,96
CANADA	20 % - 1,06	19 % - 1,27	24 % - 1,83
BELGIQUE	7 % - 0,60	8 % - 1,55	7 % - 2,21
SUISSE	5 % - 0,87	5 % - 2,78	6 % - 2,17
AUTRES	8 % - 0,84	8 % - 1,38	9 % - 3,10

Le Canada (et particulièrement le Québec) a été un pionnier dans la production de contenus sur la Toile, et c'est pour cette raison que sa productivité apparaît en déclin. Par opposition, la France a eu une production tardive, mais qui a connu une forte expansion en 2005.

Deux tendances peuvent être remarquées : premièrement, une décroissance générale de la productivité (sauf pour la France) et deuxièmement une forte décroissance en Belgique et en Suisse, témoignant de l'arrivée d'un grand nombre de nouveaux internautes, mais de peu de production de nouveaux contenus ou de pages.

37 Calculés comme étant le rapport entre le % de production et le % d'internautes par langue

Tableau 11 : Pages de la Toile en français : production par région

		11/2007	5/2005	3/2003
EUROPE		75 %	79 %	71 %
AMÉRIQUE		22 %	21 %	25 %
AFRIQUE/MOYEN-ORIENT		0,3 %	0,4 %	0,4 %
ASIE/OCÉANIE		0,2 %	0,4 %	0,4 %
NON CLASSÉ		2,11 %	0,19 %	3,32 %

La triste réalité de la fracture numérique est évidente dans le tableau ci-dessus si l'on regarde les résultats concernant l'Afrique. À ce jour, aucun changement n'a pu encore être constaté. Les pays francophones en développement qui apparaissent comme étant les plus productifs en 2007 sont le Maroc et le Sénégal, mais il est important de noter que l'Allemagne ou le Royaume-Uni produisent plus de pages en français que tous les pays d'Afrique réunis.

En ce qui concerne l'espagnol, le tableau suivant montre les principaux producteurs de contenus ainsi que leur productivité :

Tableau 12 : Principaux pays producteurs de pages de la Toile en espagnol : pourcentage total de pages, suivi de la productivité

	2007	2005	2001
ESPAGNE	56 % - 3,4	48 % - 2,4	54 % - 2,7
ÉTATS-UNIS	10 % - 0,4	14 % - 0,4	5 % - 0,12
ARGENTINE	9.4 % - 0,9[16]	10,6 % - 1,9	9,6 % - 1,3
MEXIQUE	8.4 % - 0,45	7,4 % - 0,5	8,6 % - 0,45

En 2001, les États-Unis avaient plus d'internautes hispanophones que l'Espagne, mais celle-ci a produit 54 % du total de production de contenus en espagnol et les États-Unis seulement 5 %. Depuis, la productivité étasunienne s'est améliorée, mais sans jamais atteindre la valeur moyenne de 1. Notons que le Mexique, qui a la plus forte population d'immigrés aux États-Unis, a une productivité tout aussi faible. En termes de politique publique de création de contenus en espagnol, cela pourrait se traduire par

une plus grande concentration sur la frontière virtuelle entre les États-Unis et le Mexique.

La plus forte productivité se trouve à Cuba - de 3,4 en 2001 à 4,3 en 2007 - avec le Nicaragua qui suit de près. Ceci traduit à la fois le faible nombre d'internautes et une politique visant à systématiser les publications sur la Toile dans le monde universitaire dans ces pays.

Tableau 13 : Principaux pays producteurs de pages de la Toile en anglais : pourcentage total de pages, suivi de la productivité

	11/2007	5/2005
ÉTATS-UNIS	66 % - 1	51 % - 0,8
ROYAUME-UNI	6,5 % - 0,6	7,2 % - 0,6
CANADA	3,5 % - 0,7	5 % - 0,7
AUSTRALIE	1,5 % - 0,3	1,8 % - 0,4
ALLEMAGNE	1,2 % - 39	1,9 % - 57

L'inclusion de l'Allemagne dans le tableau ci-dessous, comme grand producteur de pages de la Toile, illustre un phénomène qui était prévisible : de nombreux pays dans lesquels l'anglais n'est pas la langue principale contribuent de manière considérable à la production de contenus en anglais.

Certains domaines nationaux de premier niveau de très petits états insulaires montrent des résultats anormalement élevés de production de contenus en anglais, du fait que leurs noms de domaine nationaux ont été vendus à l'étranger à des fins commerciales (comme le Tuvalu avec .tv, Nioué avec .nu, la Micronésie avec .fm et les Samoa avec .ws).

Des considérations sur la diversité linguistique en ligne et sur la production de pages et la productivité des internautes à la fois par pays et par région démontrent l'étendue de la fracture numérique. L'ensemble des résultats montre que la production totale de pages de la Toile produites par les pays africains en anglais et en français atteint à peine les 0,33 % de la production mondiale pour ces langues. Concernant les statistiques pour l'Afrique, 97 % du contenu est généré par l'Afrique du Sud. Les autres pays non anglophones de l'Organisation de coopération et de

développement économiques (OCDE) produisent plus de 0,1 % de ce total, ce qui représente un tiers de la production africaine. En outre, de nombreux pays individuels, comme l'Allemagne, la France, l'Italie ou le Japon produisent plus de contenus en anglais que tous les pays d'Afrique réunis, y compris l'Afrique du Sud.

Avec les indicateurs conçus par le Language Observatory Project (LOP)[38], cette situation projette un message que les groupes de TIC pour le développement résistent toujours à intégrer dans leurs actions : la fracture numérique concerne tout autant, si ce n'est plus, la production des contenus que l'accès à l'Internet (voir Accessing content dans les références). La fracture des contenus, qui est une fracture linguistique et culturelle, est un indicateur inquiétant du risque d'acculturation des populations qui parviennent à accéder à l'Internet et ne disposent pas de contenus dans leur langue maternelle. Cela devrait conduire au rééquilibrage des politiques sur la fracture numérique et donner une plus grande priorité à l'alphabétisation numérique et informationnelle (qui sont d'évidents éléments déclencheurs de la production de contenus et de l'éthique de l'information). La lutte contre la fracture numérique n'est pas qu'une simple question d'accès et d'infrastructure.

Tableau 14 : Principaux pays producteurs de pages de la Toile en portugais : pourcentage total de pages, suivi de la productivité

	11/2007	5/2005
BRÉSIL	71 % - 0,90	71 % - 0,95
PORTUGAL	15 % - 0,98	17 % - 1,0
ÉTATS-UNIS	4 % - 5,0	8 % - 5,4
ESPAGNE	3,8 % - 3,7	2,3 % - 1,2

Le tableau 14 ci-dessus indique que le Brésil occupe une place dominante en termes de production de pages de la Toile en portugais, avec une production de contenus stable, produisant 71 % de l'ensemble des pages en portugais de l'Internet. Il

38 http://www.language-observatory.org/

convient également de souligner que les États-Unis produisent plus de pages de la Toile en portugais que l'Espagne.

5.3 AUTRES ESPACES DE DIVERSITÉ LINGUISTIQUE

La diversité linguistique peut être étudiée dans de nombreux autres espaces du cyberespace. Dans les premières années, des mesures ont été effectuées sur l'espace Usenet (forums) et ont donné lieu à d'intéressants résultats[39]. Plus récemment, la diversité linguistique des blogues a également été évaluée. Les résultats obtenus avec différents moteurs de recherche de blogues sont si hétérogènes qu'ils ne méritent pas d'être publiés. Le fait est qu'aujourd'hui chaque moteur de recherche de blogues est associé à un serveur particulier qui effectue seulement des recherches dans l'index de ces blogues en particulier. Dans le futur, un métamoteur de recherche additionnant les résultats des différents moteurs de recherche de blogues pourrait certainement se révéler utile.

Les résultats obtenus à partir de l'étude de Wikipédia, dont les statistiques multilingues sont impressionnantes[40], méritent d'être cités. Ils confirment qu'il s'agit bien de l'un des espaces les plus linguistiquement diversifiés de l'Internet. Le tableau ci-après informe sur la production d'articles dans diverses langues (source : Wikipédia, juillet 2008) :

Tableau 15 : Articles Wikipédia par langue

anglais	2 259 431	23,078 %
allemand	715 830	7,312 %
français	629 004	6,425 %
polonais	475 566	4,857 %
japonais	472 691	4,828 %
italien	418 969	4,279 %
néerlandais	413 325	4,222 %
portugais	363 323	3,711 %
espagnol	337 860	3,451 %

39 http://www.funredes.org/lc2005/english/L4index.html

40 http://fr.wikipedia.org/wiki/Wikipédia:Statistiques_internationales

5.4 DIVERSITÉ CULTURELLE

La méthodologie utilisée pour évaluer la présence et la distribution de la diversité culturelle sur l'Internet est assez simple, voire même simpliste. Elle ne peut être prise en considération que de manière approximative. Elle ne traduit pas vraiment toute la complexité du sujet en question, qui peut être mesuré, entre autres, sur une base aussi bien thématique que « national ». Pour fournir une indication de la diversité culturelle, plusieurs thèmes ont été sélectionnés et pour chacun d'entre eux une longue, quoique non exhaustive, liste de personnalités pertinentes a été définie (par exemple, Albert Einstein pour la science ou Pablo Picasso pour les arts graphiques). L'« index de citation sur la Toile » a ensuite été calculé pour chaque personnalité et les résultats ont été compilés. Un indicateur simple a alors été conçu et utilisé pour suivre l'évolution de la diversité culturelle en ligne au cours de cinq campagnes de mesures menées en 1996, 1998, 2001, 2005 et 2008. Considéré dans son ensemble, cela fournit une mise en perspective de la diversité culturelle sur les 12 dernières années.

Les thèmes pris en considération pour mesurer la diversité culturelle sont les suivants :

- Littérature
- Science
- Musique (tous genres)
- Cinéma
- Arts graphiques
- Politique
- Personnalités (personne célèbre ou présente dans les médias pour une raison quelconque)
- Histoire
- Fiction (Dracula ou Cendrillon, par exemple)
- Un mot (personnalités extraites de différents thèmes en un seul mot, comme Einstein ou Picasso)

Des calculs ont été effectués sur un total d'environ 1 200 personnalités[41]. Ci-dessous sont présentés quelques exemples de résultats obtenus (l'ensemble des résultats peut être consulté sur la Toile[42]). Un code de couleur a été mis en place dans les résultats afin de mieux identifier leur catégorisation culturelle. La seconde colonne indique les pertes et gains de places dans le classement par rapport aux précédentes mesures.

Tableau 16 : Premières places en littérature

	2008	
1	WILLIAM SHAKESPEARE	0
2	OSCAR WILDE	2
3	VICTOR HUGO	3
4	CHARLES DICKENS	4
5	AGATHA CHRISTIE	21
6	PAULO COELHO	3
7	J.R.R. TOLKIEN	-2
8	ERNEST HEMINGWAY	15
9	EDGAR POE	9
10	JULES VERNE	1

	2005	
1	WILLIAM SHAKESPEARE	0
2	RENÉ DESCARTES	26
3	GABRIEL GARCÍA MÁRQUEZ	34
4	OSCAR WILDE	-1
5	J.R.R. TOLKIEN	7
6	VICTOR HUGO	-4
7	LORD BYRON	14
8	CHARLES DICKENS	-4
9	PAULO COELHO	53
10	IMMANUEL KANT	20

	2001	
1	WILLIAM SHAKESPEARE	0
2	VICTOR HUGO	1
3	OSCAR WILDE	-1
4	CHARLES DICKENS	2
5	WILLIAM JAMES	0
6	JAMES JOYCE	2
7	ERNEST HEMINGWAY	7
8	WALT WHITMAN	-1
9	EDGAR POE	-5
10	HENRY JAMES	1

41 Une modification a été effectuée lors de la deuxième campagne de mesure, afin d'obtenir un échantillon de personnalités plus complet. Par la suite, le même échantillon a été conservé.

42 Voir : http://funredes.org/lc/espanol/cultura08/cultura08.htm

Tableau 17 : Premières places en science

	2008	
1	ALBERT EINSTEIN	0
2	NOAM CHOMSKY	1
3	CHARLES DARWIN	1
4	MARIE CURIE	-2
5	SIGMUND FREUD	4
6	ISAAC NEWTON	-1
7	THOMAS EDISON	5
8	CARL SAGAN	2
9	MILTON FRIEDMAN	4
10	GALILEO GALILEI	-3
11	BLAISE PASCAL	-5
12	LOUIS PASTEUR	-1

	2005	
1	ALBERT EINSTEIN	0
2	MARIE CURIE	0
3	NOAM CHOMSKY	4
4	CHARLES DARWIN	-1
5	ISAAC NEWTON	0
6	BLAISE PASCAL	4
7	GALILEO GALILEI	4
8	ALEXANDER VON HUMBOLDT	4
9	SIGMUND FREUD	-5
10	CARL SAGAN	-1
11	LOUIS PASTEUR	-3
12	THOMAS EDISON	-6

	2001	
1	ALBERT EINSTEIN	0
2	MARIE CURIE	9
3	CHARLES DARWIN	0
4	SIGMUND FREUD	0
5	ISAAC NEWTON	-3
6	THOMAS EDISON	0
7	NOAM CHOMSKY	0
8	LOUIS PASTEUR	0
9	CARL SAGAN	-4
10	BLAISE PASCAL	-1
11	GALILEO GALILEI	-1
12	ALEXANDER VON HUMBOLDT	2

Tableau 18 : Premières places en arts graphiques

	2008	
1	LEONARDO DA VINCI	0
2	ANDY WARHOL	1
3	SALVADOR DALI	-1
4	PABLO PICASSO	6
5	VINCENT VAN GOGH	6
6	CLAUDE MONET	1
7	FRIDA KAHLO	-3
8	GUSTAV KLIMT	1
9	EL GRECO	-1
10	JOAN MIRO	4
11	PAUL GAUGUIN	4

	2005	
1	LEONARDO DA VINCI	0
2	SALVADOR DALÍ	1
3	ANDY WARHOL	-1
4	FRIDA KAHLO	7
5	PAUL CÉZANNE	9
6	HENRI MATISSE	6
7	CLAUDE MONET	-1
8	EL GRECO	-1
9	GUSTAV KLIMT	6
10	PABLO PICASSO	-6
11	VINCENT VAN GOGH	-6

	2001	
	LEONARDO DA VINCI	0
2	ANDY WARHOL	0
3	SALVADOR DALÍ	0
4	PABLO PICASSO	0
5	VINCENT VAN GOGH	0
6	CLAUDE MONET	0
7	EL GRECO	1
8	MARC CHAGALL	4
9	DIEGO RIVERA	-2
10	PAUL KLEE	1
11	FRIDA KAHLO	-2

Tableau 19 : Premières places en un mot[43]

	2008	
1	WASHINGTON	0
2	CLINTON	2
3	DALÍ	34
4	DISNEY	1
5	LINCOLN	-2
6	CHRIST	3
7	KENNEDY	-5
8	MADONNA	27
9	JEFFERSON	-3
10	BACH	18

	2005	
1	WASHINGTON	0
2	KENNEDY	5
3	LINCOLN	1
4	CLINTON	-1
5	DISNEY	0
6	JEFFERSON	2
7	NEWTON	-1
8	EINSTEIN	5
9	CHRIST	-7
10	DARWIN	18

	2001	
1	WASHINGTON	0
2	CHRIST	1
3	CLINTON	1
4	LINCOLN	-2
5	DISNEY	0
6	NEWTON	0
7	KENNEDY	2
8	JEFFERSON	-1
9	GORE	7
10	DALÍ	39

Que révèlent ces mesures ?

Premièrement, là où culture et commerce sont étroitement liés, comme c'est le cas de la musique et du cinéma, le biais en ligne vers la culture américaine est évident. Cependant, dans les thèmes où la culture est seule concernée, comme en littérature, en science ou dans les arts graphiques, la représentation culturelle sur l'Internet, mesurée à partir de personnalités, n'est pas biaisée. La présence d'auteurs français en littérature ou de chercheurs français est aussi évidente que la présence de peintres hispanophones. Les deux premières campagnes ont révélé un handicap pour la représentation culturelle française, et encore plus pour la culture espagnole, mais il a été surmonté en 2005. Depuis, il n'y a plus eu de changement significatif et c'est pour cela que la campagne 2008 sera probablement la dernière à suivre cette méthodologie.

Deuxièmement, l'Internet est un média hautement réceptif qui reflète rapidement les événements de la vie réelle et qui se « désintéresse » aussi vite de certains évènements et personnalités. Ceci explique l'apparition puis la chute de certaines personnalités qui sont propulsées par un événement historique (comme la sortie d'un film sur Che Guevara en 2008, donnant

43 Il est à noter que « Bush » ne fait pas partie de l'échantillon, dans le cas contraire, en 2008, il serait arrivé second devant « Clinton ».

soudain une plus grande importance à cette personnalité) ou plus subtilement, par une tendance sociologique rendant ces personnalités plus ou moins à la mode (l'exemple des évolutions respectives de la cybercélébrité d'Albert Camus et de Jean-Paul Sartre est intéressant).

Troisièmement, une certaine « culture mondiale » peut être perçue sur l'Internet. Il est fort probable que cette culture ait tendance à exclure des éléments ou personnalités importants et extrêmement pertinents pour les cultures locales, mais qui n'ont pas su trouver leur place à une échelle mondiale. Pourtant, pour en revenir à une considération de la langue dans la représentation de la diversité culturelle, la question de savoir comment la culture des minorités est représentée en ligne est largement ouverte. Cela inclut les cultures comme celles des populations autochtones, qui ne sont pas réellement des minorités, mais demeurent malgré tout des victimes de la fracture numérique en termes d'accès aux TIC.

6. ÉVALUATION DE LA MÉTHODE

6.1 SON CARACTÈRE UNIQUE ET SES AVANTAGES

Une considération objective de la méthodologie et des résultats décrits ci-dessus fait ressortir les atouts suivants :

- La méthode décrite ci-dessus utilise de façon logique et productive des outils en ligne extrêmement polyvalents, à savoir les moteurs de recherche. Au cours de la période de recherche, la seule limite dans la mesure des citations des mots concepts dans le cyberespace résidait dans leur connectivité normative aux index des moteurs de recherche. La question de savoir s'il est efficace de s'appuyer sur de tels index pour en déduire des résultats sur le cyberespace trouve sa réponse dans une autre question : quel est l'intérêt pratique d'une page qui n'est pas indexée ?
- En outre, le projet de recherche a constamment et sur une longue période de temps été l'un des rares projets parmi ses pairs à révéler avec transparence sa façon de procéder, les résultats et la méthode détaillée.
- Même s'il est effectivement impossible de proposer une sélection parfaitement culturellement neutre des mots concepts, toutes les précautions nécessaires ont été prises, tant du point de vue linguistique que culturel, pour minimiser les biais. La liste de mots concepts (et de personnalités) a été choisie afin d'offrir les résultats les plus fiables possible.
- Une méthode statistique rigoureuse et normalisée a été utilisée afin d'assurer la cohérence des 13 campagnes de mesures menées entre 1996 et 2008. Cela a permis de donner encore plus de crédibilité au projet.
- Contrairement aux autres méthodes utilisées dans d'autres projets de recherche en matière de mesure de la diversité linguistique en ligne, celle-ci a permis d'évaluer le contenu d'autres espaces que la Toile et d'obtenir des résultats précis

par langue et par pays - ce qui représente l'unique approximation de la sorte et qui a mené à la création de puissants indicateurs.

- Jusqu'à présent, cette étude a été la seule à offrir des séries de mesures cohérentes donnant une large perspective du sujet, tous les autres projets de recherches menés évoqués s'étant limités à une seule évaluation, ou tout au plus à des évaluations à court terme.

6.2 SES FAIBLESSES ET SES LIMITES

La méthode proposée et expliquée ci-dessus présente néanmoins certaines faiblesses :

- Elle se limite à un très petit nombre de langues et le coût marginal pour ajouter une nouvelle langue est relativement élevé. Elle n'est pas une approche pratique pour la généralisation de la mesure de la présence de la plupart des langues sur la Toile : des algorithmes de reconnaissance de langue appliqués sur des bases de données obtenues par l'exploration automatique (***crawling***)[44] de la Toile (comme pour l'étude du LOP) constitueront très probablement la méthode standard du futur.
- Elle ne fournit pas directement de valeur absolue par langue. L'estimation de la présence absolue de l'anglais, utilisée pour calculer les valeurs des autres langues, est établie par un processus non systématique devenant chaque jour de plus en plus problématique en raison de l'évolution des technologies des moteurs de recherche et de la croissante diversification des langues dans le cyberespace.
- Elle ne mesure la présence des langues que dans la partie indexée des moteurs de recherche. Cet inconvénient n'avait pas d'importance jusqu'en 2005. À cette époque, les moteurs de recherche couvraient plus de 60 % de la Toile visible, rendant l'extrapolation des résultats des mesures pertinente. Cependant, après 2005, cet inconvénient est devenu

44 Processus automatique et systématique qui consiste à parcourir les pages de la Toile (et éventuellement à en stocker une représentation) comme le font les moteurs de recherche. Voir : http://fr.wikipedia.org/wiki/Robot_d%27indexation.

prépondérant, et aujourd'hui l'évolution des moteurs de recherche requiert une nouvelle méthode basée sur une exploration automatique de la Toile et un comptage direct.

- Certains indicateurs se fondent sur des chiffres qui peuvent être controversés (comme le nombre de personnes parlant une langue dans le monde) ou plutôt peu fiables (comme le nombre de personnes parlant une langue donnée et utilisant l'Internet) et cela a bien entendu des conséquences sur l'intervalle de confiance de certains des indicateurs.

7. ÉVALUATION D'AUTRES MÉTHODES

Diverses méthodes alternatives pour mesurer l'espace occupé par les langues dans le cyberespace ont été utilisées et publiées au cours de la vie du projet. De nombreux résultats ont été publiés par des mercaticiens, sans que les méthodes utilisées soient clairement décrites. Ce qui suit est une sélection des actions considérées comme les plus pertinentes sur le domaine.

Babel Team : une initiative commune d'Alis Technologies et de l'Internet Society

La première initiative, présentée comme étant pionnière en la matière, bien qu'effectuée plusieurs mois après les premières études de Funredes, a été menée par la société canadienne Alis Technologies. Elle a été publiée avec le soutien de l'Internet Society en juin 1997. Même si le rapport[45] promettait d'effectuer deux mesures par an, elle se limita à cette seule expérience. Il est intéressant de noter que cette étude proposait la première tentative de ce qui deviendrait la méthode utilisée plus tard, et à deux reprises (1992 et 2002), par l'OCLC, afin de soutenir le débat médiatique de la présence constante de l'anglais à 80 % sur la Toile pendant cette période (voir How «World Wide» is the Web? et Trends in the Evolution of the Public Web: 1998 – 2002 dans les références).

La méthode Alis est fondée sur un échantillon aléatoire de 8 000 pages d'accueil de sites internet[46] sur lesquels un algorithme de reconnaissance linguistique capable d'identifier 17 langues différentes est appliqué afin d'obtenir la répartition des langues puis d'extrapoler les résultats à l'ensemble de la Toile. Avant extrapolation, une vérification manuelle est effectuée sur un sous-ensemble de sites pour détecter des erreurs de reconnaissances

45 http://alis.isoc.org/palmares.en.html

46 Le nombre effectif de sites analysés dépasse légèrement les 3 000.

et ainsi corriger les résultats d'une manière qui n'est pas décrite dans la documentation.

La méthode et même certaines données sont exposées de façon transparente, comme la liste des adresses IP analysées. Ses principaux inconvénients résident dans le fait, premièrement, qu'elle n'a pas été reproduite malgré la promesse initiale et deuxièmement, que les résultats ont été publiés après une seule mesure. Ce second point ôte tout crédit aux résultats[47] et il est important de comprendre ces limites pour évaluer le projet de l'OCLC.

1) Cette méthode présuppose que la page d'accueil est représentative de la répartition des langues sur la totalité du site alors que de nombreux sites en d'autres langues possèdent des pages d'accueil en anglais ou bilingue.
2) Les algorithmes de reconnaissance de langues n'étaient pas - et ne sont toujours pas - complètement fiables, même s'ils se sont améliorés depuis 1997 et ont tendance à donner des résultats favorisant l'anglais[48].

Mais ces limites sont mineures par rapport aux deux suivantes :

3) D'un point de vue statistique, on peut mettre en doute le fait que 3 000 serveurs sélectionnés aléatoirement puissent être représentatifs d'un univers qui en comptait alors environ un million. En d'autres termes, comment un échantillon aléatoire équivalent à 0,3 % du nombre total de serveurs dans le monde peut-il refléter efficacement la diversité de l'univers du cyberespace ? Certes, une telle méthode utilisée pour les sondages politiques permet de prévoir avec une bonne précision les résultats d'une élection, mais dans ce cas, l'échantillon n'est pas pris au hasard. Au contraire, il est construit avec soin afin d'obtenir une juste représentation de la structure du corps électoral (âge, sexe, lieu d'habitation, etc.).

En outre, la seule façon d'évaluer sérieusement cette hypothèse de travail n'a pas été effectuée, ce qui conduit à la dernière et principale limite de cette étude :

47 Les résultats donnent un pourcentage d'anglais sur la Toile supérieur à 80 %.

48 Le Language Observatory Project estime la marge d'erreur à 10 %.

4) Pour avoir une certaine validité statistique et une rigueur scientifique, il convient de procéder à des mesures répétées avec un échantillon aléatoire différent d'adresses IP permettant l'analyse et la distribution de la variable aléatoire, afin de comprendre son comportement statistique, comme cela a été fait dans la méthode Funredes/Union latine, par l'addition de 57 concepts. Cependant, cela aurait rendu les vérifications manuelles du processus trop laborieuses et coûteuses en temps.

Projet OCLC de caractérisation de la Toile

L'OCLC est un projet célèbre en termes de mesure de la diversité linguistique en ligne. Il offre des services intéressants pour les documentalistes et a fourni des données cohérentes et fiables sur le profil démographique d'Internet avec le projet de caractérisation de la Toile jusqu'en 2003[49]. Cependant, les données sur la présence des langues sur la Toile ont été obtenues en reprenant et poursuivant la méthodologie d'Alis et souffraient donc des limites précédemment décrites. Il annonçait ainsi la même valeur de 72 % pour la présence de l'anglais sur la Toile en 1999 et 2002, lors des deux seules campagnes de mesure réalisées avec la même méthodologie[50].

La dernière publication réalisée en 2002 conclut d'une part que « la croissance sur la Toile publique, mesurée grâce au nombre de sites Internet, a atteint un plateau » et d'autre part « qu'il n'y a aucun signe que la répartition du contenu centré sur les États-Unis et dominé par l'anglais ait tendance à se mondialiser » (voir Trends in the Evolution of the Public Web: 1998 - 2002 dans les références).

Au même moment, l'étude Funredes/Union latine a montré une présence de l'anglais à 50 % et une tendance prononcée vers la diversification linguistique de la Toile. Étant l'unique source d'information étasunienne sur le sujet et bénéficiant du soutien prestigieux de l'OCLC, cette étude a soutenu, sur la base d'une méthode défectueuse, l'idée selon laquelle la présence de

49 http://www.oclc.org/research/projects/archive/wcp/default.htm

50 http://www.oclc.org/research/projects/archive/wcp/stats/intnl.htm

l'anglais restait constante, aux alentours de 80 %, contre toutes tendances évidentes et visibles. Ainsi, elle a servi involontairement de référence aux médias, qui ont colporté l'idée selon laquelle la Toile était dominée par l'anglais.

Même après quelque temps, alors qu'il devenait évident que la démographie de la Toile évoluait à une vitesse incroyable, elle est restée la référence pour les articles sur le sujet jusqu'en 2005, comme celui de Paollillo (voir Mesurer la diversité linguistique sur Internet dans les références), alors qu'on avait enregistré une chute de la proportion des utilisateurs anglophones d'Internet de 60 % à 30 % entre 1999 et 2005.

Étude Inktomi

En février 2000, Inktomi, un des principaux moteurs de recherche[51] de l'époque, promouvait grâce à une campagne publicitaire très réussie, les résultats de son « étude »[52] sur la présence de l'anglais sur la Toile, avec les résultats suivants :

Tableau 20 : Résultats de l'étude Inktomi

LANGUE	PROPORTION (%)
anglais	86,54
allemand	5,83
français	2,36
italien	1,55
espagnol	1,23
portugais	0,75
néerlandais	0,54
finnois	0,50
suédois	0,36
japonais	0,34

Le total des pourcentages des langues mentionnées dans le tableau ci-dessus atteint 100 %, alors que de nombreuses autres langues étaient présentes sur la Toile. Cela affecte évidemment la

51 Racheté en 2002 par Yahoo!, il a aujourd'hui disparu.

52 Jamais aucune référence sur la méthode mise en œuvre n'a été citée.

valeur absolue réelle de l'anglais[53]. Cette opération commerciale a gravement contribué à la désinformation sur la place de l'anglais sur la Toile en 2000. L'estimation médiatique de la prédominance de l'anglais aux environs de 80 % était le témoin du manque d'intérêt scientifique sur le sujet. Si l'on estimait la proportion des pages de la Toile du reste des langues en plus de celles mentionnées dans le tableau ci-dessus comme étant de plus de 20 %, la proportion de pages en anglais serait de moins de 70 %.

La méthode du complément de l'ensemble vide

À partir de mars 1998 avec Altavista (et ensuite avec Google), il a été découvert que certains moteurs de recherche avaient la courtoisie de donner des informations sur la composition des pages indexées en fonction des langues telles qu'ils les percevaient. En faisant une recherche sur une expression telle que « -bhjvfvvj » - indiquant une recherche pour tout sauf rien, d'où le nom de ***méthode du complément de l'ensemble vide***, la réponse serait l'ensemble de l'index avec la quantité totale de pages indexées. La même requête dans une langue donnée produirait une estimation du nombre de pages indexées dans cette langue. De toute évidence, cette méthode reflète l'important biais des algorithmes de reconnaissance de langue en faveur de l'anglais[54], et doit être considérée uniquement comme une approximation grossière des langues sur la Toile. Dans tous les cas, cette méthode a été fréquemment utilisée pour vérifier l'évolution de l'anglais et l'apparition de nouvelles langues sur la Toile. Le 3 juillet 2008, Google indiquait la répartition suivante dans sa base de données :

53 Les pourcentages auraient dû être calculés en fonction du nombre total de langues ou il aurait fallu mettre une proportion pour les autres langues dans le tableau.

54 Pour comprendre ce biais, il suffit de faire quelques expériences de recherche de mots en anglais et de constater le taux élevé de pages d'autres langues qui sont prises pour de l'anglais.

Tableau 21 : Estimation du nombre de pages de la Toile par langue dans Google

LANGUE	PAGES	POURCENTAGE
arabe	340 000 000	0,68 %
bulgare	169 000 000	0,34 %
catalan	46 400 000	0,09 %
chinois (simplifié)	3 770 000 000	7,49 %
chinois (traditionnel)	796 000 000	1,58 %
croate	113 000 000	0,22 %
tchèque	269 000 000	0,53 %
danois	249 000 000	0,49 %
néerlandais	583 000 000	1,16 %
anglais	25 580 000 000	50,82 %
estonien	129 000 000	0,26 %
finlandais	225 000 000	0,45 %
français	1 750 000 000	3,48 %
allemand	2 470 000 000	4,91 %
grec	148 000 000	0,29 %
hébreu	290 000 000	0,58 %
hongrois	278 000 000	0,55 %
islandais	27 100 000	0,05 %
indonésien	132 000 000	0,26 %
italien	951 000 000	1,89 %
japonais	3 040 000 000	6,04 %
coréen	968 000 000	1,92 %
letton	43 200 000	0,09 %
lituanien	95 600 000	0,19 %
norvégien	255 000 000	0,51 %
polonais	675 000 000	1,34 %
portugais	828 000 000	1,65 %
roumain	254 000 000	0,50 %
russe	1 470 000 000	2,92 %
serbe	61 800 000	0,12 %
slovaque	181 000 000	0,36 %
slovène	97 500 000	0,19 %
espagnol	2 180 000 000	4,33 %
suédois	116 000 000	0,23 %
turc	835 000 000	1,66 %
arménien	2	0,00 %

LANGUE	PAGES	POURCENTAGE
biélorusse	959 000	0,00 %
esperanto	3 740 000	0,01 %
perse	116 000 000	0,23 %
tagalog	8 300 000	0,02 %
thaïlandais	418 000 000	0,83 %
ukrainien	69 100 000	0,14 %
vietnamien	301 000 000	0,60 %
TOTAL	50 332 699 002	100,00 %

Le suivi de ces informations a permis de constater que dès 2000, un grand nombre de sources avaient utilisé cette méthode, sans fournir d'explications, afin d'établir leur propre estimation des langues sur la Toile[55]. Dans tous les cas, depuis 2005, les résultats présentés par Google avec cette méthode ne sont pas du tout fiables, car ils varient de façon considérable d'une fois sur l'autre.

Étude Xerox

En 2001, une étude a été publiée utilisant une technique linguistique inédite fondée sur la fréquence d'apparition de mots usuels dans des corpus linguistiques donnés. Le but était de pouvoir prédire la présence de langues données sur la Toile (voir ***Estimation of English and non-English Language Use on the WWW*** dans les références). Cette étude présente les résultats pour 1996, 1999 et 2000 et mérite donc d'être considérée comme étant la première étude sur la présence des langues sur la Toile. Les résultats produits par l'étude Xerox sur la présence des langues sont légèrement en dessous des pourcentages produits par Funredes, comme le montre le tableau ci-après :

55 C'est le cas par exemple de Vilaweb qui a réalisé une bonne publicité en diffusant de tels résultats en 2001.

Tableau 22 : Résultats de l'étude XEROX

Par rapport à l'anglais	XEROX 10/96	XEROX 8/99	XEROX 2/2000	*FUNREDES 8/2000*
allemand	3,8 %	7,1 %	6,9 %	*11 %*
français	3,7 %	5,4 %	5,7 %	*7,33 %*
espagnol	1,7 %	4,0 %	3,9 %	*8,41 %*
italien	2,0 %	2,9 %	2,8 %	*4,60 %*
portugais	1,7 %	2,1 %	2,4 %	*3,95 %*

Language Observatory Project (LOP)

Depuis 2003, le LOP est mené à la Nagaoka University of Technology du Japon. Un programme de reconnaissance linguistique a été développé dans le cadre de ce projet. Il permet d'identifier les langues, scripts et encodages d'une page de la Toile à l'aide d'une méthodologie statistique largement utilisée dans le traitement de texte, censée offrir une marge d'erreur de 9 % et identifier 350 langues. En 2006 et 2007, à l'aide d'un robot d'indexation (*crawler*) développé par l'université de Milan, le LOP a collecté et identifié près de 100 millions de pages avec des noms de domaines nationaux de premier niveau asiatiques (excepté pour la Chine, le Japon et la Corée, afin d'éviter le traitement de masse) et africains.

L'étude montre qu'en 2006 et 2007, respectivement 40 % et 41 % des pages de la Toile ayant un nom de domaine national asiatique et respectivement 73 % et 82 % des pages de la toile ayant un nom de domaine national africain étaient écrites en anglais (voir *A Language and Character Set Determination Method Based on N-gram Statistics* et *Analysis of the Asian Languages on the Web Based on N-gram Language Identification* dans les références). À l'avenir, le projet sera étendu à des domaines nationaux d'autres régions et à des domaines génériques.

Les résultats suivants ont été obtenus par le LOP pour les langues qui sont également celles étudiées par Funredes et l'Union latine en 2007 :

Tableau 23 : Résultats de l'étude du Language Observatory Project

	anglais	espagnol	français	italien	portugais	Roumain	allemand	catalan
Asie	41,5 %	0,02 %	0,25 %	0,01 %	0,03 %	0,03 %	0,21 %	0,04 %
Afrique	82,1 %	0,11 %	7,0 %	0,07 %	1,2 %	0,03 %	0,82 %	0,04 %

Le LOP, qui regroupe un large consortium de scientifiques et de spécialistes d'universités et d'organisations de 20 pays différents, a certainement le potentiel d'évoluer en un projet de référence pour la mesure de la diversité linguistique sur l'Internet. Son orientation actuelle fait de lui un outil unique pour le contrôle de la présence des langues minoritaires. Si et quand l'espace exploré s'étendra à l'intégralité de la Toile, on obtiendra le résultat qui a été attendu pendant des années, dans la limite de l'algorithme de reconnaissance linguistique qui a une marge d'erreur proche de 10 %.

Projet de l'Universitat Politècnica de Catalunya (UPC) / Institut d'Estadística de Catalunya (IDESCAT)

Ce projet a démarré en 2003 avec l'IDESCAT et a été réalisé par l'UPC. Il a consisté en la collecte d'une base de données de quelque 30 millions de noms de domaine à partir desquels il en a été extrait un sous-ensemble de 2 millions auquel ont été appliqués des algorithmes de reconnaissance linguistique, comme dans le projet du LOP.

Le projet montre les résultats suivants pour 2005 – lesquels sont relativement proches de l'étude Funredes/Union latine :

Tableau 24 : Résultats de l'étude UPC en 2005

Langue	UPC	FUNREDES
anglais	42,9 %	45 %
catalan	0,16 %	0,14 % (2007)
espagnol	3,4 %	4,60 %
allemand	6,2 %	6,94 %
français	6,2 %	4,95 %
italien	7,9 %	3,05 %

Et les résultats suivants pour 2006 – lesquels sont très différents de l'étude Funredes/Union latine :

Tableau 25 : Résultats de l'étude UPC en 2006

Langue	UPC
anglais	71,8 %
catalan	0,47 %
espagnol	2,14 %
allemand	14,5 %
français	3,8 %
italien	0,7 %

Diversité culturelle sur la Toile

L'équipe de recherche n'a trouvé qu'une seule autre tentative de mesure de la diversité culturelle sur la Toile, à savoir celle entreprise par des chercheurs espagnols en 2003 (voir ***Iconos culturales hispanos en Internet*** dans les références). Le fait est que ce travail s'est contenté de reprendre la méthodologie simpliste de Funredes en tentant de l'améliorer, mais ces améliorations marginales[56], qui semblent avoir nécessité des investissements importants[57], ne montrent pas de nouveaux résultats significatifs.

56 Comme mentionné précédemment, la méthode est trop simple pour refléter la complexité de la culture, et les quelques améliorations apportées par ces chercheurs n'ont pas permis de changer cette réalité ni d'offrir un changement de paradigme pour l'interprétation des résultats.

57 L'équipe de recherche de Funredes a été surprise lorsqu'elle a lu l'article et découvert que le groupe de chercheurs avait préféré établir un contrat avec une entreprise privée pour refaire le programme de Funredes, plutôt que de contacter Funredes pour établir une possible collaboration !

8. PERSPECTIVES

Le domaine de recherche de la diversité linguistique sur l'Internet sort d'une période de difficultés et de manque d'intérêt de la part des organismes internationaux et du monde universitaire. Le besoin de politiques linguistiques pour protéger et promouvoir les langues sur la Toile est de plus en plus évident et la nécessité de disposer d'indicateurs fiables vient naturellement de concert. Des projets tels que le LOP proposent de nouvelles approches pouvant produire une vaste quantité d'informations sur la présence des langues sur l'Internet. Cependant, les exigences sont bien plus complexes et vont au-delà de la question de la proportion des langues dans les différents cyberespaces. L'utilisation des langues dans les courriels, dans les sessions de clavardage et dans les sites visités est encore une donnée inconnue. Pourtant, elle est tout aussi importante puisqu'elle démontre la dynamique des langues dans le comportement des internautes[58].

Dans ce contexte, l'étude de Funredes/Union latine serait symptomatique de la période préhistorique de la mesure de la diversité linguistique sur l'Internet. L'évolution actuelle des moteurs de recherche montre qu'elle a atteint ses limites dans sa forme actuelle. Néanmoins, l'opportunité d'une niche pour une méthode alternative permettant de valider les résultats d'autres méthodes fondées sur l'utilisation d'algorithmes de reconnaissance de langues existe toujours. Afin d'obtenir des résultats pertinents, une restructuration de la partie de la méthodologie basée sur les moteurs de recherche est cependant nécessaire.

En d'autres termes, une telle évolution de la méthode de recherche nécessite un système d'exploration de la Toile avec application du contrôle linguistique directement durant le processus

58 L'idée d'utiliser des approches comme celle utilisée par Alexa.com pour mesurer le comportement linguistique des utilisateurs semble très prometteuse. Alexa installe un logiciel espion volontaire dans les ordinateurs personnels des personnes qui le souhaitent, et effectue un compte rendu de leurs choix de navigation. De là, Alexa est en mesure de calculer des données intéressantes à partir des comportements des utilisateurs, y compris un classement des sites Internet visités.

d'exploration. Des compteurs indiqueraient les résultats à la fin de l'exploration. Cela permettrait d'ajouter des éléments d'évaluation de la qualité, en plus du comptage des mots de notre méthode. Ceci annonce un domaine de recherche prometteur qui n'a pas encore été exploré.

La faisabilité et le coût de cette nouvelle méthode de recherche évoluée sont actuellement examinés. Funredes recherche un partenariat, notamment pour l'aspect relatif à l'exploration de la Toile de ce projet potentiel. En outre, il est envisagé, en collaboration avec l'Université Antilles-Guyane d'ajouter les langues créoles à base française dans la liste des langues traitées.

Indépendamment de l'avenir de la méthode Funredes/Union latine, la diversité linguistique dans le cyberespace devient une question de plus en plus prioritaire pour la construction des sociétés du savoir. Ce fait à lui seul va créer un plus grand besoin de conception d'indicateurs professionnels capables d'aider à contrôler les politiques linguistiques dans le cyberespace.

RÉFÉRENCES

Crystal, D. Language and the Internet. Cambridge University Press, 2001 , ISBN-10: 0521802121 , ISBN-13: 978-0521802123

Bergman, M.K. White Paper: The Deep Web: Surfacing Hidden Value. In Ann Arbor, MI: Scholarly Publishing Office, University of Michigan, University Library, vol. 7, no. 1, August 2001. http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104

Paolillo,J.; Pimienta D.; Prado, D.; et al. Mesurer la diversité linguistique sur Internet. Institut de statistique de l'UNESCO Montréal, Canada - UNESCO, 2005 (CI.2005/WS/06). http://portal.unesco.org/ci/fr/ev.php-URL_ID=20882&URL_DO=DO_TOPIC&URL_SECTION=201.html

Lavoie, B.F.; O'Neill, E.T. How "World Wide" is the Web? Annual Review of OCLC Research, 1999. http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003496

O'Neill, E. T.; Lavoie B.F.; Bennett, R. Trends in the Evolution of the Public Web: 1998 - 2002. D-Lib Magazine, 9.4., 2003. http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

Grefenstette, G.; Noche, J. Estimation of English and non-English Language use on the WWW, Technical Report from Xerox Research Centre Europe, 2000. http://arxiv.org/ftp/cs/papers/0006/0006032.pdf

Suzuki I.; Mikami Y.; and ale. A Language and Character Set Determination Method Based on N-gram Statistics. In ACM Transactions on Asian Language Information Processing, vol.1, no.3, September 2002, pp.270-279.

Nandasara S.T.; et al. Analysis of the Asian Languages on the Web Based on N-gram Language Identification. In The International Journal on Advances in ICT for Emerging Regions (ICTer), volume 1, issue 1, 2008.

Monràs F.; et al. Estadística de la presència del català a la xarxa d'Internet i de les característiques dels Webs Catalans in Llengua i ús: Revista tècnica de política lingüística, ISSN 1134-7724, nº. 37, 2006, pags. 62-66.

Cueto L.; Soler C.; Noya J. Iconos culturales hispanos en Internet (lo que ven los buscadores). In El Español en el Mundo : anuario del Instituto Cervantes 2004 / coord. por Paz Lorenzo, ISBN 84-01-37892-3 , pag. 127-190, 2004. http://www.realinstitutoelcano.org/publicaciones/109/109.pdf

Diki-Kidiri, M.; Comment assurer la présence d'une langue dans le cyberespace ? UNESCO, CI.2007/WS/1. 2007. http://unesdoc.unesco.org/images/0014/001497/149786F.pdf

Pimienta D.; Accessing content in Global Information Society Watch, 2008, APC, ITEM, HIVOS Editors. http://www.giswatch.org/gisw2008/thematic/Accessing Content.html

Notes de bas de page des tableaux

1 Estimation du nombre de locuteurs (première ou seconde langue) pour les principales langues parlées dans le monde :

X = NOMBRE DE LOCUTEURS (MILLIONS)	**LANGUES**
X > 500	CHINOIS, ANGLAIS, LANGUES INDIENNES
200 < X < 500	ESPAGNOL, RUSSE, LANGUES ARABES
100 < X < 200	BENGALI, PORTUGAIS, JAPONAIS, INDONÉSIEN, ALLEMAND, FRANÇAIS

2 http://fr.wikipedia.org/wiki/A%C3%AFnou_(langue_de_l%27ethnie_du_Japon)

3 http://en.wikipedia.org/wiki/Yagan

4 http://funredes.org/lc2005/L6/francais/evol.html

5 Source : Union latine (2000)

6 Estimation de la population mondiale. Le nombre total de locuteurs devrait cependant être plus élevé, compte tenu des personnes parlant plus d'une langue.

7 30 % serait une estimation intuitive de la population parlant plus d'une langue. Ce chiffre est probablement proche de la réalité dans les pays de l'OCDE, mais pas dans la plupart des pays en développement ou une personne moyenne parle 2 ou 3 langues (comme en Afrique).

8 Source Internet Word Stats (2005).

9 Estimation du nombre total d'internautes.

10 Pourcentage de la population mondiale ayant accès à l'Internet.

11 Source FUNREDES/Union latine (2005).

12 Ratio entre le % de pages de la Toile par langue et le % d'utilisateurs d'Internet par la langue.

13 Somme des langues étudiées hormis l'anglais.

14 Somme de la totalité des autres langues du monde.

15 En millions. Source : Global Reach jusqu'en 2005, puis Internetworldstats.

16 Une baisse de la productivité avec un tel pourcentage de production de contenus indique une augmentation du nombre d'internautes non suivie par une augmentation de la production de contenus.

Secrétariat
UNESCO
Secteur de la Communication et de l'Information
Division de la Société de l'Information
1, rue Miollis
75732 Paris cedex 15
France

Tél. : + 33.1.45.68.45.00
Fax : + 33.1.45.68.55.83

www.unesco.org/webworld

Paris: UNESCO, 2010