Sommet mondial sur la société de l'information

Le texte complet des ouvrages
est disponible à :
http://www.unesco.org/wsis

UNESCO

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

Sommet mondial sur la société de l'information

Mesurer la diversité linguistique sur Internet

Mesurer la diversité linguistique sur Internet

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

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

2005

Mesurer la diversité linguistique sur Internet

Un ensemble d'articles signés par :

**John Paolillo, Daniel Pimienta,
Daniel Prado et autres**

Révisé et accompagné
d'une introduction de l'Institut
de statistique de l'UNESCO
Montréal (Canada)

Publié en 2005
Par l'Organisation des Nations Unies
pour l'éducation, la science et la culture
7, place de Fontenoy, 75352 PARIS 07 SP
Composé et imprimé dans les ateliers de l'UNESCO

Printed in France

(CI-2005/WS/06 CLD 24822)

Table des matières

1.

Introduction

L'UNESCO a mis en lumière le concept de « société du savoir », qui met l'accent sur la pluralité et la diversité, plutôt que sur l'uniformité généralisée, comme étant susceptible de réduire le fossé numérique et de donner naissance à une société de l'information inclusive. Le multilinguisme est l'un des thèmes importants que sous-tend ce concept, pour assurer une diversité culturelle et une participation de toutes les langues dans le cyberespace. Il existe une inquiétude croissante à l'effet que, dans la foulée des efforts de réduire le fossé numérique, des centaines de langues locales puissent être laissées de côté, bien que de façon non intentionnelle. Il en découle l'importance qui est accordée à la diversité linguistique et au contenu local dans le cadre d'une ligne d'action du Plan d'action du Sommet mondial sur la société de l'information (SMSI) dont la responsabilité de la coordination a été confiée à l'UNESCO.[1]

De plusieurs façons inattendues, l'enjeu de la diversité linguistique sur Internet se révèle au cœur du débat qui entoure la société de l'information. De prime abord, la question semble tourner autour des communautés qui utilisent Internet, leur permettant ainsi de se parler les unes avec les autres dans leurs langues maternelles, mais d'autres questions ne tardent pas à surgir.

Par l'entremise de quels canaux la communication s'exprime-t-elle sur Internet ? La Toile mondiale s'apparente à une série de sources d'information générant peu d'interactivité. Les forums de discussions et le courriel permettent des échanges plus directs. Cependant, il existe trop peu de renseignements au sujet des langues utilisées dans les courriels ou les forums de discussion (consultez certains échanges sur ce sujet dans l'article de John Paolillo au chapitre 3, y compris les efforts de Sue Wright).

1 Voir le discours de l'UNESCO à la Délégation permanente au SMSI prononcé, le 8 juillet 2005, par Koïchiro Matsuura, Directeur Général.

Pour la plupart des chercheurs en analyse du langage, il faut par conséquent se tourner vers les pages Web. Dans ce domaine, comme dans toutes les formes de communications, nous devons prendre en considération les caractéristiques de l'auditoire. Une page Web ne pourra être lue que par des gens qui disposent d'un accès à Internet. Conséquemment, alors que la diversité linguistique pourrait bénéficier de l'existence de pages Web dans la langue « en voie de disparition » d'une tribu fort lointaine, très peu de gens les liraient car il est peu vraisemblable que les membres d'une tribu puissent disposer d'un accès à Internet. Par contre, des pages au sujet de la langue de la tribu rédigées dans une langue plus internationale pourraient jouer un rôle important en attirant l'attention sur la valeur culturelle de la langue visée et, possiblement, susciter un soutien pour le groupe linguistique dont il est question. De plus, il s'agirait d'une contribution à la préservation de langues en voie de disparition.

Les articles de ce volume illustrent l'existence de nombreux problèmes techniques au niveau de l'évaluation de la diversité linguistique sur Internet. Nous pouvons facilement obtenir un compte aléatoire de pages sur Internet en utilisant un nombre quelconque de moteurs de recherche commerciaux, mais nous ne pouvons pas évaluer à quelle fréquence ces pages Web sont lues ou encore si la lecture d'une page a aidé le lecteur d'une façon ou d'une autre. Aussi il est nécessaire de s'assurer que les éléments qui font l'objet d'une recherche dans différentes langues possèdent une valeur, une signification et une utilisation équivalentes (voir Pimienta).

Les langues et la société de l'information

L'Institut de statistique de l'UNESCO s'est engagé à adopter une approche d'évaluation de la société de l'information qui se situe au-delà d'un aperçu technocentriste pour considérer l'impact social d'Internet et d'autres canaux de diffusion de l'information. D'énormes problèmes se doivent d'être surmontés en ce qui a trait à

— la standardisation des définitions pour obtenir une comparabilité internationale ;

— l'identification d'indicateurs pertinents pour les politiques des pays développés et en voie de développement ;

— le renforcement des capacités aux niveaux national et international pour permettre la collecte régulière de données de qualité.

La langue est le médium qui permet tous les échanges suscités par la société de l'information. La langue est un médium fondamental de toute communication, le fondement grâce auquel les individus et les collectivités s'expriment que ce soit sous la forme d'une tradition orale ou d'un texte écrit. Pour l'UNESCO, l'enseignement de la langue maternelle s'apparente à un droit pour tous les enfants. L'UNESCO soutient aussi la diversité linguistique en s'assurant que la richesse culturelle représentée par la diversité sera préservée dans tous les pays et dans le monde entier.

L'enjeu culturel des langues sur Internet s'oppose une perception des problèmes entourant la société de l'information centrée sur les technologies de l'information et de la communication (TIC) et leurs répercussions. L'Institut de statistique de l'UNESCO souhaite présenter un point du vue qui soulève des questions au sujet de l'importance des problèmes associés au « contenu » et à l'environnement favorable qui, par la même occasion, lèverait le voile sur les problèmes techniques relatifs à la mesure de la culture et du contenu dans la société de l'information.

Les articles apparaissant dans ce volume présentent une variété de perspectives diverses sur la nature de ce problème. L'étude que signe le professeur John Paolillo présente le point de vue d'un linguiste professionnel oeuvrant dans l'univers anglophone. L'étude comporte quatre grands chapitres. Le premier chapitre traite du cadre éthique relatif à l'évaluation des biais des systèmes informatiques et il établit un lien entre ce cadre et le statut, sur Internet, des langues parlées dans le monde. Le deuxième chapitre porte sur les enjeux des biais préexistants au niveau du développement récent d'Internet, et fait appel à des statistiques relatives à la croissance d'Internet et leurs rapports avec la diversité linguistique à l'échelle mondiale. Le troisième chapitre examine les enjeux des biais linguistiques qui apparaissent dans le sillage d'Internet. Le quatrième chapitre se penche sur de tels biais dans les systèmes techniques d'Internet.

Comme compléments à ce texte, sont présentés un ensemble d'articles plus brefs issus de contextes non anglophones. Ils ont été recueillis et organisés par Daniel Pimienta de FUNREDES, un projet non gouvernemental dans le cadre duquel a été élaboré un système d'énumération des langues dans une perspective

qui privilégie les langues latines. Pimienta adoptant le point de vue d'une ONG de la société civile, décrit les obstacles auxquels sont confrontés les groupes locaux au niveau de l'accès à Internet et un aperçu des indicateurs actuellement disponibles. Son article est suivi d'une note, en provenance de Daniel Prado, présentant la réaction de la communauté linguistique « néo-latine » à la domination apparente de l'anglais. Ces articles plus brefs comportent aussi un point de vue asiatique fort intéressant de Yoshiki Mikami et autres, ainsi qu'une note sur la situation en Afrique signée par Xavier Fantognan qui résume la situation qui prévaut en Afrique dans une perspective africaine.

Le volume n'offre aucune réponse définitive quant à la façon de mesurer les langues sur Internet, mais il tente de réfuter nombre des mythes entourant les chiffres qui ont couramment été publiés. Il précise que le simple fait de compter des pages Web ne suffit pas et qu'il faudra que les fournisseurs de service Internet et les gouvernements consentent plus d'efforts de développement. Chaque auteur présente, dans une perspective qui lui est propre, un certain nombre de suggestions sur les façons d'aborder ces nécessaires efforts de développement.

Diversité linguistique sur Internet : un aperçu

La diversité linguistique peut en soi faire l'objet de différentes interprétations. L'anglais est une langue parlée assez uniformément à la grandeur des pays où elle joue un rôle prédominant. La Papouasie-Nouvelle-Guinée compte plus de 830 langues. Les résidents des pays anglophones peuvent posséder de nombreuses autres aptitudes langagières, mais rares sont les pays qui peuvent rivaliser avec la Papouasie en ce qui concerne la diversité au sein d'un même pays. Même si le nombre de locuteurs de langues néo-latines, y compris ceux aux États-Unis, peut équivaloir au double du nombre de personnes de langue maternelle anglaise (voir Daniel Prado), néanmoins les États-Unis contrôlent en grande partie les rouages qui sous-tendent la Toile mondiale (voir John Paolillo, Yoshiki Mikami). Le rapport entre les langues sur Internet et la diversité linguistique au sein d'un pays indique que, même en présence d'un réseau mondial, les Etats nations continuent d'avoir un rôle à jouer au niveau de l'encouragement à apporter à la diversité linguistique dans le cyberespace. La diversité linguistique peut être perçue autant à l'intérieur même d'un pays qu'à l'échelle d'Internet dans son ensemble.

Il est communément admis que l'anglais joue un rôle dominant sur Internet. Les articles dans ce volume diffèrent en ce qui concerne l'interprétation à donner à cette question. John Paolillo se rallie à la proposition et il prend pour acquis, comme la plupart des gens qui perçoivent l'anglais comme la langue dominante, que cela pose problème. Daniel Pimienta considère que l'anglais englobe environ la moitié de toutes les pages Web, et que cette proportion est en baisse à mesure que d'autres nations et groupes linguistiques accroissent leur présence sur la Toile. John Paolillo s'attarde sur la domination des États-Unis sur les forces qui sous-tendent la Toile, tant de nature commerciale que réglementaire, dans la mesure où cette dernière existe. Mikami appuie Paolillo sur ce point et met l'accent sur les difficultés qu'il y a à réconcilier les technologies, états-uniennes ou occidentales, de l'information et des communications et les conventions linguistiques avec les scriptes asiatiques. Cependant, Mikami laisse entrevoir, tout comme Pimienta, qu'un changement est sur le point d'intervenir à mesure qu'augmentent les clientèles indienne et chinoise. Cette divergence d'opinion sur la domination de la langue anglaise et l'avenir des langues sur la Toile ne peut être facilement résolue. Au bout du compte, cette division pourrait amener à faire l'illustration de la difficulté de mesurer l'utilisation des langues sur Internet à l'égard de laquelle, malgré la myriade de plates-formes des TIC, mais en partie à cause d'une pénurie de réglementation et d'une croissance phénoménale, nous ne possédons aucun indicateur statistique valable. Pimienta laisse à penser que le domaine des indicateurs d'Internet a été en grande partie pris en charge par les entreprises commerciales et qu'il existe un besoin pour des analyses universitaires de grande qualité.

Paolillo allègue que les compagnies de télécommunications qui profitent de la demande pour des services de technologie et de communication se doivent de garder à l'esprit qu'ils possèdent une responsabilité spéciale en ce qui concerne la diversité linguistique des pays dont ils desservent les marchés. Les sociétés actives dans les domaines du matériel et des logiciels informatiques exercent une influence de même type sur la mise en place linguistique d'Internet, en produisant des ordinateurs qui disposent de claviers, de présentations et de systèmes d'exploitation qui favorisent certaines langues en particulier. Les gestes posés par les sociétés informatiques tournées essentiellement vers la concurrence pour la domination du marché ont des effets nuisibles sur le climat de l'informatique multilingue et de la diversité linguistique en ligne. Dans de telles circonstances, la prise de conscience ethnolinguistique des compagnies de télécommunications, des sociétés informatiques et des autorités qui régissent Internet ne commencera

à s'étendre que si une masse critique de groupes ethnolinguistiques sous-représentés puisse retenir leur attention. Par conséquent, l'enjeu général des biais linguistiques émergents exige une surveillance étroite à l'échelle mondiale, régionale et locale.

La mesure des langues sur Internet peut être utilisée à titre de paradigme pour de nombreux enjeux relatifs à la mesure du contenu. Mais, à proprement parler, si nous ne pouvons pas mesurer cette dimension apparemment simple du contenu d'un site Web, que pouvons-nous mesurer au juste ? Toutefois, nous ne devrions pas faire preuve d'autant de pessimisme. Le projet de Mikami offre de grandes possibilités de composer avec plusieurs des problèmes techniques soulevés par les articles précédents et qui, de son propre aveu, adoptaient un point de départ non anglophone.

Il nous faut opter pour la mise au point d'indicateurs plus intelligents. Le fait de mesurer la présence des langues sur un nombre global de pages Web suscite de plus en plus de défis à cause de l'ampleur même de leur contenu, et la présence d'une page sur le Web ne signifie pas pour autant qu'elle soit utilisée ou même qu'elle soit « visitée ». Si nous voulons vraiment mesurer l'impact de la société de l'information, nous avons besoin de statistiques sur les modalités d'utilisation d'Internet et sur ses utilisateurs. À cet égard, les pages Web se présentent tout simplement comme des mesures visant l'offre, dans toute sa diversité ou homogénéité linguistique, et pas nécessairement comme un outil de réflexion de l'utilisation et de la demande. Dans un marché d'offre excédentaire de pages Web, par exemple en anglais, qui offrent une variété de services, il est possible que de nombreux sites médiocres ne reçoivent que peu ou pas de visiteurs. Il est aussi de notoriété publique que, pendant des années, de nombreux sites Internet ne sont pas mis à jour ou demeurent tels quels.

D'un point de vue économique, la Toile présente certains aspects d'un marché libre et un certain nombre d'échecs du marché (voir Paolillo). Les sites Web sont élaborés pour répondre aux besoins d'un auditoire particulier. Si l'accès à Internet est peu développé sur le marché intérieur, les sites Web commerciaux seront conçus en fonction d'un marché étranger extérieur et, par conséquent, seront écrits dans une langue internationale comme l'anglais. D'autre part, une faible utilisation d'un site Internet ainsi que les coûts peu élevés d'entretien de sites Web signifient qu'ils peuvent continuer d'exister et d'être enregistrés auprès de moteurs de recherche bien après la dernière visite d'un utilisateur éventuel.

D'un point de vue idéal, il nous faut une analyse de sites « utiles » et des visiteurs qui les fréquentent.

Même en tenant compte des limites des présentes études, ces dernières révèlent à quel point les statistiques sur le pourcentage des personnes qui possèdent un ordinateur ou sur le nombre d'abonnements à Internet (deux indicateurs des Objectifs de développement du millénaire) sont peu révélateurs sur les changements fondamentaux en matière d'échange d'information auxquels a donné lieu la société de l'information. Si nous mettons de côté les arguments à l'appui ou à l'encontre de la domination de la langue anglaise, nous pouvons constater dans ce volume la rapide expansion de l'utilisation d'Internet en Asie et, conséquemment, la croissance des sites Web en langues asiatiques (voir Mikami) et, dans la foulée de l'expansion du Web, les modalités de rapprochement des communautés « néolatines » afin d'examiner la place qu'elles occupent dans une société du savoir mondiale (voir Prado). Il est important de souligner que l'univers numérique fournit un environnement porteur à autant de langues que possible. Cela pourrait assurer une véritable inclusion linguistique numérique.

Les prochaines étapes

Il est à souhaiter que ce rapport relève le besoin, tel que suggéré ci-dessus, pour tous les organismes d'œuvrer aux niveaux national et international. Le Sommet mondial sur la société de l'information offre un contexte favorable à des discussions portant à la fois sur la politique linguistique et les normes technologiques, ainsi que sur les objectifs d'une politique à venir par la promotion d'une libéralisation des échanges d'information.

Les études montrent à quel point il est important de comprendre le contexte culturel propre à l'échange d'information. Étant donné cette situation, il semble improbable qu'une quelconque perspective mondiale soit susceptible de fournir des données comparables ou pertinentes par rapport à la politique qui s'avèrent suffisamment sensibles aux enjeux soulevés sur le plan technique et sur celui de la politique. Il y aurait plutôt lieu de confier à des initiatives régionales la direction de la surveillance, et les résultats de leurs études pourraient ensuite être fusionnés dans une perspective globale à l'échelle mondiale. Le projet FUNREDES et l'Observatoire de Mikami constituent deux projets éventuels susceptibles de nous montrer les modalités de fonctionnement d'un tel réseau régional.

En conclusion, ainsi que l'énonce si adroitement Paolillo dans son rapport, il se peut que des gestes soient nécessaires pour s'assurer que les valeurs de l'accès numérique et de l'alphabétisme numérique soient confirmées, tout spécialement pour le compte des nombreux pays en voie de développement et touchés par une diversité linguistique.

L'UNESCO recommande aux instances nationales, régionales et internationales de travailler ensemble afin de fournir les ressources nécessaires et à prendre les mesures qui s'imposent pour alléger les barrières linguistiques et promouvoir l'interaction humaine sur Internet en favorisant la création, le traitement et l'accès à un contenu éducatif, culturel et scientifique sous forme numérique, de façon à s'assurer que toutes les cultures puissent s'exprimer et avoir accès au cyberespace dans toutes les langues, y compris les langues indigènes.[2]

2 Pour plus de plus amples renseignements, veuillez consulter : La Recommandation de l'UNESCO relative à la promotion et à l'utilisation du multilinguisme et l'accès universel au cyberespace et le document 32 C/27, 2003, Déclaration sur la diversité culturelle de l'UNESCO, Paris, 02.11.2001.

2.

Modèles et approches

a. Diversité linguistique dans le cyberespace : modèles de développement et de mesure

Daniel Pimienta, FUNREDES

Introduction

Il est un mot que les acteurs et actrices de la société civile sur le thème de la société de l'information, spécialement, ceux et celles qui pensent que l'essence des nouveaux paradigmes qu'appelle la société des savoirs partagés et la démocratie participative réside dans une *éthique des processus,* utilisons pour traduire notre vision : **la cohérence.**

La cohérence entre le dire et le faire est pour nous ce qui permet de croire aux déclarations et de pardonner les erreurs qui, dans une approche de processus, deviennent des occasions d'apprendre, de tirer les leçons et de continuer à croître. Cette démarche, propre de la recherche-action, particulièrement adaptée pour traiter des questions de développement est celle qui nous habite dans ce document dont la prétention, plus qu'apporter des solutions pour une question aussi complexe que la diversité linguistique dans Internet, est de questionner les fausses évidences, d'apporter des points de vue provocateurs, pour ouvrir des pistes de réflexion et d'action qui sortent des sentiers battus et des jugements préconçus et puissent rendre compte de la complexité du sujet traité ; cela avec à la fois la modestie du chercheur qui tâtonne et la fermeté de la personne d'action qui s'est engagée sur le terrain.

La cohérence s'exprimera dans ce document de plusieurs manières :

— le choix de la langue maternelle, un droit élémentaire après tout, pour l'expression ;

— une volonté de laisser la diversité s'exprimer dans la sélection des personnes, compétentes sur le thème, invitées à s'exprimer. Nous avons essayé de couvrir aussi bien que possible les lieux géographiques, les cultures, les langues, les profils, les secteurs, les âges et les genres. A l'évidence, nous n'avons pas réussi complètement (nous regrettons, par exemple, que la place faite aux textes au féminin n'ait pas été plus grande) mais la cohérence s'exprime surtout dans l'authenticité de l'intention ;

— la décision de ne pas faire un texte « langue de bois » et de prendre le risque de la provocation, jamais gratuite, parfois gratifiante, toujours assise sur l'expérience de terrain et avec l'intention de déranger pour ouvrir les esprits, pas pour le plaisir de déranger.

Un approche structurée pour l'intégration des TIC et du développement humain

La « fracture numérique » est un concept qui est devenu très à la mode et a engendré beaucoup de réflexions et de réunions internationales. La vision plutôt consensuelle de la société civile (Pimienta, 2002, Communauté MISTICA, 2002) est qu'il ne faut pas se tromper de fracture et éviter la simplification qui consiste à tout mettre sur le dos de la technologie. Nous proposons ci-après une grille originale de lecture et analyse de l'utilisation des TIC pour le développement pour illustrer le fait que la résolution de la fracture numérique n'est pas, loin de là, une simple question d'accès à la technologie et que la question de la langue y joue également un rôle essentiel.

Le principe de la grille est d'identifier les obstacles successifs à surmonter pour permettre l'utilisation des TIC pour le développement humain. La grille sous-entend une progression dans l'énumération des obstacles, à partir des infrastructures vers l'infoculture en passant par l'infostructure. Il est probable que cette progression ne corresponde pas exactement à la réalité vécue par chaque personne ou groupe social et que l'ordre des facteurs dépende des contextes. Néan-

moins, pour des raisons pratiques et pédagogiques nous acceptons de simplifier cette réalité complexe de cette manière, en forme d'une série d'obstacles successifs à surmonter ou de niveau progressifs à atteindre.

Tableau 1. TIC pour développement : le long chemin semé d'obstacles de l'accès au développement humain

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
ACCES	*La possibilité pour une personne ou un groupe de personnes de détenir un moyen physique d'utiliser les TIC.*	
	Les obstacles à surmonter pour obtenir un accès sont multiples et peuvent également se présenter sous forme de couches progressives :	**– existence d'une infrastructure.** Les interfaces doivent permettre l'accès dans la langue maternelle de l'utilisateur et d'une manière adaptée à sa culture.
	– existence d'une infrastructure. *côté service :* fournisseurs d'accès TIC et fournisseurs d'accès aux réseaux de télécommunications dimensionnés de manière à servir la quantité d'utilisateurs avec des temps de réponse et des taux de congestion acceptables.	La question linguistique se retrouve, pour le matériel, dans les claviers des ordinateurs mais aussi, en ce qui concerne les logiciels, dans la gestion des caractères associés à une langue et qui doivent être codifiés pour le traitement informatique. Cependant la partie logiciel opérationnelle qui concerne les langues ne s'arrête pas à la codification :

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	côté utilisateurs : le matériel informatique requis pour cet accès avec les caractéristiques adéquates pour offrir des performances acceptables. Cela peut être fait de manière individuelle (station de travail personnelle) ou collective (télécentres ou kiosques Internet).	les programmes d'édition nécessitent, pour leur fonctionnement optimum dans une langue donnée, des corpus et dictionnaires pour la correction orthographique et de syntaxe. Une vision à long terme plus ambitieuse pourrait d'ailleurs considérer que les programmes de traduction automatique font partie de la couche opérationnelle (et non de la couche applicative). Un énorme travail reste à faire au niveau des programmes de traduction pour les étendre au-delà des langues dites dominantes. C'est un espace tout à fait indiqué pour le développement en logiciel libre mais malheureusement cet espace est pratiquement vide et un très grand effort de sensibilisation et d'encouragement doit encore être réalisé. Un aspect linguistique, qui est maintenant considéré par l'ICANN (Webopedia, 2005b), est celui des noms de domaine Internet dans toutes les langues (Wikipedia, 2005a)

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	– accès économique à l'infrastructure Que les prix pour l'utilisation de l'infrastructure soient accessibles aux utilisateurs. Il y a évidemment plusieurs éléments directs ou indirects dans l'équation de prix[3] et l'accès collectif et l'accès individuel présentent des paramètres différents. Il suffit de comparer, par exemple, l'ordre de grandeur des prix pour un accès ADSL (Webopedia, 2005a) (entre 10 et 50 $EU par mois) et les salaires dans la pyramide sociale pour découvrir que ceci représente plus d'un an de salaire pour une proportion importante de l'humanité (celle qui vit en dessous du seuil de pauvreté), une valeur de l'ordre d'un mois de salaire pour une autre proportion importante (une proportion notable des peuples des pays du Sud), une valeur de l'ordre de 10% du salaire	**– accès économique à l'infrastructure** Le principe de « l'accès universel » doit inclure la considération sur un prix d'accès cohérent avec le niveau économique des populations concernées.

3 Directs, comme le prix du poste d'accès, celui du fournisseur d'accès, dans certains cas, celui de la liaison téléphonique ou celui du fournisseur d'information, celui du logement d'un serveur ou d'un domaine Internet (car l'accès c'est aussi la production de contenus) ; ou indirects, comme les économies que permettent un accès (par exemple, téléphone IP ou facture de déplacement évitée) ou les coûts de maintenance des équipements et de formation du personnel.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	mensuel pour les classes moyennes des pays en développement et une valeur de l'ordre de 1% pour les classes moyennes des pays développés. La première fracture n'est finalement pas numérique elle est économique et sociale ... La résolution des deux premières couches mentionnées devrait[4] représenter ce qu'il est convenu d'appeler, par l'UIT et les organismes régulateurs des télécommunications (UIT, 2003), **« l'accès universel ».** Mais, s'il s'agit d'une condition nécessaire pour résoudre la fracture numérique, elle est très loin d'être une condition suffisante ...	**– accès économique à l'infrastructure** Le principe de « l'accès universel » doit inclure la considération sur un prix d'accès cohérent avec le niveau économique des populations concernées.
	– alphabétisation fonctionnelle Que la personne qui utilise l'infrastructure ait la capacité fonctionnelle de lire et écrire	**– alphabétisation fonctionnelle** Il n'est certes pas exclu de tirer parti de la composante multimédia

4 Nous écrivons «devrait» car trop souvent l'aspect économique est négligé dans les plans d'accès universel et le concept est compris comme une couverture physique totale des accès aux infrastructures, ce qui fait certainement l'affaire des vendeurs de matériel mais pas forcément celui des utilisateurs.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	dans sa langue. Il s'agit probablement de la seconde fracture qu'il faut résoudre quand on prétend offrir, par exemple, « Internet pour tous ».	des TIC pour adapter des interfaces permettant un certain nombre de possibilités aux personnes analphabètes. Cependant, il faut se rendre à l'évidence s'il s'agit d'accès à la connaissance et non simplement d'accès aux technologies, l'alphabétisation fonctionnelle est une priorité au dessus de l'accès technologique pour les populations non alphabétisées. Ici se pose aussi la question des langues seulement orales pour lesquelles l'espace numérique représente un handicap fatal sauf à réaliser l'effort d'inventer une forme écrite et codifiable.
	– numérisation de l'alphabet Que la langue maternelle de la personne qui utilisera l'infrastructure puisse se prêter à un traitement informatique. Pour cela il faut qu'elle existe sous forme écrite et que les caractères de son alphabet soient convenablement codifiés.	**– numérisation de l'alphabet** C'est aujourd'hui encore un obstacle majeur pour une très grande proportion des langues et cela doit représenter une priorité initiale majeure. Des efforts sont en cours dans le cadre de UNICODE (Wikepedia, 2005b) et

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	Ce n'est malheureusement pas le cas pour la majorité des langues encore en usage.	doivent être maintenus et amplifiés.
UTILISATION	*La possibilité de faire une utilisation efficiente (qui conduise à l'objectif fixé) et efficace (que le processus soit optimum dans l'utilisation du temps) des TIC.*	
	Pour cela il faut que la personne dispose d'un grand nombre de capacités de gestion des outils numériques et de compréhension des éléments conceptuels, méthodologiques et culturels associés à l'espace numérique. Il ne faut pas sous-estimer l'ampleur des capacités requises qui nous conduit au concept d'**alphabétisation numérique** (en anglais, « digital literacy »). L'apprentissage de l'espace numérique, qui ne doit pas être un simple entraînement à l'utilisation de certains programmes d'ordinateurs mais devrait inclure une vision holistique des considérations et impacts sociétaux[5] de l'utilisation des TIC pour le développement, est sans	**– alphabétisation numérique** L'effort formidable nécessaire pour une éducation numérique (apprentissage) doit impérativement être conçu et réalisé dans les langues maternelles des populations concernées et en tenant compte de leurs cultures. Il est important de noter que ce critère impératif s'applique également aux interfaces des applications de gouvernement électronique.

5 Impact politique, économique, social, culturel, linguistique, organisationnel, éthique, biologique, psychologique.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	aucun doute le nœud le plus difficile à résoudre, l'élément à la fois le plus important et le plus négligé, de l'effort à consentir pour surmonter la fracture numérique. Les trois piliers de la société de l'information à construire ne sont pas, contrairement à la croyance la plus répandue, les télécommunications, les équipements et les logiciels mais l'éthique de l'information, l'éducation et la participation ...	
APROPRIATION TECHNOLOGIQUE	*Quand la personne qui utilise est suffisamment habile pour que la technologie soit* ***transparente de son utilisation personnelle.***	
	Par exemple, une paire de lunettes, une technologie optique que l'on met sur son nez le matin et que l'on oublie totalement toute la journée ou encore, dans le champ des TIC, la personne qui fait usage de son téléphone sans que l'existence de ce média participe d'aucune manière du dialogue à distance. De manière évidente, pour les TIC, cette appropriation	Comment rendre transparente la technologie si son accès demande de passer par une langue autre que la langue maternelle ? Ce niveau renforce clairement les arguments avancés pour les niveaux précédents.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	demande des capacités plus sophistiquées qui concernent l'usage d'un PC et des applications informatiques qui interviennent dans les processus, ainsi, bien entendu, qu'une certaine expertise dans la recherche d'information ou la manière de communiquer par courrier électronique et de se comporter en communauté virtuelle. En plus d'une bonne éducation numérique, une **pratique** minimum est nécessaire pour atteindre ce stade.	
USAGE PORTEUR DE SENS	*La capacité de faire un usage des TIC qui possède une signification sociale pour la personne dans son contexte personnel, professionnel et communautaire.*	
	Il s'agit de dépasser l'utilisation ludique et de simple outil de communication interpersonnelle et d'orienter l'usage vers des fins de développement humain. C'est ici que doivent apparaître des capacités fondamentales pour ne pas être un simple consommateur et passer du côté de la	Le thème linguistique est essentiel dans ce niveau et renvoie à la possibilité et la motivation à produire des contenus et des communautés virtuelles localisées. Il pose aussi clairement la question du multilinguisme et de la nécessité de dispositif de passerelles entre les langues.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	production (de **contenus** par exemple) et de **création** (de c**ommunautés virtuelles** par exemple).	
APPRO-PRIATION SOCIALE	*Quand la personne qui utilise est suffisamment habile pour que la technologie soit* ***transparente de son utilisation sociale.***	
	Ce niveau évoque une compréhension lucide des impacts sociétaux de l'usage des TIC pour le développement et des **implications culturelles et éthiques** propres à cet usage (culture/éthique de réseau, culture/éthique de l'information et une connaissance des aspects méthodologiques liées aux usages productifs de développement). En plus d'une bonne éducation numérique une pratique orientée vers le développement est nécessaire pour atteindre ce stade.	Les aspects éthiques et culturels des réseaux ne sont pas entièrement neutres et doivent passer par le filtre du métissage (voire même d'une certaine forme de syncrétisme) avec les cultures et les éthiques locales. La langue étant un des vecteurs de transport des cultures n'est pas indifférente aux questions complexes et délicates qui se posent.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
«EMPOWERMENT»[6]	*Quand la personne et/ou la communauté est en mesure de* ***transformer sa réalité sociale*** *grâce à l'appropriation sociale des TIC à des fins de développement.*	
	Ici, il ne s'agit plus seulement des capacités elles-mêmes mais de leur **mise en pratique** aussi bien au niveau individuel que collectif. Cette mise en pratique demande évidemment l'application des valeurs associées à la culture de réseau et la culture de l'information : l'organisation en réseau, la propension au travail collaboratif, la transparence active, la participation proactive.	Clairement, plus on s'approche de la fin de la chaîne qui conduit de l'accès vers le développement plus il est clair que c'est l'aspect culturel qui prend de l'importance, sans perdre de vue qu'il est souvent impossible de le dissocier complètement de l'aspect linguistique. Que signifie « l'empowerment » et comment se manifeste-t-il dans chaque culture ?
INNOVATION SOCIALE	*Quand l'action de transformation de la réalité sociale est porteuse de* ***solutions originales*** *créées par la personne ou la communauté.*	
	Le nouveau paradigme de travail en réseau porte les germes de l'innovation, en particulier sociale (nouvelles formes d'organisation, réponses nouvelles à problèmes connus ...).	Que signifie « l'innovation » et comment se manifeste-t-elle dans chaque culture ?

6 Ce mot anglais rassemble à la fois les sens de recevoir et de prendre la capacité ainsi que la notion de prise de pouvoir à travers cette capacité.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
DEVELOP-PEMENT HUMAIN	*Quand les options de libertés individuelles et collectives s'ouvrent à la personne ou la communauté et peuvent s'exercer sous la forme de* ***« capacités »***.[7]	
	Il s'agit là de la finalité du processus, mais il doit rester clair que dans tout processus social on ne peut retrouver à la fin que ce que l'on a entretenu tout au long du processus depuis sa conception. Ainsi les options de libertés ne pourront s'épanouir que si la **participation** des personnes et des communautés a été une réalité dans tout le processus décrit.	***options de libertés*** *en forme de* ***« capacités »***. Que signifie « la participation » et comment se manifeste-t-elle dans chaque culture ? Une réelle « participation » dans des processus sociaux est-elle possible si une langue différente de la langue maternelle est imposée ?

Société de l'information : enjeux croisés pour les langues et cultures

Il est une discipline essentielle qui a vu le jour ces dernières années et pour laquelle l'UNESCO a apporté de nombreuses contributions : celle de l'éthique de l'information. Le croisement de cette discipline avec la question de la diversité culturelle et linguistique ouvre des perspectives et des réflexions tout à fait

7 « Le développement peut être vu comme un processus d'expansion des libertés réelles dont les personnes bénéficient. Considérer les libertés humaines (ou les capacités) diffère des visions plus étroites du développement, comme celles qui l'identifie avec la croissance du PNB, l'augmentation des revenus personnels, l'industrialisation, l'avance technologique ou la modernisation sociale. » (Sen, 2005).

pertinentes de notre débat. Un congrès a été consacré à ce thème en 2004[8] par l'ICIE (International Center for Information Ethics) et un livre sera publié à la fin de l'année 2005 avec les textes du Congrès qui sont autant de contributions pertinentes par rapport au sujet qui nous préoccupe (Capuro, 2005).

Parmi celles-ci, Charles Ess (2004) nous fait remarquer que contrairement aux hypothèses fréquentes selon lesquelles les TIC sont culturellement neutres, un grand nombre d'études ont pu montrer que les TIC, ayant leur origine dans les cultures occidentales, et plus spécialement nord-américaine, transportent et d'une certaine manière font la promotion de leurs valeurs culturelles et leurs préférences en termes de communication. Ceci est manifeste, selon Charles Ess, dans les multiples façons avec lesquelles ces valeurs et préférences rentrent en conflit avec celles des cultures qui reçoivent les technologies (plus particulièrement les cultures indigènes, asiatiques, latines et arabes). Ces conflits se traduisent dans les échecs parfois spectaculaires d'efforts de bonne volonté pour surmonter la pauvreté et la marginalisation (Postma, 2001). Ess va encore plus loin en soulignant le danger d'une « colonisation assistée par ordinateur » qui pourrait être le produit d'un plan naïf pour « brancher le monde » qui ne prête pas attention aux risques avérés d'affecter les valeurs et cultures locales par une implantation imprudente des TIC.

Charles Ess nous rassure cependant en indiquant que de tels conflits sont évitables, tout d'abord en adoptant une attitude consciente des enjeux culturels et il nous indique des pistes pour structurer un design des interactions homme-machine qui réponde à ce critère (Hall, 1976).

Si l'on convient que l'éducation numérique est l'un des enjeux essentiel du passage à une société de l'information inclusive, il devient également clair que cette éducation doit répondre à ce critère éthique fondamental de respect de la diversité culturelle et linguistique et donc éviter l'ethnocentrisme et la colonisation implicite par les technologies.

Il est une autre question essentielle et transversale parmi les enjeux de la société de l'information : celle d'un domaine publique de la connaissance qui

8 « Localizing the Internet: Ethical Issues in Intercultural Perspective », 4-6 October, 2004 – Karsluhe - http://icie.zkm.de/congress2004

devrait échapper à la logique du marché, et en dérivation celle des contenus et des logiciels ouverts. Cette question se croise également avec celle de la diversité linguistique dans la société de l'information.

José Antonio Millán (2001), le spécialiste espagnol du thème des langues et Internet, nous rappelle que nos langues restent l'interface le plus complet qui existe et que, sous la forme orale ou écrite, elles sont de plus en plus utilisées pour rentrer en relation avec une variété de programmes, comme par exemple dans le cas de la recherche de l'information. Le savoir linguistique qui est incorporé dans les programmes (correction automatique, fabrication de synthèse, transformation texte/voix, etc.) n'est pas forcément visible à l'utilisateur; pourtant son importance économique est énorme. Les ressources élémentaires qui ont servi de substrat aux programmes proviennent le plus souvent de recherches financées par des fonds publics. Pourtant, elles bénéficient souvent à des logiciels commerciaux dont la source n'est pas ouverte qui ne peuvent donc pas être améliorés et étendus (par exemple pour se préoccuper des variantes minoritaires des langues les plus répandues) ni servir de base pour que des langues minoritaires puissent créer leur propres logiciels. La démocratisation des logiciels linguistiques passe, selon Millán, par la libération (sous licences GPL ou similaires - Wikipedia, 2005c) des ressources linguistiques produites avec des fonds publics ou qui font simplement partie du domaine public.

En tout état de cause, les logiciels libres qui, par leur nature, devraient jouer un rôle particulièrement important dans le secteur linguistique n'y ont qu'une présence modeste et un effort de sensibilisation vers les communautés de développeurs est nécessaire.

Le thème des contenus ouverts nous conduit naturellement à considérer les changements requis par un système d'édition scientifique qui est considéré, par les acteurs de la société civile qui travaillent sur le thème de la société de l'information (Guédon, 1998) comme obsolète parce que représentant un frein au partage de la connaissance scientifique en particulier vers les pays du Sud. Ce système commence à être remis en question par des initiatives comme « Public Library Of Science » et la déclaration de Berlin sur l'accès ouvert au savoir dans les Sciences (ZIM, 2003). La diversité linguistique a tout à gagner d'une évolution du système d'édition scientifique vers des modèles tirant meilleur parti des TIC et basés sur les notions de contenus ouverts.

Derrière cette situation et un certain immobilisme des États concernés se cachent l'absence de politiques linguistiques et, en fait, la lacune critique à combler, comme le souligne José Antonio Millán, est celle d'une véritable politique des contenus numériques (qui inclut bien entendu une politique linguistique dans le monde numérique). A ce sujet, le rôle des organisations internationales comme l'UNESCO pourrait être de sensibiliser les États membres sur l'importance de politiques volontaristes de promotion du multilinguisme.

Les mesures et les indicateurs

Est-il raisonnable de définir et conduire des politiques linguistiques dans l'espace numérique sans détenir des indications amples, fiables et précises sur la situation de la langue et son évolution ?

Très paradoxalement, le monde des réseaux qui est né et s'est développé au sein de l'université a pendant longtemps abandonné la mesure de la place des langues à des entreprises de marketing répondant à des logiques distinctes de celle de la publication scientifique (et donc peu soucieuses de documenter leurs méthodes). Il en a résulté un désordre et une confusion sur l'état des langues dans l'Internet qui a pu faire le lit de la désinformation. Ainsi, alors que le nombre de locuteurs de langue anglaise qui utilise le réseau a pu passé de plus de 80%, l'année de la naissance du Web, a environ 35% aujourd'hui, les chiffres qui circulent dans les médias sur le pourcentage de pages Web en anglais continuent, contre toute évidence, à se situer de manière stable entre 70 et 80% !

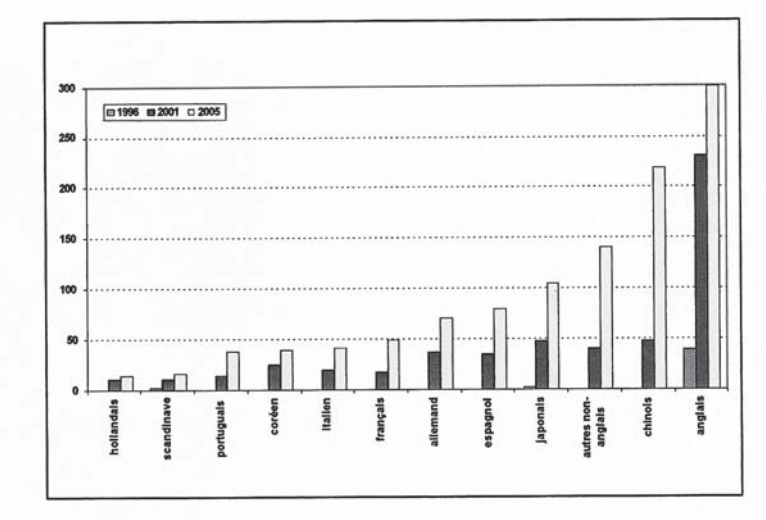
Il est urgent que l'académie reprenne son rôle dans cette affaire (ainsi que les institutions gouvernementales nationales et internationales) et les signes sont clairs que cette évolution est en cours, enfin ! Pour s'en rendre compte, il faut consulter les présentations en ligne de la réunion organisée par l'UNESCO (avec l'ACALAN et l'AIF) à Bamako sur le multilinguisme dans le cyberespace[9].

En attendant que cette évolution porte ses fruits (des indicateurs fiables, documentés et mis à jour à la vitesse de l'évolution du média), obtenir une perspective sur la situation et les tendances est extrêmement difficile.

9 http://portal.UNESCO.org/ci/en/ev.php-URL_ID=19088&URL_DO=DO_TOPIC&URL_SECTION=-465.html ou http://www.UNESCO.org/webworld/multilingualism.

I - En ce qui concerne les données sur la proportion des internautes dans chaque langue, une source a réussi à s'imposer depuis plusieurs années. Global Reach fournit avec une grande régularité des chiffres qui reposent, certes, sur des sources multiples et non cohérentes sur le plan méthodologique, mais au moins elles sont connues (Figure 1). Les chiffres ne sont pas d'une totale fiabilité mais ils ont le mérite d'exister et d'être maintenu à jour avec fréquence; si on leur accorde une confiance relative (plus ou moins 20% d'erreur), ils permettent d'obtenir une perspective raisonnable de l'évolution de la population d'internautes en termes de langue.

Figure 1 : Nombre d'internautes par langue d'utilisation



Source : Global Reach 2005.
(http://global-reach.biz/globalstats/index.php3)

II - Pour la place des langues sur le Web il y a un certain nombre d'approches qui cohabitent :

1) L'une consiste à extrapoler les chiffres des moteurs de recherche par langue. C'est la plus facile et elle donne des ordres de grandeur acceptable mais pas de chiffre assez fiable pour maintenir une veille sérieuse, étant donné les faiblesses des algorithmes de reconnaissance des langues et les comportements erratiques des moteurs sur les totalisations.

2) Une autre a été lancée par une des premières études sur le sujet, qu'Alis Technologies a réalisée en juin 1997, avec le soutien de l'Internet Society et dont la méthode a été reprise par d'autres, en particulier l'étude de l'OCLC (« Online Computer Library Center ») qui semble être la référence sur laquelle s'appuie de nombreux auteurs et médias pour continuer à proposer une valeur de plus de 70% pour les pages Web en anglais (O'Neill, 2003). La méthode consiste à créer un échantillon de quelques milliers de sites Web par le jeu du hasard sur les adresses IP (Wikipedia, 2005d), à appliquer les moteurs de reconnaissance des langues sur cet ensemble de site et à en généraliser les résultats.

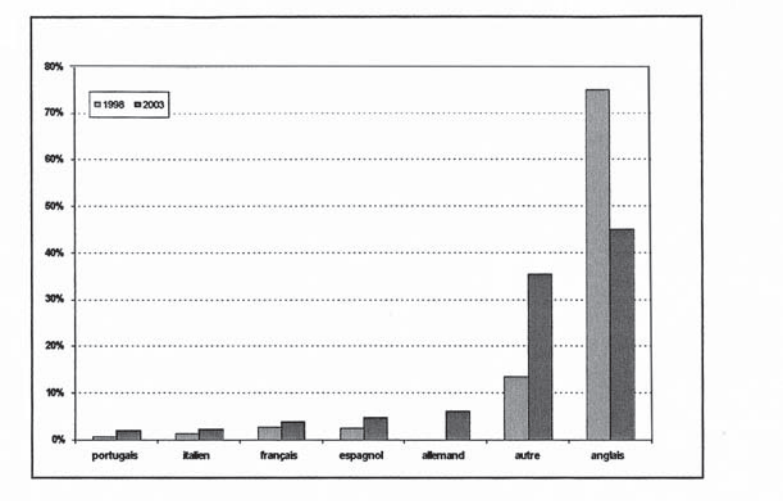
Elle partage avec la première approche la limitation des algorithmes de reconnaissance des langues, quoique l'on puisse espérer que des progrès importants aient été réalisés depuis 1997 et que dans le futur les techniques augmenteront de manière décisive la fiabilité des résultats.

La seconde limitation nous préoccupe beaucoup plus car elle est d'ordre statistique. Le traitement mathématique prévu pour une variable aléatoire (comme c'est le cas de l'échantillon de sites Web pris au hasard sur lequel est appliqué la reconnaissance des langues) est d'en étudier la distribution statistique pour en extraire la moyenne, la variance et en déduire l'intervalle de confiance. Une seule prise faite au hasard ne peut fournir aucun résultat crédible (que représentent 8000 sites Web en face des 8000 millions de pages indexées par Google ?). A travers le peu de documentation publié il semble pourtant que les chiffres soient produits de cette manière par OCLC.

3) Il existe une ample catégorie où des chiffres sont avancés et aucune méthode n'est révélée. Il est impossible de valider les résultats. C'était le cas de l'étude de Inktomi en 2001 qui était lancée avec un grand fracas de marketing et qui en plus comportait des erreurs grossières (elle annonçait des pourcentages globaux de pages Web dans un nombre limité de langues et le total de ces pourcentages était de 100% ...) !

4) Enfin la dernière catégorie regroupe quelques rares méthodes qui sont documentées comme l'approche très originale des chercheurs de Xerox en 2001 (Grefenstette & Nioche, 2001), parmi celles-ci, l'approche que FUNREDES et l'Union Latine ont utilisée depuis 1996 (voir Figure 2).

Figure 2 : Proportion de pages Web composées dans une langue donnée



Source: FUNREDES 2003, http://funredes.org/lc

Le principe de la méthode est le suivant : les moteurs de recherche permettent d'obtenir la valeur du nombre d'occurrence d'un mot donné dans l'espace recherché (pages Web ou groupes de discussion, par exemple). Un échantillon de mots-concepts dans chacune des langues étudiées a été construit avec un souci de fournir la meilleure équivalence sémantique et syntaxique entre les mots-concepts. Les valeurs d'apparition de chaque mot mesurées par les moteurs de recherche sont compilées pour chaque concept dans chaque langue. Ces valeurs sont traitées comme une variable aléatoire dont la distribution mathématique est étudiée avec les outils traditionnels de la statistique (moyenne, variance, intervalles de confiance, loi de Fisher) et le résultat consiste, pour chaque langue étudiée, en une estimation du poids de sa présence relativement à l'anglais qui est pris comme langue de référence. Cette estimation est de plus validée quantitativement par les instruments statistiques (intervalle de confiance). La répétition de la méthode à intervalles successifs permet d'obtenir une vision de l'évolution de la présence des langues dans les espaces considérés et en même temps d'apprécier la valeur de la méthode qui a donné des résultats cohérents tout au long des mesures.

Si la méthode publiée intégralement depuis son origine n'a pas reçu à ce jour d'arguments l'invalidant, elle présente un certain nombre de limitations :

— Elle fournit une valeur du pourcentage de pages Web dans une des langues travaillées (allemand, espagnol, français, italien, portugais et roumain) par rapport à l'anglais mais pas de valeur absolue. Pour l'obtenir, il faut établir une estimation du poids absolu de l'anglais à partir de recoupements de plus en plus difficiles et incertains avec la multiplication des langues ;

— Il est difficile (sur le plan linguistique) et coûteux de rajouter une nouvelle langue ;

— Elle donne une valeur qui correspond à l'espace des pages indexées par les moteurs et ne prend pas en compte le Web invisible (Bergman, 2001). Mais quelle « existence » ont réellement les pages non indexées ? ;

— Mais surtout elle est très dépendante des possibilités de comptage fiable qu'offrent les moteurs de recherche[10], ce qui à terme risque de la disqualifier puisque les moteurs prennent de plus en plus de liberté avec le traitement de la recherche par mot[11].

Du côté des avantages, la méthode a permis de maintenir un suivi d'observation cohérent sur une longue période, d'examiner d'autres espaces que le Web[12] et surtout, en bénéficiant des techniques de recherche par pays et par domaine, de produire une série d'indicateurs originaux et très signifiants (Pimienta, 2001).

Perspectives pour de nouvelles approches

Le projet maintenant avancé de l'Observatoire des Langues (voir l'article de Yoshiki Mikami, plus loin) porte de nombreux espoirs pour occuper ce vide et

10 La majeure partie du travail pour les mesures consiste aujourd'hui à vérifier le comportement des moteurs, sélectionner les plus fiables et compenser leurs comportements erratiques, en particulier dans le traitement des signes diacritiques.

11 Il est probable que d'ici peu les moteurs offriront des résultats comportant des textes avec la traduction des mots de recherche dans d'autres langues.

12 Elle a également permis une première approximation certes grossière mais intéressante sur le plan des évolutions de la présence des cultures dans l'Internet.

apporter les réponses dont les politiciens ont besoin pour établir leur choix et en mesurer l'impact.

Notre expérience de terrain nous fait penser qu'une approche très prometteuse et qui ne semble pas encore exploitée consisterait en une méthode similaire à celle qu'utilise Alexa pour dresser le hit parade des sites visités et pour apporter de précieux renseignements. Alexa compile les données de comportement d'un grand nombre d'utilisateurs qui ont accepté le chargement d'un programme espion dans leur ordinateur et en tire des statistiques extrêmement riches. Sur le même principe, il est possible d'imaginer un programme qui soit capable de mesurer les langues utilisées dans divers contextes pertinents pour les indicateurs comme : langues de lecture et écriture des courriels, langues des sites visités, etc.

Bibliographie

Bergman, M.K. 2001. The Deep Web: Surfacing Hidden Value. *Bright Planet – Deep Web.* http://www.brightplanet.com/technology/deepweb.asp

Capurro, R. & al. (Eds.) 2005. Localizing the Internet. Ethical Issues in Intercultural Perspective. *Schriftenreihe des ICIE* Bd. 4, München: Fink Verlag.

Communauté MISTICA. 2002. « Travailler l'Internet avec une vision sociale ». http://funredes.org/mistica/francais/cyberotheque/thematique/fra_doc_olist2.html

Ess, C. 2004. Moral Imperatives for Life in an Intercultural Global Village in The Internet and Our Moral Lives, ed. R. Cavalier, State University of New York Press, Albany. pp. 161-193.

Ess, C. 2005. Can the Local Reshape the Global? Ethical Imperatives for Human Intercultural Communication Online, in (Capurro, 2005).

Ess, C. 2006. From Computer-Mediated Colonization to Culturally-Aware ICT Usage and Design, In P. Zaphiris and S. Kurniawan (eds.), *Human Computer Interaction Research in Web Design and Evaluation.* Hershey, PA: Idea Publishing.

Ess, C. & Fay S. 2005. Introduction: Culture and Computer-Mediated Communication – Toward New Understandings, *Journal of Computer-Mediated Communication Vol. 11, No. 1.* < http://jcmc.indiana.edu/>

Grefenstette, G. & Nioche, J. 2001. Estimation of English and non-English Language. Use on the WWW. Xerox Research Centre Europe, Meylan.

Guédon, J.C. 1998. « La bibliothèque virtuelle : une antinomie ? » conférence prononcée à la National Library of Medicine. Washington. http://sophia.univ-lyon2.fr/francophonie/doc/nlm-fr.html

Hall, E.T. 1976. Beyond Culture. Anchor Books, New York.

Millán, J.A. "How much is a language worth: A Quantification of the Digital Industry for the Spanish Language". *Language Diversity in the Information Society International Colloquium.* Paris, France. http://jamillan.com/worth.htm

O'Neill & al. 2003. Trends in the Evolution of the Public Web: 1998 – 2002 http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

Pimienta, D. 2002. « La fracture numérique, un concept boiteux. » *Communauté Virtuelle MISTICA.* http://funredes.org/mistica/francais/cyberotheque/thematique/fra_doc_wsis1.html

Pimienta, D. & Lamey B. 2001. "Lengua Española y Culturas Hispanicas en la Internet: Comparación con el inglés y el francés." *II Congreso Internacional de la Lengua.* Valladolid. http://www.funredes.org/LC/L5/valladolid.html

Postma, L. 2001. "A Theoretical Argumentation and Evaluation of South African Learners". Orientation towards and Perceptions of the Empowering Use of Information. *New Media and Society.* Vol. 3 No. 3. pp. 315-28.

Sen, A. 2005. *Human Development and Capability Association.* http://www.fas.harvard.edu/~freedoms/

UIT. Union Internationale des télécommunications. 2003. *Competitive Markets Required to Bridge Digital Divide : Regulators map 'Universal Access' route to Information and Communication Technology.* http://www.itu.int/newsarchive/press_releases/2003/33.html

UNESCO. 2000. "Infoethics". *UNESCO WebWorld News.* http://www.UNESCO.org/webworld/news/infoethics.shtm

UNESCO. 2005. Multilinguisme pour la diversité culturelle et la participation de tous dans le cyberespace. http://portal.unesco.org/ci/fr/ev.php-URL_ID=17688&URL_DO=DO_TOPIC&URL_SECTION=201.html

ZIM. 2003. "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities". *Conference on Open Access to Knowledge in the Sciences and Humanities.* Berlin. http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

Glossaire

Webopedia.2005a. *ADSL.* http://www.webopedia.com/TERM/A/ADSL.html

Webopedia. 2005b. *ICANN.* http://www.webopedia.com/TERM/I/icann.html

Wikipedia. 2005a. *Internationalized Domain Name.* http://en.wikipedia.org/wiki/IDNA

Wikipedia. 2005b. *Unicode.* http://en.wikipedia.org/wiki/Unicode

Wikipedia. 2005c. *GNU General Public License.* http://en.wikipedia.org/wiki/GNU_General_Public_License

Wikipedia. 2005d. *IP Address.* http://en.wikipedia.org/wiki/IP_address

b. Le contexte politique et juridique

Daniel Prado, Union Latine

En règle générale, les grandes langues occidentales connaissent un recul important dans la communication scientifique et technique au profit de l'anglais. A l'exception de certaines langues de moindre diffusion qui ont su reprendre une place ces dernières années, les grandes langues d'origine européenne comme l'allemand, l'espagnol, le français, l'italien, le portugais, le russe et les langues scandinaves sont touchées (Hamel, 2002).

Parmi ces langues européennes, les langues néolatines sont particulièrement touchées, que ce soit dans l'édition spécialisée, dans les congrès scientifiques, dans les organisations internationales, dans les médias ou dans l'enseignement, etc.

En novembre 2002, le premier Congrès international sur la place des langues néolatines dans la communication spécialisée (UNILAT, 2002a) réunissait des spécialistes des politiques linguistiques de trois espaces linguistiques : la francophonie, la lusophonie et l'hispanophonie.

Lors de ce congrès, des statistiques et des constatations ont montré la perte vertigineuse de vitalité des langues d'origine néolatines dans plusieurs secteurs touchant aux sciences et techniques. Malgré le fait d'être langues officielles dans plus d'un quart des pays de la planète (27,53 %) selon Calvet (2002) et d'être parlées par près d'un milliard de locuteurs, des langues comme le français, l'espagnol, le portugais, l'italien, le roumain, le catalan et une vingtaine d'autre langues de moindre diffusion, ne produisent qu'un dixième des publications scientifiques par rapport à l'anglais, en suivant les bases de données internationales les plus importantes[13]. En effet, selon ce que nous rappelle Hamel, l'anglais représenterait entre 80 et 90 % des publications scientifiques en sciences naturelles et entre 74 et 82 % en sciences humaines et sociales tandis que les trois langues néolatines les mieux

13 Il est souvent considéré que les journaux scientifiques en langue anglaise sont surreprésentés dans ces bases de donnés internationales, et qu'en contrepartie les journaux des pays au dehors de ceux de l'OCDE sont sous-représentés [UIS].

représentées proposeraient 12 % des publications en sciences sociales et 18 % en sciences humaines. Mais Hamel nuance ses propos, rappelant que ces statistiques proviennent des bases de données des publications scientifiques et que l'édition de livres est tout aussi vigoureuse que les revues scientifiques. Il est intéressant de noter que le monde de l'édition des pays latins se porte bien, avec 18,9 % de la production mondiale (Rousseau, 2002), mais c'est la littérature qui est concernée majoritairement par ce chiffre (Leáñez Aristimuño, 2002).

Bien entendu, par comparaison avec la plupart des langues de la planète, la situation des langues néolatines dans la diffusion de connaissances n'est pas la pire. En effet, pour 100 pages Web mesurables en anglais, on trouve près de 38 pages (UNILAT, 2005) en langues latines[14] ; le français est la deuxième langue d'usage international; l'espagnol prend une confortable troisième place dans cet univers et son enseignement croît dans le monde entier ; le portugais a une belle implantation démographique et intercontinentale et l'italien reste une langue de prestige culturel malgré sa faible démographie et son cantonnement géographique (Italie, Suisse et Saint-Marin).

Mais, il ne faut pas oublier que l'anglais, avec deux fois et demie moins de locuteurs que l'ensemble des locuteurs latins a deux fois et demie plus de pages Web que toutes les langues latines réunies. Il ne faut pas non plus oublier que les publications scientifiques éditées en anglais représentent plus des deux tiers de l'ensemble mondial, tandis que toutes les langues latines réunies ne représenteraient qu'environ une publication scientifique sur dix.

Loin de notre étude l'intention d'ignorer la situation de déclin scientifique ou technique que vivent d'autres langues comme celles du Nord de l'Europe (langues scandinaves, notamment) pour lesquelles des pans de vocabulaire scientifique disparaissent du fait du monolinguisme anglais que pratiquent les spécialistes de certaines disciplines (Nilsson, 2005). Également loin de nous l'intention de vouloir dramatiser la situation des langues européennes lorsque, comme nous le rappelle Leáñez, 98 % des langues de cette planète ne disposent même pas de certains vocabulaires spécialisés de base, qu'ils soient administratifs, scientifiques, techniques, juridiques ou commerciaux. Il s'agit de tirer la sonnette d'alarme sur

14 L'étude a été réalisée sur les cinq premières langues néolatines en nombre de locuteurs, soit espagnol, français, italien, portugais et roumain.

une situation inquiétante qui n'épargne pratiquement aucune langue en dehors de l'anglais.

Pour revenir sur la présence des langues sur l'Internet, même si les statistiques Funredes/Union Latine nous montrent qu'en 2003 près de 14 % des pages Web étaient éditées en au moins une langue latine, près de 45 % le sont en anglais. Même l'allemand, avec dix fois moins de locuteurs, avait à peine deux fois moins de pages que l'ensemble des langues romanes. Mais ce qui est le plus inquiétant sur la place des langues latines sur l'Internet ce sont les données non publiées, l'Internet invisible, les Intranet, les bases de données, les listes de diffusion, les forums, etc. Nous ne disposons pas de statistiques sur ce sujet, mais une simple pratique quotidienne montre la prédominance majeure de la langue anglaise dès qu'une discussion technique internationale s'engage dans un forum électronique ou dès qu'une base de données scientifiques a une portée internationale ou même dans une conversation de jeunes sur leur star préférée. Ce phénomène s'expliquait bien aux débuts des réseaux télématiques, car ils s'adressaient à un public de chercheurs internationaux, et il est inutile de rappeler que l'anglais est perçu dans le milieu scientifique comme la langue principale de communication. Mais ce qui est regrettable, c'est que ce modèle n'a pas su évoluer, excluant de ce fait des populations ou des collectifs moins habitués à manier la langue anglaise.

Leáñez nous rappelait qu'« une langue qui a peu de valeur est peu utilisée et une langue peu utilisée a peu de valeur » [traduction libre] et affirmait que si nos langues ne couvrent pas nos besoins, nous apprenons et en enseignons une autre.

Face à cette affirmation, le plan d'action de l'UNESCO (2005) pour le SMSI tombe à point nommé. En effet, dans le premier chapitre, l'une de ses lignes d'action concerne la diversité culturelle et linguistique et il y est recommandé « d'élaborer des politiques qui encouragent le respect, la préservation, la promotion et le renforcement de la diversité culturelle et linguistique et du patrimoine culturel dans le contexte de la société de l'information … ». A l'heure actuelle, aucun Etat latin ne s'est doté d'une politique qui permette un usage des langues latines dans leur plénitude et notamment dans la Société de la Connaissance et du Partage du Savoir.

En effet, en matière de politiques linguistiques, les pays latins (sauf à de rares exceptions) sont trop concentrés sur les aspects exclusivement administratifs

d'une part, sur la protection des langues endogènes, d'autre part, et plus rarement, sur la protection du consommateur. Ne créant pas les dispositifs de contrôle nécessaires et ne se donnant pas les moyens pour mettre en pratique ce que les textes législatifs prônent, ils ne disposent pas des ressources suffisantes pour développer leur langue et laissent vacante une place vite reprise par l'anglais, notamment dans le discours scientifique, dans la documentation technique, dans l'enseignement supérieur, dans l'Internet, etc.

À l'exception du Québec, de la Catalogne et de la France, aucun organisme d'État ne prend en charge, dans les pays latins, toutes les composantes permettant une politique globale de développement, d'enrichissement, de modernisation et de diffusion d'une langue. En Belgique, en Suisse, en Espagne, au Portugal des institutions existent mais ne s'occupent que partiellement de cette tâche. Et encore, dans les régions ou pays les plus développés en matière de politiques linguistiques, une politique de soutien au multilinguisme numérique fait défaut. Trop souvent, ce sont des associations de droit privé (ayant peu de moyens) ou des organismes intergouvernementaux (n'ayant pas un mandat clair pour ce faire) qui doivent venir compléter ces actions.

Heureusement, beaucoup de langues minoritaires ou « minorisées », contrairement à ce qui se passe avec les grandes langues, prennent une place dans la communication spécialisée qu'elles ne connaissaient pas auparavant. C'est notamment le cas du catalan, mais aussi du galicien, du basque, voire du sarde et autres. Cependant il reste encore beaucoup à faire et il n'est pas dit qu'elles pourront couvrir toutes les sphères nécessaires à l'épanouissement de leurs populations.

Reste l'épine principale de l'accès à l'information lorsqu'elle a été produite dans une langue que nous ne maîtrisons pas. Les traductions, nous le savons, sont chères. Pour certains processus (la traduction d'un appel d'offre d'une OIG, par exemple) la traduction est lente.

La traduction automatique, qui, rappelons-le, ne remplacera jamais la traduction humaine, (simplement l'aidera à être plus performante, rapide et abordable) est l'instrument indispensable à une transformation nécessaire du monde de l'édition numérique et papier.

Aucun système actuel ne permet des traductions satisfaisantes pour les couples de langues les plus usitées. Toute traduction pour ces couples a besoin

d'une révision. Mais le plus grave, c'est que la plupart des systèmes de traduction automatiques ou de TAO ne prennent en charge qu'un nombre dérisoire de couples de langues.[15]

La qualité des systèmes existants doit s'améliorer et voyant leur évolution, ceci se fera sans doute, mais rien ne laisse présager que ce pourcentage fatidique de moins de 1 % de couples de langue puisse être dépassé prochainement. Des initiatives volontaristes doivent montrer le chemin de la traduction entre des langues qui ne présentent aucun débouché pouvant intéresser les compagnies commerciales. L'Union latine a initié certaines démarches dans ce sens[16], l'Université des Nations Unies également. Il est à attendre que d'autres puissent également se produire pour les langues les moins favorisées.

Que faire alors pour parvenir à un monde numérique multilingue ? La récente discussion franco-française reprise par la presse internationale sur un « Google » européen a suscité certaines idées (Millán, 2005) et l'UNESCO insiste sur le rôle des bibliothèques et des collections. Une idée pourrait être celle de mettre en place de vastes programmes d'informatisation des collections, faisant appel autant aux Etats qu'aux OIG ou ONG ou bien aux fournisseurs de services Internet privés, mais seulement ceux qui pourraient s'engager à respecter une charte éthique dans l'utilisation de cette information. Il faut évidemment empêcher l'appropriation à des fins commerciales de l'information numérisée ou exigeant des droits de diffusion ou d'exploitation de cette information. L'objectif est de diffuser librement et gratuitement les contenus numérisés, seul moyen de garantir une véritable diversité linguistique.

L'Internet nous montre dans son quotidien, de façon spontanée, de nouvelles voies : des organes de presse indépendants et autonomes, des blogues, des initiatives citoyennes voient le jour de façon quotidienne et elles démontrent que d'autres voies aux monopoles monolingues existent. Il faudrait peut-être mieux observer ces initiatives alternatives, les soutenir et s'en inspirer.

15 En effet, l'on recense bien moins de 100 langues traitées par des systèmes de traduction automatique ou de TAO sur près de 6000 langues existantes.

16 Notamment en introduisant la langue roumaine dans le projet Atamiri (http://lux0.atamiri.cc/forum/init.do).

En règle générale, les Etats latins sont en retard par rapport aux enjeux que représente la présence de leurs langues dans la société numérique. En ce sens, plusieurs actions s'imposent : la création d'une politique volontariste de numérisation des fonds et des catalogues existant, à l'heure actuelle seulement sur papier et d'une politique constante de production scientifique en langue nationale ou, à défaut, de traduction de cette production si elle est réalisée en anglais, et de son immédiate diffusion sur l'Internet; la mise en place d'une charte de respect du droit des citoyens de s'informer dans leur langue et donc une obligation respectée de multilinguisme sur les sites des organisations internationales, des compagnies internationales et bien entendu, une obligation de diffusion en langue locale pour les corporations nationales; et finalement, une proposition de dynamisation des projets de traduction automatique, notamment pour les couples de langues inexistants.

L'Union latine prépare une deuxième rencontre sur la place des langues latines dans la communication spécialisée pour pouvoir mettre en pratique les recommandations que la première rencontre avait proposées (UNILAT, 2002b). Elles prévoient des mécanismes de consultation, de suivi, de statistiques, d'action visant à encourager l'édition en langues latines, à favoriser la recherche en langues latines et à développer des outils linguistiques performants. Cette rencontre devrait avoir lieu en 2006 en Espagne, en étroite relation avec les institutions des Trois Espaces Linguistiques et il est à espérer que des solutions aux problèmes soulevés seront trouvées.

Bibliographie

Calvet, L.J. 2002. Le marché aux langues. Plon, Paris.

Hamel, R.E. 2002. "El español como lengua de las ciencias frente a la globalización del inglés. Diagnóstico y propuestas de acción para una política iberoamericana del lenguaje en las ciencias" au *Congrès international sur les langues néolatines dans la communication spécialisée.* Mexique. http://unilat.org/dtil/cong_com_esp/comunicaciones_es/hamel.htm#a

Leáñez Aristimuño, C. 2002. "Español, francés, portugués: ¿equipamiento o merma?" au *Congrès international sur les langues néolatines dans la communication spécialisée.* Mexique. http://unilat.org/dtil/cong_com_esp/comunicaciones_es/leanez.htm#a

Millán, J.A. 2005. « A quoi bon un projet européen concurrent ? ». *Courrier International.* http://www.courrierint.com/article.asp?obj_id=51004&provenance=hebdo

Nilsson, H. 2005. « Perte de domaine, perte de fonctionnalité : indicateurs et enjeux ». *Au Lexipraxi.* http://www.ailf.asso.fr/presentation.htm

Rousseau, L.-J-. 2002. « Le français dans la communication scientifique et technique » au Congrès international sur les langues néolatines dans la communication spécialisée Mexique. http://unilat.org/dtil/cong_com_esp/comunicaciones_es/rousseau.htm#a

UNESCO. 2005. *Plan d'action du SMSI.* http://portal.UNESCO.org/ci/fr/ev.php-URL_ID=15897&URL_DO=DO_TOPIC&URL_SECTION=201.html

UNILAT. 2002a. *Congrès international sur les langues néolatines dans la communication spécialisée.* http://www.unilat.org/dtil/cong_com_esp/es/index.htm

UNILAT. 2002b. *Recommandations. Congrès international sur les langues néolatines dans la communication spécialisée.* http://www.unilat.org/dtil/cong_com_esp/es/index.htm

UNILAT. 2005. *Etude sur La place des langues latines sur l'Internet* (http://www.unilat.org/dtil/LI/2003_2005.htm)

3.

Diversité linguistique sur Internet : examen des biais linguistiques

John Paolillo, School of Informatics, Indiana University

Plus de deux décennies après l'arrivée d'Internet dans le monde anglophone, la représentation des différentes langues sur Internet reste largement biaisée en faveur de l'anglais. Cette langue reste en effet la plus répandue sur Internet, alors que certaines langues très parlées sont peu ou pas représentées. Dans quelle mesure une telle situation constitue-t-elle un biais en faveur de l'anglais et au détriment des autres langues ? Cet article[17] aborde cette question en présentant le cadre éthique de Friedman et Nissenbaum (1997) afin d'évaluer le biais dans les systèmes informatiques, lié au statut sur Internet des langues parlées à travers le monde. Ce cadre conceptuel nous aide à interpréter les causes probables ainsi que les solutions de ce biais éventuel. Les revendications actuelles relativement au statut linguistique international sur Internet sont aussi présentées et reformulées dans l'optique de leur signification dans ce cadre, nous amenant à examiner non seulement la distribution et l'usage des langues sur Internet, mais aussi des institutions sociales guidant la gouvernance et le développement d'Internet pouvant mener à ce que Friedman et Nissenbaum appellent le « biais émergent ». Enfin, nous examinons les enjeux liés au biais linguistique dans les systèmes techniques d'Internet.

17 Ont collaboré à ce rapport : ELIJAH WRIGHT et HONG ZHANG, Indiana University, Baskaran, S., G. V., Ramanan, S. V., Rameshkumar, S., SHOBA NAIR, L., VINOSHBABU JAMES, VISWANATHAN, S. Anna University, Chennai, Inde. On peut accéder à la version complète du rapport sur le site: http://ella.slis.indiana.edu/~paolillo/paolillo.diversity.pdf.

Biais, multiculturalisme et systèmes informatiques

La « fracture numérique », c'est-à-dire la distribution inégale de l'accès aux sources et aux services d'information numérique, s'avère l'un des principaux enjeux politiques à notre époque d'information numérique. Les gouvernements, agences internationales, groupes de citoyens, sociétés et autres cherchent tous à profiter des promesses de moindres coûts et d'accès instantané à l'information en migrant plusieurs de leurs systèmes de communications sur des ordinateurs en réseaux. Mais si les barrières sociales traditionnelles, tels que le statut socio-économique, l'éducation, l'origine ethnique, le genre, etc. entravent l'accès à l'information numérique, les politiques doivent alors être formulées en vue d'égaliser l'accès pour que ces avantages se concrétisent.

Les questions relatives au statut linguistique international en ligne peuvent s'exprimer sous forme de fracture numérique. Dans certaines langues, le contenu informatique est déjà facilement accessible en grand nombre. Les internautes qui parlent, lisent et écrivent ces langues ont beaucoup moins de difficultés à accéder et à partager de l'information utile que ceux qui parlent des langues moins bien représentées. Une telle situation soulève évidemment la question à savoir si les systèmes d'information numérique, leur configuration, ou leur usage constituent une forme de biais envers les langues moins bien représentées. La différence linguistique est-elle devenue un obstacle à l'accès à l'information, constituant un avantage injuste pour certains et un désavantage pour d'autres ? Par définition, les questions de cette nature sont fondamentalement d'ordre éthique et moral, et le cadre conceptuel doit en tenir compte.

UNESCO et diversité culturelle

En 2001, les Etats membres de l'UNESCO ont adopté une Déclaration universelle sur la diversité culturelle.[18] L'article 6 « Vers une diversité culturelle accessible à tous », énonce :

18 http://unesdoc.UNESCO.org/images/0012/001271/127160m.pdf.

> Tout en assurant la libre circulation des idées par le mot et par l'image, il faut veiller à ce que toutes les cultures puissent s'exprimer et se faire connaître. La liberté d'expression, le pluralisme des médias, le multilinguisme, l'égalité d'accès aux expressions artistiques, au savoir scientifique et technologique - y compris sous la forme numérique - et la possibilité, pour toutes les cultures, d'être présentes dans les moyens d'expression et de diffusion, sont les garants de la diversité culturelle.

En ce sens, l'UNESCO favorise clairement l'accès égal à l'information numérique, autant à la production qu'à l'utilisation, pour tous les groupes linguistiques et culturels. La déclaration développe cette position en énumérant plusieurs orientations concrètes pour sa mise en œuvre. Trois aspects concernent directement les questions liées aux moyens numériques et à la technologie de l'information.

> 9. encourager « l'alphabétisation numérique » et accroître la maîtrise des nouvelles technologies de l'information et de la communication, qui doivent être considérées aussi bien comme des disciplines d'enseignement que comme des outils pédagogiques susceptibles de renforcer l'efficacité des services éducatifs ;
>
> 10. promouvoir la diversité linguistique dans l'espace numérique et encourager l'accès universel, à travers les réseaux mondiaux, à toutes les informations qui relèvent du domaine public ;
>
> 11. lutter contre la fracture numérique – en étroite coopération avec les institutions compétentes du système des Nations Unies - en favorisant l'accès des pays en développement aux nouvelles technologies, en les aidant à maîtriser les technologies de l'information et en facilitant à la fois la circulation numérique des produits culturels endogènes et l'accès de ces pays aux ressources numériques d'ordre éducatif, culturel et scientifique, disponibles à l'échelle mondiale (UNESCO, 2001, p.8).

Ces principes et orientations concrètes déterminent les valeurs permettant d'évaluer les attributs de la société de l'information en termes éthiques, ainsi que ses objectifs de développement. Ils ne fournissent cependant pas un aperçu suffisant des causes possibles de tout biais pouvant survenir. Et en ce sens, il s'avère difficile de faire des recommandations d'action appropriées dans des cas précis.

À titre d'exemple, les Maori de la Nouvelle-Zélande n'ont pas bien accepté les bibliothèques numériques. Plutôt qu'un simple problème d'alphabétisation numérique, une étude attentive a révélé que plusieurs enjeux d'ordre culturel nuisent au succès de cette ressource, notamment le fait que la bibliothèque est une forme d'institution « Pakeha » (Européen de l'Ouest de race blanche) supposant un accès à l'information méconnu dans la culture Maori (Dunker, 2002). La grande disponibilité de l'information, traditionnellement protégée dans la culture Maori (notamment l'information généalogique) constitue un aspect essentiel du problème pour ce peuple. Par définition, les bibliothèques permettent un libre accès à l'information, peu importe le contenu, et ignorent donc cette valeur. C'est pourquoi il faut revoir le modèle d'accès à l'information aux bibliothèques numériques avant qu'une telle institution ne soit mise en place et acceptée chez les Maori.[19]

Un cadre éthique

Friedman et Nissenbaum (1995, 1997) fournissent un cadre conceptuel utile pour analyser le biais dans les systèmes informatiques, en aidant à concentrer l'attention sur les causes du biais. Ces auteurs identifient trois principales catégories de biais : préexistant, technique et émergent. Le biais préexistant est ancré dans les institutions, les pratiques et les attitudes sociales, et existe indépendamment des systèmes informatiques. Le biais technique est issu des propriétés techniques des systèmes utilisés, quand les hypothèses ne correspondent pas à tous les aspects auxquels ils sont appliqués. Quant au biais émergent, il survient lors de l'utilisation concrète avec les usagers ; ce biais n'est pas inhérent à la conception du système ni au contexte social, mais survient plutôt à la suite de l'interaction des deux dans un cas particulier.

Des exemples de ces trois formes de biais peuvent être trouvés lors de l'étude des langues. Le biais préexistant s'avère évident lorsqu'un gouvernement, une industrie ou une puissante société refuse de rendre l'information, les technologies ou les produits disponibles aux personnes parlant une ou plusieurs langues. Ainsi, au milieu des années 90, Microsoft Inc. refusa de fabriquer des versions de ses produits pouvant s'avérer compatibles avec des systèmes d'écriture non

19 Cette situation est similaire aux problèmes soulevés lorsque des dossiers médicaux personnels deviennent accidentellement publics par Internet.

romaine, tel que WorldScript, de Apple Computer Inc. Microsoft justifia sa décision en invoquant que le marché des applications non romaines était trop limité pour justifier une nouvelle version de leur produit ; par conséquent, cet exemple de biais pré-émergent était dicté par des raisons d'ordre économique.[20] Le biais technique survient avec les séquences de code de texte tel Unicode UTF-8, faisant en sorte qu'un texte en format non romain exige de deux à trois fois plus d'espace qu'un texte comparable en format romain. Ici, la raison provient d'aspects de compatibilité entre les anciens systèmes romains et les systèmes Unicode plus récents. Et finalement, le biais émergent survient lorsque des systèmes informatiques créés à une fin sont utilisés à d'autres. C'est le cas du système de bibliothèque numérique développé pour un contexte urbain et blanc en Nouvelle-Zélande, et qui fut mal accueilli par la population rurale des Maori.

Ces trois types de biais doivent être abordés de différentes façons. Le biais préexistant doit l'être par les ressources éducatives, juridiques et institutionnelles des pays, industries ou sociétés. Le biais technique peut être abordé dans la conception des principes sous-jacents aux systèmes informatiques eux-mêmes. Et les biais émergents doivent être abordés à la fois par l'éducation et le design, à partir des informations obtenues sur l'utilisation concrète des systèmes informatiques.

Étant donné que le développement d'Internet implique l'interaction de technologies, de conditions préalables, d'objectifs, d'industries et d'intervenants, ces trois formes de biais sont impliquées dans le développement linguistique sur Internet, à plusieurs périodes et endroits différents.

Internationalisation et Internet : conceptions populaires

Le contenu des médias populaires relativement au potentiel de biais linguistique sur Internet a tendance à refléter deux perspectives opposées. Wasserman a décrit cette opposition dans les termes suivants :

> Puisque Internet contribue à … l'augmentation de la prise de conscience au fait que la planète est interconnectée et interdépendante, il pourrait

20 Depuis cette époque, Microsoft a modifié sa position et créé des versions de ses produits pour les autres marchés linguistiques.

> s'agir de l'un des plus récents développements accélérant la globalisation… Parce que la globalisation est perçue comme une force émanant du monde dit « développé », certaines critiques entrevoient la destruction des lieux et de spécificités culturelles au sein des pays et communautés minoritaires. D'autre part, certaines critiques font valoir que les forces internationales et locales interagissent dans le processus de globalisation, en faisant un processus multidirectionnel pouvant s'avérer bénéfique aux cultures et aux langues locales, et même favoriser leur autonomisation (Wasserman, 2002:2).

Ceux appuyant cette deuxième perspective tentent à défendre les droits des minorités, alors que ceux en faveur de la première soutiennent les nouvelles technologies de réseaux d'information. La deuxième perspective constitue en quelque sorte une réaction aux changements rapides et profonds résultant de la popularité d'Internet, tandis que la première est largement favorisée depuis ses débuts par les partisans de la technologie.

Il est assez facile de trouver des comptes-rendus connus des équipes d'ingénierie ayant travaillé sur les premières versions d'ARPANET (le premier réseau informatique) et qui présentent l'organisation de façon idéalisée, démocratique et décentralisée (par ex. Hafner et Lyon, 1996) ou le Whole Earth Lectronic Link (aussi connu sous l'acronyme WELL) disséminant des communautés virtuelles à travers le monde par le biais d'Internet (Rheingold, 2000). À partir de cette perspective, il est facile d'extrapoler que la domination linguistique serait une forme d'inégalité que la technologie Internet permettra d'éliminer rapidement. Tout d'abord (selon cet argument), Internet est international et décentralisé ; aucun usager ou groupe d'usagers ne peut posséder un contrôle hiérarchique sur un autre usager ou groupe d'usagers, parce que Internet permet une liberté complète d'association. En ce sens, n'importe qui peut utiliser n'importe quelle langue, à condition qu'une autre personne soit disposée à faire de même. Ensuite, la croissance des internautes non anglophones, et notamment les personnes parlant chinois, devrait dépasser le taux de croissance actuel des internautes parlant anglais. En d'autres mots, l'anglais ne dominera éventuellement plus Internet, parce que beaucoup plus de gens parlent les autres langues. La question sur la détermination de quelle langue domine en ligne est simplement une affaire de distribution démographique. Et finalement, les partisans font valoir que les capacités suggestives d'action d'Internet tel Unicode pour le texte multilingue et les systèmes comme BabelFish pour la traduction d'instance de documents Web, peuvent

résoudre tous les problèmes que les internautes parlant d'autres langues peuvent avoir en utilisant l'information sur Internet. Il est à noter que cette perspective caractérise largement la position retenue dans le document *La diversité culturelle et linguistique dans la société de l'information*, une publication de l'UNESCO préparée pour le Sommet mondial sur la société de l'information (UNESCO, 2003).

Chacun de ces arguments possède une perspective opposée qui, de façon plus spécifique, soutient que la langue anglaise – et dans une certaine mesure d'autres langues européennes – domine les communications sur Internet. Les raisons invoquées sont en partie sociales et techniques. D'abord, on fait valoir que Internet est basé sur une infrastructure de télécommunications économiquement dominée par des sociétés américaines. Le centre géographique de connectivité du réseau global de télécommunications est situé aux États-Unis, de sorte que tout ce qui le favorise profitera de façon démesurée aux États-Unis, par le biais de coûts de communications moindres et d'un nombre accru de destinations atteignables. Ensuite, en dépit des tendances récentes, les internautes utilisant l'anglais restent le plus important groupe d'usagers sur Internet. À tout le moins, la proportion d'usagers parlant anglais sur Internet est disproportionnée par rapport aux populations parlant d'autres langues. Et en dernier lieu, la plupart des technologies sur Internet sont mieux adaptées à l'anglais. Les interfaces pour les alphabets non romains sont complexes ou n'existent pas encore pour certaines langues. Même des systèmes tel que Unicode comportent des biais techniques au profit de l'anglais, tandis que les systèmes de traduction ne sont pas suffisamment fiables pour fonctionner à l'échelle requise.[21]

Ces perspectives diffèrent dans la manière dont les trois types de biais identifiés par Friedman et Nissenbaum (1997) sont perçus. La démographie linguistique des usagers d'Internet soulève des questions de biais préexistant L'aspect de la disponibilité des capacités suggestives d'action (affordances) pour différentes langues soulève des questions de biais techniques. De plus, les enjeux liés à la décentralisation en opposition au contrôle central de facto soulèvent la question de biais émergeant dans un système ayant dépassé ses frontières nationales d'origine.

21 Des variantes de ces deux positions, ainsi que leurs rapports avec des perspectives semblables sur la globalisation sont discutées dans Block (2004).

Malgré les divergences d'opinions et de vifs débats parfois suscités, il existe une pénurie de recherche empirique portant directement sur ces questions de biais linguistique préexistant, technique et émergent sur Internet. Ceci s'explique en partie par l'étendue et l'évolution rapide d'Internet. Ces deux conditions compliquent l'obtention de données fiables. Et même si des sondages linguistiques sont parfois effectués par des entreprises de marketing comme Jupiter Research (http://www.jupiterresearch.com/), et Global Reach (http://www.glreach.com/), ces données ont une valeur discutable sur le plan du biais linguistique, en raison des intérêts économiques sous-jacents des spécialistes du marketing et de leurs clients. De plus, un sondage fiable et effectué à grande échelle sur le multilinguisme en ligne serait dispendieux, au-delà des budgets limités ou des recherches non financées.

Sources de biais préexistant

Les biais préexistant concernent les institutions, pratiques et attitudes sociales indépendantes des technologies. Les sources de biais préexistant incluent la répartition historique des populations linguistiques, les ententes économiques favorisant des langues plus répandues, ainsi que les politiques institutionnelles des états nations. Au chapitre de la diversité linguistique sur Internet, les biais préexistant se retrouvent à la disposition des gouvernements, institutions et entreprises envers les personnes de différentes origines linguistiques, face à la mise en application d'une politique sur les technologies de l'information. La compréhension de tels biais s'avère complexe, mais puisque Internet est un phénomène international, cette compréhension doit s'effectuer dans le contexte de la diversité linguistique globale.

Diversité linguistique globale

Toute discussion sérieuse sur la diversité linguistique à l'échelle internationale ou régionale requiert un indice quantitatif de diversité. Malheureusement, de telles mesures quantitatives de diversité linguistique sont rarement utilisées à l'heure actuelle en recherche linguistique, et aucune mesure reconnue n'est utilisée à grande échelle. Les mesures déjà existantes ont tendance à être plutôt simplistes, tel le nombre de langues ou le nombre de groupes linguistiques, utilisés par Barrera-Brassols et Zenck (2002) ainsi que Smith (2001). Des mesures de diversité

plus élaborées furent proposées par le passé (par ex. Greenberg, 1956 ; Lieberson, 1964), mais leur valeur statistique n'était pas toujours bien fondée et elles sont devenues désuètes. L'approche retenue dans le présent rapport suit celle de Nettle (1999) et utilise une mesure de la variance comme indice de diversité.

Un indice de diversité linguistique satisfaisant doit tenir compte de plusieurs facteurs. D'abord, il doit comporter une certaine unité d'analyse, tel un pays, un continent ou Internet. Ensuite, cette diversité linguistique devrait tenir compte des probabilités de trouver des usagers d'une langue particulière. Le minimum naturel devrait être zéro, dans le cas d'une population entièrement homogène, et ne comporter aucune valeur maximale fixe. Une variété accrue de langues devrait augmenter la valeur de l'indice, mais à mesure que la proportion du groupe linguistique diminue, sa contribution à la diversité devrait aussi diminuer. De cette façon, les pays où l'on retrouve plusieurs groupes linguistiques d'importance semblable (par ex. la Tanzanie ; Mafu, 2004) démontreront une diversité linguistique plutôt élevée, tandis que les pays ayant un nombre comparable de langues, mais avec seulement une ou deux langues dominantes (par ex. les États-Unis), afficheront une diversité linguistique relativement peu élevée. Une mesure qui possède ces propriétés est la construction information-théorique appelée « entropie », sur laquelle nous pouvons baser notre mesure de diversité linguistique. En termes statistiques, l'entropie est une mesure de variance (écart). L'entropie est calculée à partir de la proportion estimée de la population du pays pour chaque langue, multipliée par son logarithme naturel et en faisant la somme de toutes les données pour une unité particulière (pays, région). La valeur de l'indice final représente - 2 fois cette somme.

Le Tableau 1 et la Figure 1 présentent les chiffres pour cette mesure de diversité basée sur l'entropie, dans différentes régions du monde, en fonction des 7 639 chiffres sur les populations linguistiques présentés dans Ethnologue (www.ethnologue.com) et allant de la diversité linguistique la plus faible à la plus élevée. Les États-Unis, d'où provient Internet, ont été séparés dans la première rangée à des fins comparatives. Les régions bien connues pour leur diversité linguistique (par ex. l'Afrique, l'Océanie) font voir la plus grande diversité linguistique, tandis que les régions ayant des langues nationales très répandues (Asie de l'Est, Amérique du Nord) affichent la plus faible diversité. Ces deux dernières régions sont particulièrement importantes pour comprendre la diversité linguistique sur Internet. Les États-Unis et la Chine sont sans doute les deux joueurs les plus importants sur Internet (certaines prévisions estiment que le nombre d'usa-

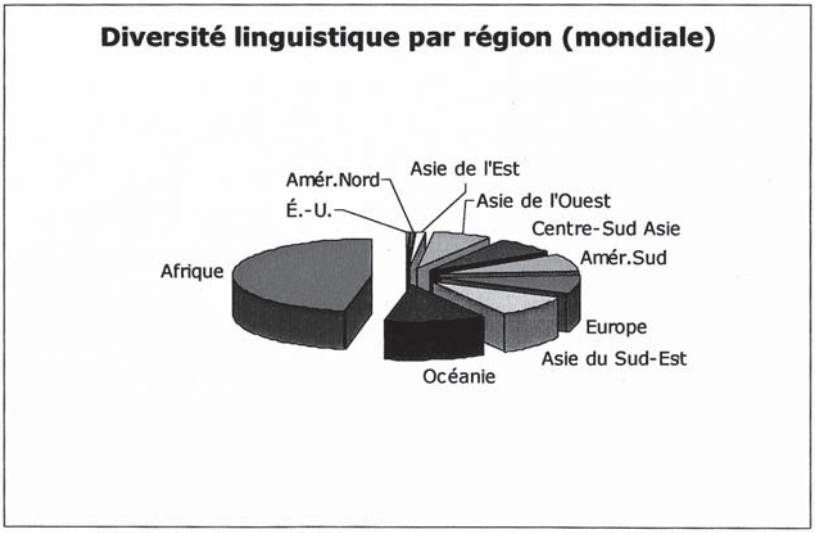
gers en Chine pourrait supplanter celui aux États-Unis au cours des prochaines années), et aucun de ces deux pays n'est très diversifié sur le plan linguistique, en comparaison à l'Océanie ou à l'Afrique. Dans la mesure où ces deux pays dominent Internet (ou par extension, la discussion de la diversité linguistique sur Internet), ce dernier ne peut prétendre refléter la diversité linguistique à l'échelle internationale.

Tableau 1. Résultats aux indices de diversité linguistique par région

Région	Langues	Indice de diversité	Proportion de population mondiale
États-Unis	170	0,7809	0,0020
Amérique du Nord (incl. Etats-Unis)	248	3,3843	0,0086
Asie de l'Est	200	4,4514	0,0112
Asie de l'Ouest	159	26,1539	0,0659
Centre Sud de l'Asie	661	29,8093	0,0752
Amérique du Sud	930	30,5007	0,0769
Europe	364	32,4369	0,0818
Asie du Sud-est	1 317	37,6615	0,0949
Océanie	1 322	46,5653	0,1174
Afrique	2390	185,6836	0,4681

Source : Ethnologue.

Figure 1. Indice de diversité linguistique par région

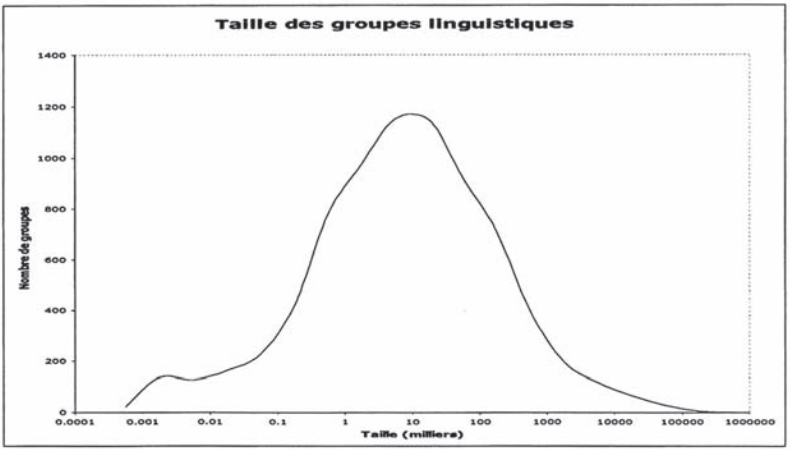


Source : Ethnologue.

Évolution de la diversité linguistique

Pour obtenir une perspective sur la signification de la diversité linguistique, on peut examiner l'importance des populations linguistiques. La Figure 2 illustre différents groupes linguistiques de différentes tailles, aussi issus des données d'Ethnologue. L'axe horizontal est à l'échelle logarithmique, ce qui signifie que la courbe en forme de cloche reflète une distribution normale (Grimes, 1986). L'importance habituelle d'un groupe linguistique se chiffre dans les dizaines de milliers de personnes – soit la taille d'une petite communauté urbaine. Les langues parlées par centaines de millions de personnes telles l'anglais, le chinois, l'espagnol, le français, etc. sont plutôt atypiques, tout comme le sont les plus petits groupes linguistiques regroupant quelques centaines de personnes. En termes d'expérience humaine, la situation est différente : près de la moitié de la population de la planète parle une langue parlée par des centaines de millions d'autres personnes.

Figure 2. Taille des groupes linguistiques



Source : Ethnologue, UNPD.

La diversité linguistique internationale et régionale n'est pas statique mais évolue avec le temps. Elle subit l'influence d'événements socio-historiques telles l'immigration massive, la colonisation, les guerres, les épidémies, et autres. À l'heure actuelle, la diversité linguistique globale est en déclin et ce depuis longtemps. Pour les linguistes qui étudient la diversité de la parole humaine, la situation actuelle est une catastrophe. La disparition de centaines de langages à une époque récente signifie que des pans entiers de connaissances sur cette capacité propre à l'humain sont perdus à jamais, tout comme le sont les littératures, l'histoire et les cultures des populations qui parlaient ces langues. La situation est catastrophique pour les populations concernées. Selon certaines estimations, près de la moitié des langues parlées dans le monde auront disparu d'ici l'an 2050 (Dalby, 2003 ; Krauss, 1992 ; Nettle et Romaine, 2000). Alors que la diversité linguistique disparaît avec l'extinction des plus petits groupes linguistiques, la proportion de personnes parlant une langue très répandue augmente à travers le monde.

La perte de la diversité linguistique n'est pas tributaire d'une région particulière dans le monde : beaucoup de langues ont disparu en Europe depuis la montée des états nations. En Amérique du Nord, en Amérique du Sud ainsi qu'en Australie, la colonisation européenne fut suivie de pertes tragiques qui continuent à notre époque. Dans les îles du Pacifique et en Indonésie, l'anglais et l'indonésien

remplacent les langues autochtones. Et en Asie, les principales langues parlées en Chine, au Japon, en Inde et en Russie se sont développées depuis des siècles au détriment d'autres langues (Crystal, 2000 ; Muhlhausler, 1996).

Certaines causes de disparition linguistique sont évidentes. Par exemple, l'extermination accidentelle ou voulue d'un groupe de gens peut entraîner la disparition de sa langue (Wurm, 1991). La majeure partie de la diversité linguistique nord-américaine disparut de cette façon : les guerres avec les immigrants européens et l'apparition de maladies étrangères qui se répandirent au contact des Européens décimèrent les populations autochtones à un point tel que leur langue disparut. D'autres causes de disparition linguistique sont moins évidentes, notamment quand ces changements sont attribuables à l'écologie culturelle.

Diversité linguistique globale et Internet

La faible diversité linguistique, notamment en Amérique du Nord, en Amérique latine et dans les Caraïbes, en Europe ainsi qu'en Asie de l'Est, facilite l'accès à Internet par le truchement d'un nombre réduit de solutions technologiques standardisées visant chaque population linguistique majeure. Dans les régions et pays ayant une plus grande diversité linguistique, des ententes plus complexes pour l'accès à Internet sont généralement requises, pouvant exiger l'adaptation des ressources à chacune des nombreuses langues minoritaires. En ce sens et dès le départ, Internet s'avère biaisé en faveur des langues plus répandues. Mais même les groupes linguistiques importants ont rarement des normes techniques soutenues. Par exemple, des centaines de millions de personnes parlent hindi, mais un chercheur de l'Université Southern California estime que la plupart des sites Web en hindi possèdent leurs propres polices de caractères en hindi, non compatibles avec les autres polices en hindi. Les usagers désireux de lire le matériel en hindi sur ces sites Web doivent installer les polices de caractères requises sur chaque site individuel, et la recherche sur ces différents sites s'avère extrêmement ardue puisque les mots ne correspondent pas aux différentes représentations (Information Sciences Institute, 2003). En d'autres mots, Internet ne favorise pas d'une manière égale les grands groupes linguistiques. Les régions comme l'Afrique, l'Océanie et l'Asie du Sud-est font face à des défis encore plus sérieux, en raison du grand nombre de langues non encore en usage sur Internet. Par conséquent, des développements techniques importants restent à faire avant de parvenir à atteindre ces groupes linguistiques.

Il importe de conserver une perspective évolutionniste lorsque l'on examine les effets d'Internet. Même si Internet peut très bien avoir un impact à long terme sur la diversité linguistique, tant la nature que l'envergure de cet impact en termes historiques ne sont pas claires. Puisque Internet améliore l'accès aux langues individuelles, il contribue éventuellement à les renforcer, mais puisqu'il fait de même pour les langues plus répandues en favorisant les échanges linguistiques, il contribue également à les affaiblir. Ces deux effets pourraient être beaucoup moins importants que l'influence d'autres causes sociales tout aussi omniprésentes dans la diversité linguistique. Elles représentent notamment le développement de l'agriculture, de l'urbanisation des populations, des événements géopolitiques, etc., dont tout gouvernement ou agence de coopération telle les Nations Unies pourraient très bien ne pas pouvoir empêcher. Par la même occasion, le monde constate le déclin réel de la diversité linguistique, alors que la survie de centaines de communautés historiques et culturelles à travers le monde est directement menacée. Il importe donc que toute politique axée sur la diversité linguistique sur Internet tienne compte de ces préoccupations.

Sources de biais émergent

Le biais émergent porte sur les effets du biais survenant avec l'usage actuel des technologies Internet. Eu égard à la diversité linguistique sur Internet, le biais émergent est fondé sur l'expérience des usagers des technologies de l'information quand leurs antécédents linguistiques deviennent tributaires de leur capacité à employer la technologie ou l'information fournie. Ce biais se manifeste surtout de deux façons : d'abord dans la distribution linguistique sur Internet, et ensuite par le contrôle économique du marché des télécommunications et des technologies de l'information. Dans cette section, nous examinerons les sources de tels biais émergents. Les résultats présentés ici suggèrent à l'heure actuelle un biais important en faveur de l'anglais.

Diversité linguistique des sources d'information sur Internet

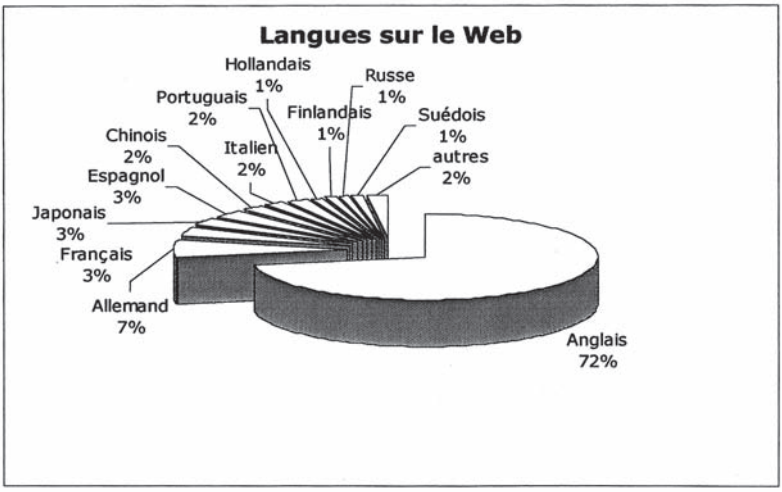
Quelques études ont effectué une analyse quantitative à grande échelle sur les langues utilisées sur Internet. Ces études portent généralement sur le Web, à l'exclusion des autres modes de communication tels le courriel et le clavardage (chat), puisque le Web s'observe plus directement et s'étudie plus facilement que

les autres formes de communication Internet. Deux recherches dignes d'être mentionnées ont produit des résultats intéressants dans ce domaine : une série d'études par Lavoie, O'Neill et des collègues du Online Computer Library Center (OCLC), ainsi qu'une étude de Nunberg (1998) à PARC sur des sites Web non anglais.

Les études du OCLC (Lavoie et O'Neill, 1999 ; O'Neill, Lavoie et Bennett, 2003) ont eu recours à un échantillon au hasard de sites Web disponibles sur Internet. Pour y parvenir, ils ont généré des numéros de protocole Internet (IP) et tenté de se relier à un site Web pour chaque adresse. Si un serveur Web répondait, les chercheurs téléchargeaient alors sa page d'accueil et activaient un système automatisé de classification linguistique sur cette page (O'Neill, McClain et Lavoie, 1997). Cette méthode d'échantillonnage a l'avantage d'être non biaisée. Toutes les autres méthodes d'échantillonnage reposent directement ou non sur des moteurs de recherche ou des « robots Web » (web spiders), soient des programmes qui découvrent de nouvelles pages Web en suivant tous les liens dans une série connue de pages Web. Les robots Web produisent ce qu'on appelle un « sondage cumulatif », c'est-à-dire un échantillon biaisé en raison de sa proximité à un point de départ arbitraire. Les moteurs de recherche dépendent des robots Web pour développer leurs indices, de sorte que les échantillons qui en sont tirés sont également biaisés. De tels échantillons biaisés sont à éviter, si l'on veut obtenir une estimation fiable de la fréquence des différentes langues sur le Web.

La première étude fut effectuée à deux périodes différentes, à intervalle d'une année, afin d'évaluer les tendances dans l'usage de ces différentes langues. En 2002, une étude subséquente chercha à confirmer ces observations. L'étude menée en 1998-1999 suggérait qu'une certaine expansion internationale était en cours sur le Web, et que l'usage de différentes langues correspondait étroitement au domaine de provenance de chaque site Web. Ainsi, l'échantillon de 1999 comportant 2 229 sites Web au hasard permit d'identifier 29 langues différentes dont la répartition est présentée à la Figure 3. Comme on pouvait le prévoir, l'anglais domine clairement dans 72% de l'ensemble des sites Web échantillonnés. L'indice de diversité de cet échantillon de pages Web équivaut à 2,47, soit moins que celui d'un pays caractéristique de l'Asie du Sud-est et plus qu'un pays caractéristique du Centre-Sud de l'Asie. Cet indice est aussi des centaines de fois plus petit que l'indice linguistique global. En ce sens, la diversité linguistique sur le Web, même si elle s'apparente à celle de plusieurs pays multilingues, constitue une faible représentation de la diversité linguistique à travers le monde.

Figure 3. Proportion de langues sur le Web à partir d'un échantillon au hasard de pages Web



Source : O'Neill, Lavoie et Bennett (2003).

En relation à l'étude précédente, l'étude de suivi effectuée en 2002 indique que la proportion de l'anglais sur le Web semble constante, même si de petites différences surviennent parmi les autres langues (O'Neill, Lavoie et Bennett, 2003). L'indice de diversité était de 2,44 en 2002, démontrant peu d'écart sur l'étude précédente, ce qui pourrait être en partie attribuable à la méthodologie utilisée. Les 29 langues identifiées dans l'échantillon des pages Web représentent en fait la limite du programme d'identification linguistique utilisé par ces chercheurs (http://www-rali.iro.umontreal.ca/SILC/SILC.en.cgi), et les nouvelles langues utilisées sur le Web ne peuvent être découvertes de cette façon. Même si le programme d'identification linguistique pouvait englober davantage de langues, celles-ci ne représentent que de faibles proportions et par conséquent, changeraient peu la diversité calculée sur le Web.

En 1999, l'étude de l'OCLC identifia aussi les proportions de pages Web multilingues à partir de chaque domaine d'origine, et quelles combinaisons de deux langues étaient employées. Si un site Web comptait plus d'une langue, l'anglais était toujours l'une d'elles : l'ensemble (100%) des 156 sites multilingues identifiés utilisait l'anglais. Le français, l'allemand, l'italien et l'espagnol étaient

chacun présents sur environ 30 % des sites Web multilingues, tandis que les autres langues étaient beaucoup moins bien représentées. De plus, 87 % des sites Web multilingues provenaient de domaines situés à l'extérieur des principaux pays anglophones (Australie, Canada, Royaume-Uni et États-Unis). Pour l'ensemble des domaines, le taux de multilinguisme allait de 6 sur 13 (42 %) sur les sites russes, à 16 sur 1 103 (1,5 %) pour les sites américains. On constate donc que le Web penche fortement vers le monolinguisme, et la plupart des sites multilingues ne viennent que refléter la domination de l'anglais. Ce résultat est directement à l'opposé de la croyance répandue selon laquelle le Web favorise la diversité linguistique.

Les tendances observées dans les études de l'OCLC ont été confirmées dans l'étude de Nunberg (1998), qui a retenu une méthodologie différente. Dans cette étude, un robot Web *(web crawl)* de 2,5 millions de pages collectées en 1997 par Alexa, une firme de services Internet, fut analysé à l'aide d'un identificateur automatique de langue préparé par Heinrich Schütze, un collègue de Nunberg. Même s'il s'agit d'un sondage cumulatif biaisé, il est néanmoins plus de mille fois plus grand que celui de l'OCLC. Nunberg a surtout constaté que les pays ayant un faible taux de pénétration d'Internet utilisent surtout l'anglais sur leurs sites Web, tandis que ceux ayant un taux plus élevé de pénétration ont davantage recours à des langues autres que l'anglais. Il est à noter que l'Amérique latine s'inscrit à contre-courant de cette tendance, avec un taux de pénétration Internet très faible en 1997 et une prédominance écrasante de sites Web dans une autre langue que l'anglais. En ce sens, l'étendue du bilinguisme anglais dans un pays non anglophone peut influencer l'expression de la diversité linguistique sur ses sites Web.

Outre les études déjà citées, quelques autres tentatives ont voulu mesurer la distribution linguistique à partir des statistiques obtenues des moteurs de recherche. Pour diverses raisons, l'information recueillie n'est pas aussi utilisable. Par exemple, FUNREDES, une ONG favorisant les technologies de l'information et de communication en Amérique latine, a mené une série d'études depuis 1995 en vue d'évaluer la distribution linguistique et les influences nationales sur Internet (Pimienta et Lamey, 2001 ; Pimienta et autres, 1995-2003). Ces études ont recensé le nombre de pages Web indexées par des moteurs de recherche bien connus, à partir de certains mots sélectionnés dans différentes langues et groupes nationaux. Ces chercheurs ont notamment recueilli une proportion beaucoup plus faible de pages anglaises (52 % en 2001, 45 % en 2003) que dans les études menées par Lavoie et O'Neill ainsi que Nunberg.

Le calcul du nombre de pages dérivées des moteurs de recherche s'avère toutefois une méthodologie non fiable en vue de déterminer la représentation linguistique sur le Web. Outre les échantillons biaisés fournissant des pages aux moteurs de recherche, on retrouve plusieurs autres influences confondantes. Les moteurs de recherche ont généralement recours à différentes méthodes d'indexation propriétaire ne pouvant être inspectées, ce qui peut biaiser le total de pages retournées de façon impossible à corriger ni même d'évaluer. Un mot qui n'est pas sur une page peut être calculé dans le total, tandis que des pages contenant le même mot peuvent ne pas être calculées. De plus, la méthode assume que la fréquence des mots reliés aux concepts « culturellement neutres » est uniforme d'une langue à l'autre. Cependant, la neutralité culturelle est inaccessible. Beaucoup de mots observés fréquemment représentent des concepts culturels, tel que le mot « cheese ». La culture anglo-américaine et la culture française continentale attribuent une signification alimentaire très différente aux mots *cheese* et *fromage* respectivement. Ces faits seront représentés par la fréquence des termes correspondants. De plus, puisque le total des pages est retourné (plutôt que le total de mots), les totaux retournés pour différentes formes linguistiques peuvent inclure des pages bilingues à multilingues, calculées plusieurs fois.

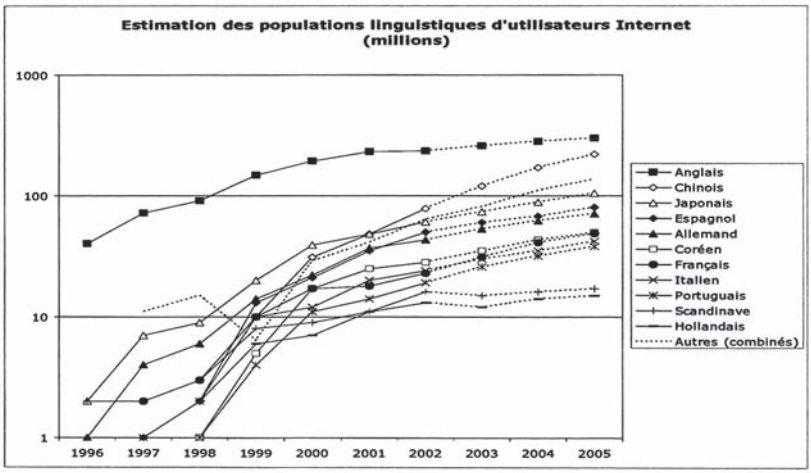
Diversité linguistique parmi les usagers d'Internet

La firme de services de traduction Global Reach a déployé les efforts les plus directs en vue d'évaluer la diversité linguistique des internautes. Ces évaluations, préparées annuellement de 1996 à 2002, sont largement citées comme montrant un Internet où la diversité linguistique s'accroît sans cesse.[22] Ces estimations sont basées sur celles de l'Union internationale des télécommunications (UIT) pour les populations d'usagers dans chaque pays, qui définissent un « usager » comme une personne ayant utilisé Internet au cours des trois derniers mois. Ces populations d'usagers sont réparties en populations linguistiques calculées à partir des estimations d'Ethnologue et réajustées avec des données démographiques de l'ONU, comme nous l'avons fait plus haut dans le calcul de la diversité linguistique. Dans certains cas, les auteurs ont complété ces sources avec des statistiques de marketing obtenues de firmes telles que Nielsen Net Ratings. Ces données n'incluent aucune forme d'étude réelle auprès des internautes, de sorte que les données de

22 Ces données sont disponibles sur le site http://global-reach.biz/globstats/evol.html.

Global Reach ne représentent pas les langues vraiment parlées par les usagers d'Internet. Puisque ces figures sont souvent citées à l'appui de la diversité linguistique des internautes, il est approprié de les examiner de plus près.

Figure 4. Estimation des populations linguistiques d'utilisateurs Internet (échelle logarithmique pour l'axe y).



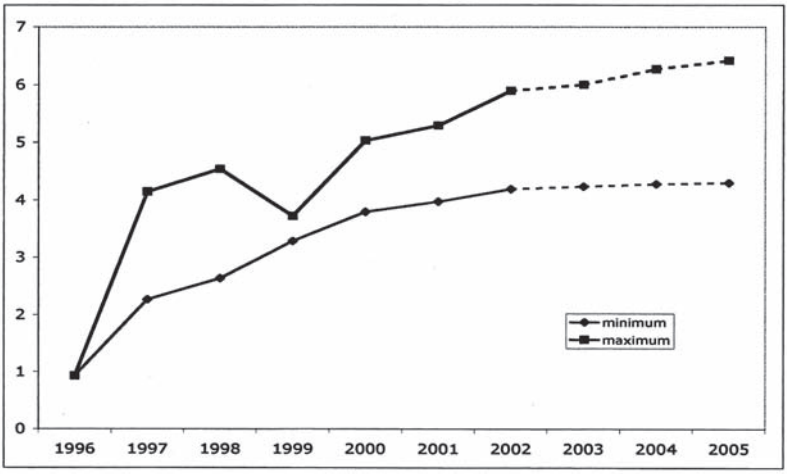
Source : Global Reach.

La Figure 4 présente l'estimation de Global Reach pour les populations des différentes langues. La période de 2003 à 2005 est indiquée par une ligne pointillée, puisqu'il s'agit de prévisions. Les langues identifiées s'apparentent à celles des études de l'OCLC. Comme il fallait s'y attendre, l'anglais avec quelque 230 millions d'usagers avait près de trois fois plus d'usagers en 2001 que la langue suivante, le chinois, avec environ 60 millions d'usagers.[23] La Figure 4 fait voir que tous ces groupes d'usagers semblent en croissance exponentielle, sauf pour l'anglais et le japonais qui semblent ralentir. On estime qu'environ 50 % de la population de ces deux groupes linguistiques utilise déjà Internet.

23 Ces estimations semblent traiter de façon semblable toutes les formes de chinois, même si les linguistes estiment que le chinois représente un groupe de neuf langues différentes (souvent appelés « dialectes » dans le grand public).

À partir des estimations de Global Reach, on peut calculer les indices de diversité linguistique pour l'ensemble des usagers d'Internet ; ces valeurs sont présentées à la Figure 5. Puisque la composition du groupe linguistique « Autres » n'est pas définie dans les données de Global Reach, nous avons calculé des valeurs minimum et maximum pour l'indice, en assumant que « Autres » représente une seule langue (diversité minimale) ou une distribution uniforme parmi 6 000 langues (distribution maximale). Il est étonnant de constater qu'en dépit d'importants gains initiaux de l'indice de diversité entre 1996 et 1999, la diversité linguistique semble se stabiliser après l'an 2000, en dépit de la croissance exponentielle de beaucoup de langues. De plus, les prévisions pour 2003-2005 continuent de démontrer cette tendance à la stabilisation ; l'augmentation prévue du nombre de personnes parlant chinois, en raison de son importance, limite dans les faits l'augmentation de diversité. Il en résulte un indice de diversité linguistique se situant entre celui d'un pays africain typique et les indices régionaux combinés de l'Amérique du Nord et de l'Europe. Ce résultat n'est probablement pas étonnant, étant donné que les hôtes Internet restent concentrés en Amérique du Nord et en Europe. Malgré tout, la diversité linguistique d'Internet n'est nulle part aussi grande que l'indice de toute autre région ou du monde dans son ensemble. Par conséquent, et contrairement à l'opinion répandue, on ne peut affirmer que Internet englobe la diversité linguistique en ce sens.

Figure 5. Estimation de la diversité des usagers d'Internet



Source : Global Reach.

En ce sens, Internet n'a pas acquis sa diversité linguistique simplement en étant international et en reliant entre eux beaucoup d'internautes. Pour s'assurer que les langues des internautes sont représentées en ligne, il faudra s'attaquer à d'autres enjeux, et comme on peut le constater ci-dessous, ces enjeux peuvent s'avérer très spécifiques aux contextes des communautés en ligne.

Internet et la pratique du multiculturalisme

L'accès à Internet est un pré requis à l'utilisation de l'information qu'il fournit. Jusqu'ici, nous avons examiné ce qu'un tel accès signifie en termes globaux. Cependant, un tel effort ne saurait réussir si les personnes parlant les différentes langues à travers le monde choisissent tout simplement quelques langues dominantes. Quels sont alors les facteurs qui dictent le choix linguistique sur Internet ?

Les langues ne servent pas seulement à transmettre des informations – il s'agit aussi de systèmes complexes de symboles comportant des évaluations riches et subtiles de leur contexte d'application. Les études sociolinguistiques sur le multiculturalisme ont largement détaillé les écologies sensibles et turbulentes des langues en contact ; des recherches récentes sur le multiculturalisme d'Internet soulignent la pertinence de ces leçons en rapport à Internet. De plus, l'intérêt international évident envers Internet repose sur les avantages économiques offerts. Internet favorise-t-il aussi de la même façon les langues plus répandues ?

Il n'est pas simple d'identifier en termes généraux quelles langues sont employées en ligne et la façon dont elles le sont. Plusieurs enjeux sont impliqués, allant des communautés linguistiques individuelles à l'accès différentiel à Internet, aux différents systèmes d'écriture et de codage informatique, jusqu'aux divers modes de communication. La majeure partie de la recherche disponible concernant les effets éventuels d'Internet sur la langue et la culture porte sur des études de cas de groupes linguistiques particuliers utilisant Internet dans des contextes précis, plutôt que dans une perspective macro sociale. Ces études de cas suggèrent que le contact linguistique sur Internet favorise les langues répandues, tout comme les contacts hors ligne. Ainsi, Wright (2004) et Holmes (2004) ont examiné le comportement linguistique en ligne d'étudiants de niveau universitaire dans huit pays. Leurs résultats indiquent que l'ampleur avec laquelle les gens utilisent leur langue maternelle en ligne varie énormément selon le contexte examiné. Par la même occasion, aucune population étudiée ne démontre qu'elle utilise son

répertoire linguistique complet en ligne. Les langues moins utilisées ne semblent pas être du tout employées sur Internet. En ce sens, les questions concernant cet enjeu sont à la fois subtiles et complexes.

Lors de recherches préliminaires, Paolillo (1996) constata que l'anglais est largement préféré au pendjabi lors de discussions de groupes Usenet regroupant principalement des internautes de langue pendjabi. Un tel comportement est en partie prévisible de la part des participants majoritairement expatriés et éduqués en anglais, mais les tendances observées marginalisent l'usage en ligne du pendjabi au point où il devient réservé à des fonctions de communications de nature très rituelle ou nationaliste, et sert davantage d'outil d'identification que de transmission de l'information. Dans un article subséquent, Paolillo (2006) compara l'interactivité et l'homogénéité linguistiques des Asiatiques du Sud en contact entre eux dans les clavardoirs (chat rooms) et des groupes de discussion sur Internet, et constata que ces deux moyens favorisent le recours à une langue minoritaire (hindi ou pendjabi, selon le forum). Ces tendances sont aussi signalées dans Peel (2004), qui indique que des clavardoirs interactifs dans les Émirats arabes unis favorisent l'arabe, alors que les courriels privilégient l'anglais. Dans un autre article, Paolillo (2001) constata que les participants centraux sur un canal de clavardage étaient plus enclins à utiliser des langues minoritaires que les participants périphériques. Puisque les clavardoirs facilitent le va-et-vient des participants, les participants périphériques et leurs préférences linguistiques prédominent. En ce sens, les aspects technologiques et sociaux de la communication Internet interagissent de façons complexes qui favorisent néanmoins les langues majoritaires, plutôt que minoritaires. Il est possible de manipuler les variables technologiques afin de limiter dans une certaine mesure les effets de la domination linguistique, mais leur efficacité est inconnue.

L'étude du grec dans les communications sur Internet par Koutsogiannis et Mitsakopolou (2004), Georgakopoulou (2004, à paraître) et Androtsopolous (1998) explore une gamme d'enjeux recoupant les études citées plus haut. Tout comme le caractère d'écriture gurmukhi du pendjabi, l'alphabet grec est difficile à utiliser sur Internet, de sorte que l'on préfère une forme « romanisée » du grec adaptée d'un alphabet hors ligne appelé « greeklish », surtout chez les expatriés vivant en milieu multilingue préférant l'anglais (Georgakopoulou, 2004) ou l'allemand (Androtsopolous, 1998). Cet état de fait corrompt en retour la norme grecque de la diglossie (Ferguson, 1959), alors que ceux qui l'utilisent ont recours à une forme vernaculaire distincte pour la langue parlée informelle et à la langue

classique à l'écrit. À différentes époques par le passé, le gouvernement de la Grèce a déployé beaucoup d'efforts pour conserver l'alphabétisme envers le Katharevousa, la langue classique pour l'écriture formelle, mais l'érosion de la diglossie grecque sur Internet risque de miner ces efforts. Dans un autre contexte de diglossie, arabe cette fois, Warschauer et autres (2002) observent que l'égyptien, l'arabe et l'anglais vernaculaires empiètent sur les fonctions traditionnelles de l'arabe classique. De tels empiétements tendent à déstabiliser les situations diglossiques, menant éventuellement à un changement linguistique vers une langue extérieure dominante. En ce sens, lorsque les normes linguistiques sont érodées sur Internet, la disposition universelle de l'accès à Internet pourrait avoir un effet potentiellement nuisible sur une telle diversité linguistique.

L'influence de l'anglais est à la fois répandue et subtile. Des situations très différentes impliquent l'usage du courriel en Suisse (Durham, 2004) et d'Internet en Tanzanie (Mafu, 2004), où les résidents bilingues de chaque pays préfèrent utiliser l'anglais plutôt que les langues locales plus évidentes. Alors qu'il existe un précédent colonial favorisant l'anglais chez les élites de la Tanzanie, ce n'est pas du tout le cas en Suisse. L'explication d'un tel phénomène se constate en raison du statut international de l'anglais (Crystal, 2003 ; Phillipson, 1992, 2003). Un autre exemple de l'influence de l'anglais sur Internet implique la diffusion de certaines caractéristiques de la langue parlée dans l'écrit, par le biais de courts messages (SMS), de messages instantanés (IM), et de clavardage sur le Web en suédois (Hård af Segerstad, 2002). De même, Torres (1999, 2001) observe plusieurs fonctions pragmatiques des émoticons (« binettes » ou smileys) en clavardage catalan. Ces formes proviennent de contextes parlés anglais, et témoignent de l'influence du contact de l'anglais au catalan par le biais d'Internet.

Ces études et d'autres font ressortir tant la richesse que la complexité des facteurs reliés à l'usage des langues minoritaires par des usagers multilingues d'Internet. Plusieurs de ces études soulignent la précarité de l'usage des langues non dominantes dans les communications Internet.

Institutions et intérêts gouvernant Internet

Contrairement à la croyance populaire, Internet n'est pas une institution ouverte et démocratique (ou anarchique). Il s'agit plutôt d'une institution ayant un réseau complexe d'intérêts puissants, dont beaucoup sont très centralisés. Ces intérêts

puissants se préoccupent peu des actions des usagers individuels, laissant croire que Internet ne subit aucunement la contrainte des intérêts civils, gouvernementaux ou corporatifs. Néanmoins, chaque niveau d'intérêt constitue l'occasion de biais linguistiques en vue de déterminer quelles langues sont utilisées sur Internet.

Plusieurs acteurs importants et différents sont impliqués dans la réglementation d'Internet. En premier lieu, on retrouve des monopoles et oligopoles en télécommunications dans différentes régions du monde. Ces sociétés maintiennent les infrastructures permettant aux individus de se brancher à Internet, et de relier entre eux les divers sites Internet. Ensuite, on retrouve les sociétés de matériel informatique et de logiciels tels que Intel, IBM, Hewlett-Packard, Cisco Systems, Sun Microsystems, Microsoft, Adobe. Ces entreprises développent et commercialisent le matériel informatique et les logiciels qui constituent l'infrastructure d'Internet. De plus, il existe des organismes de réglementation particuliers à Internet, tels que Internet Corporation for Assigned Names and Numbers (ICANN), ainsi que les Centres d'information sur les réseaux tels que l'American Registry for Internet Numbers (ARIN), Réseaux IP Européens (RIPE) ainsi que l'Asia Pacific Networking Information Centre (APNIC), qui prennent des décisions relativement à la connectivité d'Internet. Les gouvernements nationaux jouent aussi un rôle, tant dans la gestion des ressources Internet au niveau du pays que dans l'application d'autres formes de politiques d'information. Enfin, on retrouve d'autres organisations et consortiums, tels que World-Wide Web Consortium (W3C), le Consortium Unicode, ainsi que l'Organisation internationale de normalisation (ISO), qui développent des normes pour l'application des technologies Internet.

Dès le début, le réseau téléphonique a toujours été important pour Internet. Lorsqu'un hôte Internet se branche à un autre hôte, les modems, lignes louées, lignes d'abonnés numériques, réseau d'infrastructure de fibres optiques et satellites géostationnaires peuvent tous être impliqués à une certaine étape de la communication numérique, acheminant les données sur le réseau téléphonique. Plus récemment, d'autres formes de réseaux de télécommunications tels que les réseaux de télévision par câble ont été adaptées au trafic de données Internet. Tant par le passé qu'à l'heure actuelle, de larges entreprises (souvent privées ou des monopoles étatiques) ont assumé le contrôle économique de ces ressources. À l'échelle internationale, ces sociétés se retrouvent surtout aux États-Unis. Par exemple, MCI gère un réseau acheminant une écrasante majorité du trafic international d'Internet par le biais de sa filiale UUNET (voir Mapnet, http://www.caida.org/tools/visualization/mapnet). Le réseau d'infrastructure

de fibres optiques mis en place par MCI il y a plusieurs années est essentiel à ce réseau. Même si les sociétés telle que MCI sont relativement peu intéressées par les langues des internautes sur leurs lignes de données, le rôle central des États-Unis dans la distribution du trafic de données fait en sorte que les tâches administratives de haut niveau reliées au trafic du réseau d'infrastructure se feront en anglais. En ce sens, les réseaux régionaux se raccordant à ces réseaux centraux devront nécessairement embaucher du personnel maîtrisant très bien l'anglais. Même si cette situation ne semble pas très inquiétante, étant donné que les informaticiens à travers le monde tendent à être très familiers avec l'anglais, ces deux tendances symbiotiques se complètent mutuellement. Si les autorités des réseaux régionaux ne peuvent communiquer avec leurs fournisseurs dans la langue de leur choix, l'anglais restera alors par défaut la langue dominante de l'administration du réseau. Les entreprises de télécommunications, qui encaissent des profits substantiels pour la demande de services de communication et de technologie, ont la responsabilité particulière de tenir compte de la diversité linguistique des pays sur les marchés desservis.

Les entreprises de matériel informatique et de logiciels ont une influence semblable sur le caractère linguistique d'Internet, en fabriquant des ordinateurs munis de claviers, écrans et systèmes d'exploitation favorisant certaines langues en particulier. Ces produits sont fabriqués à faible coût en réalisant des économies d'échelle, permettant le marketing d'un produit standardisé sur le plus vaste marché possible. La technologie informatique, avec ses usines de puces à l'étranger, le développement externalisé de logiciels (et même de la gestion), ainsi que les marchés de marchandises, constituent l'un des secteurs globalisés originaux de l'industrie. Pour cette raison, et aussi à cause de l'importance des sociétés américaines à développer de nouveaux systèmes et de nouvelles normes, les systèmes informatiques qui parviennent dans des régions de diversité linguistique comme l'Afrique sont en grande majorité conçus pour être utilisés en anglais ou dans une langue européenne, et sont très peu, sinon aucunement adaptés aux langues locales. De telles circonstances constituent une autre forme de biais émergent à l'endroit des langues européennes sur Internet, et au détriment des langues des pays moins industrialisés. Comme pour les entreprises de télécommunications, celles de matériel informatique et de logiciels ont une responsabilité particulière envers la diversité linguistique des pays sur les marchés desservis.

Ainsi, les actions des sociétés informatiques en étroite concurrence pour la domination de marché nuisent au climat de diversité linguistique en ligne et d'in-

formatique multilingue. Pour favoriser l'informatique multilingue, des ententes sont requises afin que les intérêts internationaux puissent avoir préséance sur les objectifs concurrentiels des sociétés privées. Certaines de ces tendances sont améliorées par les activités des organisations et consortiums internationaux, telle l'Organisation internationale de normalisation (ISO), le Consortium Unicode et le World-Wide Web Consortium, qui supervisent les différents aspects du développement technologique d'Internet. Plusieurs sociétés informatiques importantes (y compris Apple et Microsoft) sont impliquées dans ces organisations. Même si certains technologues déplorent que ces organisations entravent l'innovation, leur caractère international aide à tenir compte des intérêts des différents groupes nationaux et linguistiques. En revanche, ces organisations normatives n'ont pas vraiment de mécanisme de mise en application. Par conséquent, plusieurs technologies Internet possèdent des normes qui ne sont pas largement mises en pratique. C'est notamment le cas du langage HTML utilisé sur les pages Web et du langage de programmation ECMAScript pour l'interactivité du navigateur Web. Les incompatibilités entraînées par l'absence de conformité aux normes nuisent aux progrès de l'informatique multilingue. Si ces organisations visent à promouvoir et à protéger la diversité linguistique, leurs mécanismes d'intervention doivent être renforcés.

ICANN est un autre acteur gouvernant Internet et ayant un impact important sur la diversité linguistique d'Internet. ICANN administre le protocole connu comme système de noms de domaine (DNS), sous contrat avec le Département américain du commerce. Le système DNS accomplit la fonction d'associer des noms mnémotechniques uniques à tous les hôtes Internet, une fonction essentiellement linguistique. Malheureusement, le système DNS est difficile à utiliser avec d'autres langues que l'anglais américain et ne convient pas non plus avec la manière dont les systèmes de noms fonctionnent dans le langage humain. Le système DNS est profondément intégré au fonctionnement d'Internet, puisque la plupart des autres protocoles d'application Internet dépendent de lui pour repérer les hôtes Internet. Il s'agit aussi du seul protocole en réalité administré, plutôt que simplement codifié, par une autorité centrale. ICANN réglemente le système DNS surtout par délégation, mais sa structure administrative, son réseau de contrats avec le gouvernement américain et autres parties, ainsi que ses différentes politiques, ont tous concouru à limiter le multilinguisme dans la désignation des hôtes Internet. Par conséquent, le système DNS ne peut remplir son rôle initial de fournir des mnémotechniques utiles aux hôtes Internet. Des changements à ICANN, au système DNS lui-même ainsi qu'aux politiques d'administration des noms de domaine sont tous requis pour améliorer cette situation.

Les internautes considèrent les noms des hôtes Internet de la même façon que les autres noms. Mais dans les faits, ils sont très différents. Le système DNS requiert que les noms d'hôtes sur Internet soient globalement uniques, alors que dans une langue habituelle, il est peu probable qu'un nom particulier sera unique en raison des métaphores, du symbolisme et des acronymes. Lorsqu'un domaine « acl.org » est enregistré auprès de l'Association of Christian Librarians, il n'est plus disponible pour l'Association for Computational Linguistics ou toute autre organisation au monde désireuse de s'identifier sous le même acronyme.

Afin d'appliquer l'unicité tout en permettant une flexibilité limitée, le système DNS a recours à des noms structurés hiérarchiquement : les noms d'hôtes individuels comportent des séries de noms, en ordre de spécificité plus ou moins grande. Le premier niveau de la hiérarchie est le dernier champ du nom ; il s'agira d'un domaine de premier niveau générique ou de code du pays (domaine générique de premier niveau gTLD ou domaine national ccTLD) (TLD – Top-Level Domain), qui sert de classificateur général. Cependant, on ne sait pas toujours très bien quel classificateur est pertinent à une fin particulière. En vertu de leurs ententes avec ICANN, les domaines génériques de premier niveau TLD sont supposés être administrés pour différentes fonctions : .com est réservé aux sites commerciaux, .net aux réseaux, .org pour les organisations à but non lucratif, et les codes de pays doivent être administrés par les pays associés à leurs propres fins. Les noms de domaines gTLD sont cependant plus recherchés parce qu'ils ont tendance à être courts et plus faciles à retenir. Puisqu'il n'existe qu'un petit nombre de domaines génériques de premier niveau (gTLD) et des centaines de millions d'hôtes, il est inévitable que des conflits surgissent dans l'assignation des noms de domaines.

En présence de tels conflits et leur règlement, l'approche de ICANN favorise des marques de commerce légalement reconnues. Autrement, la première partie à enregistrer un nom de domaine le conserve, en autant que l'enregistrement est maintenu. Ceci ne favorise pas les inscrivants qui ne sont pas des détenteurs de marques de commerce, ou qui proviennent d'un petit milieu ou d'une langue minoritaire. Les inscrivants internationaux ne sont pas particulièrement favorisés non plus si leur identité naturelle ressemble à un homographe d'un nom de domaine déjà enregistré. Lorsqu'un domaine est enregistré, des négociations dispendieuses ou des mesures légales sont requises pour le modifier. L'enregistrement préalable de centaines de millions d'hôtes en anglais entraîne ainsi un biais évident à l'endroit des enregistrements d'hôtes non anglophones, étant donné

que plusieurs milliers de noms d'hôtes souhaitables dans d'autres langues seront des homographes d'hôtes déjà enregistrés dans le domaine générique de premier niveau (gTLD). En ce sens, dans le système DNS, l'attribution de marques de commerce – un enjeu juridique américain relié au commerce – a préséance sur l'attribution multilingue transparente de noms, un enjeu international lié à la langue et aux communications. Cette fausse conception des priorités ne changera pas tant que le système DNS ne relèvera pas d'une autorité complètement internationale, plutôt que d'un organisme privé ayant des liens contractuels avec le gouvernement américain (ou autre).

La conception originale du système DNS était fortement biaisée en faveur de l'anglais, en ce sens qu'il ne pouvait employer qu'un codage 7 bits US-ASCII. En ce sens, même les langues européennes telles que le français, l'espagnol et l'allemand, qui ont recours à des signes diacritiques non US-ASCII, sont désavantagées quand vient le temps de choisir des noms souhaitables pour les hôtes Internet. Plusieurs organisations, telles que Multilingual Internet Names Consortium (MINC), New.net et RealNames, ont tenté pendant des années de convaincre ICANN de développer des alternatives au système DNS actuel, en vue d'offrir une meilleure assistance multilingue. Malgré le fait que ces groupes aient présenté plusieurs propositions constructives méritant une étude plus attentive, ICANN a opposé beaucoup de résistance. ICANN n'a adopté que récemment une variation de l'Unicode, connue sous punycode, pour permettre les noms de domaines multilingues, mais son déploiement a fait l'objet de lenteurs insatisfaisantes et de lourdeurs politiques.

L'aspect du nom de domaine attribué est surtout symbolique. Néanmoins, ce symbolisme est puissant et l'intransigeance de l'ICANN à l'endroit des noms de domaine multilingues a mené à la perception globale que l'organisme se préoccupe peu de l'internationalisme ou de la diversité linguistique. Même si ICANN a récemment subi une réforme en profondeur et que son conseil d'administration se veut désormais plus international, il a perdu une grande partie de la confiance publique sur la question des noms de domaines multilingues, et il n'est pas très clair si ces changements permettront un système DNS équitable, fonctionnel et international, ou si la confiance perdue pourra être rétablie.

Le rôle des organisations telles ARIN, RIPE et APNIC de même que d'autres Centres d'information de réseaux (ou *Network Information Centers :* NIC) dans le biais linguistique émergeant est plus subtil que celui de ICANN.

Ces organisations, dont l'adhésion est relativement ouverte, régissent les interconnexions physiques des réseaux régionaux et locaux. L'une de leurs tâches principales est de maintenir l'espace adresse du protocole Internet (IP). Les numéros IP sont des numéros de 32 bits servant à identifier personnellement chaque hôte. Comme les noms de domaines, les numéros IP sont assignés par l'entremise d'un processus de délégation à des intermédiaires, pouvant à leur tour déléguer l'autorité. Mais contrairement aux noms de domaines, chaque plage assignée correspond à une branche physique du réseau, dont l'équipement associé est utilisé par une seule autorité. Les numéros IP sont attribués en plages, et puisque l'espace adresse est éventuellement limité, chaque attribution possède ses limites – les mêmes chiffres ne peuvent être assignés ailleurs plus tard, à moins de manipuler cette partie du réseau.

Les rapports entre le rôle des NIC et les enjeux liés à la diversité linguistique sont attribuables à leur fonction en tant qu'autorités régionales. Les ressources de réseaux disponibles dans un pays ou un groupe linguistique particulier dépendent des plages de numéros IP disponibles à l'autorité régionale pertinente, et leur allocation à d'autres groupes et pays. Une mauvaise allocation d'adresses ou une plage réduite de l'espace disponible dès le départ sont deux conditions pouvant mener à une pénurie d'adresses pour les nouveaux hôtes. La controverse a fait rage, à savoir si l'APNIC, dont les responsabilités régionales incluent l'Océanie, l'Asie de l'Est et du Sud-est, avait suffisamment d'espace pour continuer d'attribuer des plages IP au taux nécessaire. L'APNIC nie l'existence du problème, mais le spectre d'une crise est préoccupant. On prévoit d'améliorer les problèmes d'espace adresse par la mise à niveau de la version IP 4 (IPv4) à la version IP 6 (IPv6), qui utilise une plage plus étendue de numéros d'adresses, mais cette conversion nécessitera plusieurs années en raison des incompatibilités techniques avec IPv4.

Néanmoins, l'attribution de l'espace adresse IPv4 est très inefficace. De larges plages d'espace adresse sont désignées à des fins spéciales ou entièrement inutilisables ; on les appelle « bogons » et l'on conserve soigneusement ces plages afin que les administrateurs de systèmes puissent les surveiller à des fins sécuritaires (voir http://www.cymru.com/Bogons/). Même lorsque les plages de bogons sont masquées, un échantillon au hasard de 1 107 numéros IP a retourné 203 numéros IP (18 %), apparemment alloués pour l'essai d'un protocole « multidiffusion » rarement employé. En d'autres mots, 18 % de l'espace adresse IP globalement disponible était bloqué et inutilisable en raison d'une attribution inefficace. Dans la mesure où de telles inefficacités peuvent survenir, et qu'elles nuisent à l'espace

adresse disponible aux autorités régionales, les groupes linguistiques locaux pourraient se voir privés de ressources Internet. Pour que les différentes langues aient une chance raisonnable d'être utilisées en ligne, l'administration et l'attribution de l'espace adresse Internet doivent aussi être équitables.

Les gouvernements nationaux peuvent jouer un rôle à la fois favorable et défavorable pour influencer les biais linguistiques sur Internet. Dans la mesure où les gouvernements nationaux appliquent les politiques à l'intérieur de leurs frontières en vue de protéger et de promouvoir les droits linguistiques de leurs citoyens multilingues (Skutnabb-Kanngas et Phillipson, 1995), les biais linguistiques préexistant dans ces pays sont freinés. Dans la mesure où leurs politiques linguistiques sont appliquées dans des domaines pertinents de la politique d'information, elles favorisent la diversité linguistique sur Internet. Mais les gouvernements sont généralement plus préoccupés par l'efficacité administrative et les risques du séparatisme, et beaucoup de gens à travers le monde vivent sans garantie pour leurs droits linguistiques les plus élémentaires. Quand des pays ne connectent avec Internet au niveau international et demandent la conformité à leurs langues nationales, ils favorisent les biais émergents à l'endroit de leurs propres minorités ethnolinguistiques, faisant peu en bout de ligne pour favoriser la cause de la diversité linguistique en ligne. Si les groupes linguistiques nationaux espèrent occuper leur propre niche dans l'ethnosphère des télécommunications globales, ils doivent donc reconnaître et s'attaquer à la diversité linguistique à l'intérieur de leurs frontières nationales. Plus spécifiquement, ils doivent s'efforcer d'informer les citoyens de tous les groupes linguistiques sur l'alphabétisation numérique requise pour participer pleinement sur Internet. La prise de conscience ethno-linguistique des sociétés de télécommunications, d'informatique, ainsi que les autorités régissant Internet ne se développera que si une masse critique de groupes ethno-linguistiques sous-représentés réussit à attirer leur attention. Ceci risque peu de survenir, si la portée véritable de la diversité linguistique reste sous-évaluée.

Le biais linguistique émergent est un domaine de préoccupation significatif pour la diversité linguistique sur Internet. Les aspects discutés ici ne sont que des exemples, et non une liste complète des biais émergents éventuels. Avec l'évolution des marchés des télécommunications, du matériel informatique et des logiciels, tout comme celui des autorités régissant Internet, de nouveaux biais linguistiques peuvent survenir. Puisqu'ils découlent des contextes particuliers de la technologie et de l'utilisation de la langue, les biais linguistiques émergents peuvent aussi être de portée très locale, et se manifester de façon particulière seulement dans un pays

donné. Par conséquent, l'enjeu général du biais linguistique émergent exige une surveillance étroite aux niveaux international, régional et local.

Sources de biais techniques

Trois domaines de biais techniques, ayant différents rapports à la diversité linguistique, sont pertinents aux efforts actuels d'internationalisation en vertu des trois orientations concrètes de l'UNESCO mentionnées plus haut. D'abord, on retrouve l'aspect des normes de codage, directement relié à l'orientation concrète numéro 10, favorisant la diversité linguistique et culturelle sur Internet. Les codages de textes sont les principaux moyens techniques d'obtenir la diversité linguistique avec cet outil de communication surtout textuel. De plus, on retrouve l'aspect des langages de balisage et de programmation servant à créer et maintenir les applications et le contenu Internet. Ces systèmes techniques portent directement sur l'orientation concrète numéro 9, favorisant l'alphabétisation numérique. Si celle-ci requiert l'alphabétisation dans une autre langue comme pré-requis, tant l'ouverture que l'accès universels ne sont pas assurés. Et finalement, on retrouve les aspects du biais linguistique technique dans les protocoles d'application d'Internet, relatifs aux orientations concrètes 9 et 10. Pour favoriser l'accès aux technologies de l'information dans les pays en voie de développement, les principales applications Internet (courrier électronique, navigation hypertexte, messagerie instantanée, etc.) devraient permettre d'utiliser les langues des pays concernés. Le cas contraire, les embûches à l'acceptation de la technologie peuvent s'avérer prohibitifs. Ces trois domaines de biais techniques sont discutés ci-dessous.

Codage

Les codages précisent l'attribution arbitraire de chiffres aux symboles des langues écrites. Deux codages différents peuvent s'avérer incompatibles en assignant le même chiffre à deux symboles distincts, ou vice versa. Afin de profiter de l'avantage de la capacité informatique à manipuler les textes (par ex. affichage, modification, tri, recherche et transmission efficace), les communications d'une langue donnée doivent s'exprimer sous une forme quelconque de codage. Ainsi, ce qu'Internet peut vraiment offrir en termes de diversité linguistique se résume aux codages textuels disponibles.

Le codage le plus couramment utilisé est l'American Standard Code for Information Interchange (ASCII), un code mis au point durant les années 50 et 60 sous la direction de l'American National Standards Institute (ANSI) afin de standardiser la technologie des téléscripteurs. Ce codage comprend 128 attributions de caractères et convient surtout à l'anglais nord-américain. Puisqu'il fut développé tôt et adopté à grande échelle, la plupart des codages subséquents ont été définis en fonction d'ASCII, notamment l'ISO-8859-1 de l'Organisation internationale de normalisation (aussi appelé Latin-1) qui spécifie 256 codes dont les premiers 128 codes sont identiques à ASCII. Unicode, qui vise à fournir des codages compatibles pour toutes les langues à travers le monde (Consortium Unicode 1991, 1996, 2000, 2003), retient une stratégie semblable en faisant en sorte que les 256 premiers caractères des 65536 caractères du Basic Multilingual Plane (BMP) sont identiques à ISO-8859-1. La plupart des technologies de soutien Internet reposent sur l'ASCII et ses dérivés. Des systèmes tels DNS, Usenet news et Internet Relay Chat ne permettent d'utiliser qu'un sous-ensemble des caractères ASCII. Les systèmes d'exploitation tels que Linus reposent largement sur les « fichiers textuels plats ASCII » pour certaines de leurs fonctions les plus élémentaires. Tous ces systèmes comportent un biais technique favorisant l'anglais.

L'acceptation éventuelle d'Unicode constitue l'espoir le plus sérieux d'internationaliser l'infrastructure d'Internet. Les efforts de standardisation ont été entrepris par le Consortium Unicode, en collaboration avec ISO. Les adhérents au Consortium Unicode sont d'importants vendeurs de logiciels, des groupes religieux internationaux, des organisations régionales vouées à l'éducation, ainsi que des gouvernements nationaux. La norme Unicode (maintenant à sa version 4.0) comporte plus d'un million de codes de caractères possibles, permettant d'utiliser toutes les langues modernes et anciennes dans un seul texte. Le basic multilingual plane (BMP) comprend soixante-cinq mille caractères, ce qui devrait suffire à la plupart des communications écrites. Mais une telle souplesse d'utilisation comporte des limites. Dans sa forme la plus élémentaire, UTF-32, le texte Unicode exige quatre fois plus d'espace qu'en format ASCII. Beaucoup de développeurs de logiciels soutiennent que les usagers n'accepteraient pas cet inconvénient pour les textes multilingues, surtout si l'ordinateur est principalement utilisé en contexte monolingue.[24] Unicode offre d'autres codages de longueur variable plus efficaces,

24 À savoir s'il s'agit de la vérité est une question importante qui n'a pas été abordée de façon satisfaisante dans la littérature de recherche.

mais les inconvénients s'appliquent aux textes n'étant pas en caractères romains, qui doivent occuper plus d'espace. Même si les coûts de stockage de données ont largement diminué au cours de la dernière décennie (suffisamment pour qu'Unicode soit moins problématique), le traitement d'Unicode continue de compliquer significativement la tâche des développeurs de logiciels, puisque la plupart des applications exigent une interaction avec ASCII. De plus, les formats plus gros de documents Unicode comportent des coûts de transmission, de compression et de décompression, qui constituent un inconvénient suffisant pour décourager les usagers d'Unicode dans certains cas.

Même si Unicode a permis des progrès importants pour l'internationalisation de l'informatique, les problèmes liés au texte multilingue sur Internet sont loin d'être résolus. Pour différentes raisons d'ordre technique, économique et organisationnel, le développement d'une norme technique acceptable s'est fait plus lentement que celui d'Internet lui-même. Par conséquent, le recours international à Internet a privilégié les langues basées sur l'alphabet romain et surtout l'anglais, qui a profité d'un codage standard largement reconnu avant même la popularité d'Internet. Pour qu'Internet permette l'usage équivalent de toutes les langues à travers le monde, il faudra qu'Unicode soit plus répandu. Comme c'est le cas pour le système DNS, il faudra peut-être mettre à niveau certains protocoles Internet, afin qu'ils fonctionnent conjointement avec Unicode.

Langages de balisage et de programmation

Les « codes » informatiques – les langages de balisage et de programmation – servant à configurer le contenu et les services Internet constituent un autre biais technique favorable à l'anglais et perpétué sur Internet. Le soutien au contenu multilingue constitue le premier biais technique le plus évident. Les langages de balisage tels que le langage de balisage hypertexte (HTML) et le langage de balisage extensible (XML) doivent être en mesure de décrire le texte dans une gamme complète de langues. Le World-Wide Web Consortium a stipulé ceci en exigeant le soutien Unicode dans le cadre de ses normes. Ce qui signifie que lorsque le soutien Unicode est déficient, comme c'est le cas avec la plupart des langues de l'Asie de l'Ouest, du Centre-Sud et du Sud-est, le soutien HTML et XML est aussi déficient. De la sorte, le biais envers certaines langues s'avère uniforme pour cette raison. Les langages de programmation doivent aussi devenir compatibles avec le texte multilingue. Malheureusement, plusieurs langages de programma-

tion couramment employés, tels que le langage C, n'offrent pas encore le soutien Unicode.[25] Un nombre croissant de langages conçus pour des applications Web le font (notamment Java, JavaScript, Perl, PHP, Python et Ruby, qui sont tous largement adoptés), mais le soutien des autres systèmes tels les logiciels de bases de données s'oriente davantage envers Unicode. La promesse du commerce électronique dans d'autres langues que l'anglais sous-entend que les bases de données conformes à Unicode deviendront très répandues.

Le biais en faveur de l'anglais se constate aussi dans la conception même des langages de balisage et de programmation. Les langages de programmation constituent l'interface humaine la plus élémentaire pour le contrôle informatique, agissant comme intermédiaire entre les processus cognitifs des programmeurs et les capacités logiques des ordinateurs. Une surabondance de langages de programmation existe ; les estimations vont de 2 500 à plus que le nombre de langues dans le monde. Mais en dépit de cette diversité apparente, la grande majorité des langues tracent ultimement leur origine au FORTRAN, le premier langage de programmation de haut niveau développé en 1957 par IBM (Lévénez, 2003). Ces langages ont largement recours aux mots anglais pour définir d'importantes constructions de programmation, tels les conditionnels (*if, then, else, case,* etc.) et le bouclage interactif (*while, for, until,* etc.). Même si beaucoup de langues possèdent des équivalents pour ces mots, ils ne semblent jamais se substituer aux mots anglais en code exécutable. Par exemple, Ruby, conçu par le programmeur japonais Yukihiro Matsumoto avec un souci de l'internationalisation, a aussi recours aux mots anglais.[26]

HTML et XML sont semblables à cet égard. Les balises HTML sont généralement des abréviations mnémotechniques de mots anglais (par ex. b "bold", ul "unordered list", li "list item", etc.). Même si XML n'est pas un langage de balisage en soi, il s'agit d'une syntaxe pour définir les langages de balisage et tous les langages de balisage à base XML reconnus sont basés sur l'anglais (par ex. MathML, pour les expressions mathématiques, et XML:FO pour le formatage de documents textuels), malgré le fait que la norme XML soit basée sur Unicode. Cette tendance s'est poursuivie avec le projet de développement du Web séman-

25 Le site Web de International Components for Unicode (ICU) offre une bibliothèque C libre accès qui aide au soutien Unicode (http://oss.software.ibm.com/icu/).

26 Voir http://www.ruby-lang.org/ja/uguide/uguide03.html, contenant un échantillon de programme de Rudy intégré à une page de texte japonais utilisant trois autres systèmes d'écriture.

tique (Semantic Web), visant à fournir un raisonnement « connu de tous » sur le Web. On prévoit avoir recours à d'importantes bases de données d'intelligence artificielle telles que Cyc (Reed et Lenat, 2002) et WordNet (Fellbaum et Miller, 1998) afin de développer de nouveaux balisages qui aideront les programmes Internet à trouver et à traiter l'information pour les usagers. Ces bases de données ont déjà été critiquées dans une perspective culturelle de l'hémisphère Nord comme comportant des biais sexistes et androcentriques (Adam, 1998). En outre, elles comportent sûrement aussi des biais culturels. En ce sens, des projets tels le Web sémantique, qui promettent de fournir la « prochaine génération » de services d'information Internet, menacent de renforcer encore davantage les biais linguistiques et culturels déjà existants.

Il faut tenir compte du potentiel de biais linguistique dans les langages de programmation et de balisage, tout comme de la nature culturelle du calcul informatisé. Le calcul informatisé moderne dérive de plusieurs siècles d'apprentissage mathématique, et sa diffusion actuelle est comparable à celle du système des nombres décimaux, tant par sa nature que son importance. L'invention dans le nord de l'Inde des nombres décimaux date environ du 7e siècle après J.C. et s'est répandue partout, remplaçant la plupart des autres systèmes numériques. Toutefois, la diffusion culturelle des nombres décimaux n'exigea pas l'importation du vocabulaire ; plusieurs langues modifièrent toutefois leurs vocabulaires numériques existants afin d'intégrer cette nouvelle pratique. L'informatique développe davantage le principe de nombres décimaux en automatisant leur traitement. Cependant, contrairement à la diffusion des nombres décimaux, la popularité des ordinateurs s'est accompagnée de vocabulaires anglais lourds et complexes – les langages de programmation.

Il ne fait aucun doute qu'en tant qu'artefact physique, l'ordinateur joue un rôle dans ce rapport en associant les symboles aux actions. Le couplage exact des symboles et des actions reste arbitraire, de sorte que tout langage pourrait être utilisé, mais s'avère aussi suffisamment complexe qu'y parvenir n'est pas évident. En ce sens, une vaste question pour la diversité linguistique n'a pas été adéquatement posée dans la littérature de recherche : dans quelle mesure les différentes caractéristiques des langages de programmation facilitent-elles leur acquisition et leur utilisation par les personnes parlant diverses langues ?[27] Les effets du transfert chez une personne parlant une langue et qui en apprend une autre sont bien

27 Voir Anis (1997) pour des suggestions en ce sens.

connus. On pourrait supposer que les langages de programmation, étant en soi des systèmes linguistiques formels, pourraient faire l'objet d'un transfert semblable menant à des difficultés ou des erreurs systémiques chez les personnes de diverses origines linguistiques. Les propriétés conceptuelles des langages de programmation varient grandement. Est-il possible que les personnes parlant une certaine langue soient mieux servies par des langages de programmation dont les caractéristiques correspondent à leur propre langue ? Les langages de programmation pourraient possiblement être conçus pour refléter le raisonnement de différentes traditions culturelles et linguistiques. De telles adaptations aideraient-elles ces gens à contrôler leurs propres ressources en technologie de l'information ?

L'UNESCO et les autres agences des Nations Unies ont un besoin pressant d'obtenir des réponses à ces questions, en vue d'atteindre les objectifs éducatifs requis pour favoriser la diversité linguistique. Grâce à la programmation informatique, la langue devient puissante et animée, ayant le potentiel de redéfinir les cultures. Malheureusement, c'est surtout l'anglais qui est présentement animé de cette façon. Si l'alphabétisation numérique des langages de programmation informatique exige la connaissance linguistique ou culturelle de l'anglais, les personnes parlant d'autres langues doivent ultimement porter le lourd fardeau des coûts éducatifs et possiblement culturels afin de s'approprier les ressources d'information sur Internet.

Modes de communication

Même si la plupart des gens connaissent Internet par l'entremise du Web (certains croient qu'ils sont synonymes), il s'agit en fait d'un environnement plus hétérogène offrant une variété de modes de communication. De plus, de par sa conception, Internet permet la création et le déploiement à peu de frais de nouveaux modes de communication. Alors que nous utilisons à l'heure actuelle le courrier électronique, le Web et les messages instantanés sur Internet, nous ignorons tout des utilisations éventuelles dans un avenir rapproché. Certains modes de communication sont néanmoins devenus largement répandus, et il arrive qu'ils intègrent des formes techniques de biais linguistique.

L'un de ces modes de communication est Usenet News, d'abord créé en 1978 pour mettre en réseau les systèmes informatiques de trois universités (Spencer et Lawrence, 1998). Usenet regroupe des centaines de « forums » *(newsgroups)*,

des espaces de messages publics dont les noms suggèrent un contenu local. Le serveur et le logiciel client de Usenet sont accessibles gratuitement, et sa gestion est largement ouverte. Les administrateurs de Usenet peuvent régler individuellement la quantité, le taux et la fréquence du partage des messages avec d'autres serveurs, de façon à optimiser facilement l'accès au réseau dans les régions à faible connectivité. De la sorte, les obstacles pour accéder à Usenet sont relativement faibles. Usenet constitue une ressource extrêmement importante à l'échelle internationale. En 1999, 205 pays à travers le monde avaient accès à Usenet (Smith, 1999).

Sur le plan technique, Usenet représente un microcosme d'Internet. Sa séquence de d'attribution de noms des forums est hiérarchique et a recours à un sous-ensemble d'ASCII, tout comme pour le système DNS. Usenet possède des hiérarchies de premier niveau, ainsi que des hiérarchies locales, régionales et nationales.[28] Les messages textuels doivent rester compatibles avec ASCII. Les textes chinois et japonais ont recours à des codages spéciaux sur Usenet. Comme ailleurs sur Internet, l'anglais a primauté dans les hiérarchies génériques de premier niveau. Par exemple, dans la hiérarchie comp., la catégorie générique servant à l'affichage de systèmes informatiques, on retrouve peu, sinon aucun affichage en japonais, même sur comp.lang.ruby. C'est seulement sur la hiérarchie fj.comp que l'on retrouve des discussions techniques et scientifiques sur l'informatique en japonais. La sous hiérarchie soc.culture fournit aussi de l'espace pour le trafic multilingue, mais surtout dans les langues européennes. Ainsi, en dépit de son faible coût d'accès pour les pays ayant des ressources très limitées, Usenet est faiblement internationalisé et comporte beaucoup de biais techniques favorisant l'anglais, dont certains entraînent d'autres biais émergents.

Un autre mode de communication devenu populaire au début des années 90 est le service de clavardage IRC *(Internet Relay Chat),* un mode de communication synchrone multipartite en temps réel. Les participants sur un canal de clavardage communiquent entre eux en temps réel, un peu comme lors d'une conférence téléphonique, à l'exception que la conversation est enregistrée. Les serveurs IRC en réseau peuvent héberger ces milliers de canaux et il est fréquent de retrouver sur les réseaux IRC tels EFNet ou UnderNet des canaux de clavardage abordant des thèmes culturels, régionaux ou nationaux, et d'attirer des participants de partout à travers le monde (Paolillo, 2001). Le service de clavardage IRC provient du nord de l'Europe, de sorte que certaines caractéris-

28 L'espace nom Usenet, tout comme l'espace nom DNS, a aussi fait l'objet d'abus sérieux.

tiques – notamment les caractères attribués dans les messages textuels ainsi que les noms des participants – diffèrent de ceux de Usenet. Toutefois, le soutien au texte multilingue n'est pas meilleur avec IRC qu'avec Usenet. Dans les faits, les différences d'affichage entre les ordinateurs utilisant l'anglais américain et ceux du nord de l'Europe causent des problèmes évidents (par exemple, la substitution de caractères de ponctuation en faveur des caractères à voyelle diacritique dans les noms et les mots scandinaves).

Ainsi, en dépit de l'attrait de ces deux systèmes sur le plan international, ils comportent des défauts provenant des biais linguistiques découlant de leur conception même. Évidemment, les nouveaux modes de communication tels la messagerie instantanée, le blogage, le clavardage et autres apparaissent constamment. Même si certains de ces modes de communication comportent des caractéristiques de conception particulières tels XML et Unicode, le stade de développement de ces normes est tel que seulement une faible partie de la population mondiale et des langues à travers le monde bénéficient de ces technologies à l'heure actuelle. Certains partisans de la technologie peuvent espérer encore d'autres protocoles de communications, telles que la voix sur IP, ou les interfaces multimodes. Même si ces technologies parviennent à résoudre certains enjeux linguistiques, d'autres se poursuivront, comme l'assistance aux personnes aveugles ou malentendantes. De plus, les biais techniques déjà existants renforcent les biais émergents associés à la démographie, à l'économie et autres. Afin de minimiser les biais linguistiques sur Internet, on devrait examiner de près les nouveaux modes de communications pour découvrir tout biais technique potentiel avant de permettre leur adoption à grande échelle.

Beaucoup de technophiles ont exprimé l'espoir que la traduction automatique soit la réponse aux problèmes de communications multilingues sur Internet. Les services de traduction offerts par des sociétés comme Systran, le fournisseur du système de traduction BabelFish, sont très en demande et dans certains cas, notamment du catalan à l'espagnol, la traduction automatique a été suggérée comme la réponse qui s'impose aux problèmes de communication (Climent et autres, 2004). Les gens pourront-ils un jour accéder à Internet dans leur propre langue, en recourant tout simplement à l'un des systèmes de traduction en ligne ? Cette question s'avère trop optimiste pour plusieurs raisons.

En premier lieu, un système de traduction automatique assume que les problèmes plus courants de représentation et de formulation du texte dans la

langue sont déjà réglés, alors que pour beaucoup de langues, ce n'est pas le cas. Ensuite, la conception même du système de traduction automatique nécessite énormément de travail. Des problèmes particuliers peuvent survenir en traduction entre des combinaisons de deux langues, qui doivent être résolus pour ces langues seulement. La traduction des différentes langues à travers le monde s'avère donc un défi qui n'est pas près d'être résolu dans un avenir rapproché. De plus, la conception des systèmes de traduction automatique exige de grandes quantités de contenus dans les langues à traduire[29] ; à l'heure actuelle, ces contenus sont recueillis sur les sites Web des langues à traduire (Grefenstette, 1999 ; Resnik, 1999), et doivent donc être créés par des locuteurs natifs. Ceci ne peut survenir à moins d'un soutien technique adéquat pour la langue. En dernier lieu, la traduction automatique n'est jamais de qualité équivalente à celle produite par un traducteur humain (Kay et autres, 1993). Les utilisateurs de systèmes de traduction automatique doivent adapter les restes inappropriés de vocabulaire et de séquences de mots qui ne représentent qu'une forme indirecte du biais linguistique qui a nécessité la traduction en premier lieu. Par conséquent, nous ne pouvons espérer que l'approche technologique d'un système de traduction automatique réduira les problèmes de biais linguistique de façon substantielle sur Internet.

Conclusions

L'exploration des sources potentielles de biais effectuée dans la discussion qui précède constate plusieurs sources de biais linguistique sur Internet, tant préexistant que technique ou émergent. Par conséquent, la réponse à la question posée, à savoir s'il existe un biais linguistique sur Internet, ne peut être qu'affirmative. Les principales conséquences des biais sont de favoriser les langues très répandues, ayant des normes techniques bien définies. Il est à noter que l'anglais est probablement la première de ces langues en s'avérant – non par coïncidence – la langue des inventeurs d'Internet et des projets de recherche précédents. Toutefois, il est aussi évident que les causes et les effets des biais sont subtils, diversifiés et dans beaucoup de cas, imprévisibles. Si l'UNESCO désire sérieusement s'attaquer au biais linguistique sur Internet, il faut faire davantage pour s'informer, tout comme les principaux agents de développement sur Internet, des

29 On ne peut utiliser n'importe quel texte. Règle générale, des textes bilingues alignés phrase par phrase sont requis. La préparation est coûteuse et non disponible pour toutes les combinaisons de deux langues.

manifestations des biais linguistiques ainsi que de l'importance et de l'intérêt de la diversité linguistique.

Glossaire

ACM. Association for Computing Machinery. Le plus important regroupement international de professionnels de l'informatique. L'ACM comprend plusieurs groupes d'intérêts spéciaux actifs sur les aspects techniques, sociaux et de politiques des réseaux informatiques et d'ordinateurs.

APNIC. Centre d'information du réseau Asie-Pacifique *(Asia-Pacific Network Information Center)*. Le Centre d'information de réseau supervise le fonctionnement d'Internet en Asie et dans le Pacifique. Ses activités s'étendent à l'Australie, Chine, Japon, Corée, Indonésie, Malaisie ainsi que toutes les îles indépendantes du Pacifique.

ARIN. American Registry for Internet Numbers. Centre d'information de réseau supervisant le fonctionnement technique d'Internet en Amérique du Nord.

ASCII. Code standard américain pour l'échange d'information *(American Standard Code for Information Interchange)*. Une des premières normes à sept bits pour le codage textuel informatisé et supportée de façon omniprésente par la plupart des applications informatiques. La plupart des codages textuels modernes, y compris Unicode, sont conçus pour être rétrocompatibles avec ASCII, dont les sept bits permettent le codage de 128 caractères distincts. L'ASCII étendu *(Extended ASCII)* est un prolongement à huit bits de l'ASCII qui ne possède aucune norme. Différents marchands supportent différentes versions de l'ASCII étendu qui sont mutuellement incompatibles.

Biais émergent. Pour Friedman et Nissenbaum (1997), biais résultant de l'interaction des systèmes techniques dans des contextes sociaux particuliers.

Biais préexistant. Pour Friedman et Nissenbaum (1997), tout biais entraîné par des causes exclusivement sociales, antérieures à l'application particulière d'une technologie où le biais se manifeste.

Biais technique. Pour Friedman et Nissenbaum (1997), tout biais inhérent à un système technique. Le biais de l'ASCII en faveur de l'anglais américain constitue un exemple de biais technique.

BMP. Basic Multilingual Plane. Partie des valeurs du code Unicode comprenant les codes de tous les caractères d'écriture les plus couramment utilisés à travers le monde.

ccTLD. Domaine de premier niveau de code de pays *(Country-Code Top-Level Domain)*. Domaines de premier niveau associés à des pays particuliers. Les ccTLD sont semblables aux codes de pays ISO-3166. Par exemple, .uk (Royaume-Uni) et .za (Afrique du Sud).

CMC. Communication assistée par ordinateur *(Computer-Mediated Communication)*. Communication entre humains s'effectuant par le biais d'ordinateurs en réseaux.

CNNIC. Centre d'information du réseau chinois *(China Network Information Center)*. Centre d'information de réseau supervisant le fonctionnement technique d'Internet en Chine.

DNS. Système de noms de domaine *(Domain-Name System)*. Système technique administré par ICANN et permettant l'attribution de codes mnémotechniques aux ordinateurs en réseaux.

Domaine (nom). Nom enregistré dans le système des noms de domaine (DNS) et servant à référer à un ordinateur hôte Internet. Les noms de domaine sont attribués à des organisations pouvant à leur tour les assigner à des ordinateurs ou ensembles d'ordinateurs spécifiques, en collaboration avec les fournisseurs de services de réseau sous contrat.

Registre de noms de domaine. Organisation sous contrat avec ICANN et administrant certaines parties de l'espace nom DNS. Un registre est généralement en charge de l'entretien d'un ou plusieurs domaines de premier niveau (TLD). Verisign et Educause sont des exemples de registres de noms de domaine.

Ethnologue. Base de données maintenue par SIL International (Barbara Grimes, ed.) et enregistrant les données descriptives générales de toutes les populations linguistiques connues à travers le monde.

GPL. Licence GPL *(Gnu Public License)*. Licence régissant certains logiciels libres afin de protéger les droits d'auteur tout en permettant l'accès libre du code source du logiciel aux développeurs.

gTLD. Domaine générique de premier niveau *(Generic Top-Level Domain)*. Domaine de premier niveau assigné à des fins « génériques » sans nécessairement référer à un pays en particulier. Parmi les gTLD connus, on retrouve .com, (commercial) .edu (éducation supérieure accréditée aux É.-U.), .mil (armée américaine), .net (fournisseurs de réseaux), .org (organisations sans but lucratif), etc.

Hôte, hôte Internet. Tout ordinateur relié à Internet.

HTML. Langage de balisage hypertexte *(Hypertext Markup Language)*. Langage de balisage permettant le formatage de pages Web. Langage simple bien compris par les internautes et les logiciels, et dont la norme est maintenant maintenue par le consortium W3C.

IANA. *(Internet Assigned Numbers Authority)*. Branche d'Internic, anciennement responsable de l'inscription des nouveaux sites au réseau Internet.

ICANN. *(Internet Corporation for Assigned Names and Numbers)*. Organisation en partenariat public privé qui supervise le système DNS.

Internet. Réseau informatique international résultant de la liaison de ARPANET aux autres réseaux informatiques régionaux.

IPv4. IP version 4. Version IP la plus couramment utilisée à l'heure actuelle, et caractérisée par les numéros d'adresse de 32 bits pour chaque hôte Internet. L'espace adresse sous Ipv4 est limité, en ce sens qu'Internet passe présentement de la version IPv4 à IPv6.

IPv6. Version 6 du protocole IP. Cette version de « nouvelle génération » du protocole Internet a recours à des adresses de 128 bits. Le soutien à Ipv6 s'étend à plusieurs applications en réseau, mais son déploiement reste pour l'instant limité, puisque les applications IPv4 sont incompatibles avec les hôtes IPv6.

IRC. Service de clavardage sur Internet *(Internet Relay Chat)*, un protocole d'application permettant des communications simultanées, en temps réel, entre plusieurs

internautes sur Internet. La plupart des programmes de « clavardage », y compris plusieurs programmes propriétaires, sont largement inspirés d'IRC. On compte beaucoup de réseaux IRC utilisés surtout à des fins personnelles par des millions d'internautes à travers le monde.

ISO-8859-1, Latin-1. Codage textuel standard de huit bits supportant la plupart des langues européennes dérivées de l'alphabet romain.

Langage de balisage. Système destiné à introduire le formatage ou autres codes (« balisage ») dans des documents textuels, de façon à formater ou interpréter le texte avec un appareil comprenant le balisage. HTML est un exemple de langage de balisage, mais d'autres langages comme SVG (Scalable Vector Graphics) fonctionnent de façon semblable tout en effectuant des fonctions différentes. Voir XML.

Localisation linguistique *(localisation)*. La localisation est l'adaptation culturelle d'un produit ou d'un service pour assurer son respect des exigences légales et socioculturelles spécifiques à un marché cible. La localisation implique l'adaptation d'un produit à un marché spécifique lors d'un processus qui va bien au-delà de la traduction classique et qui prend en considération les usages contemporains et familiers d'une langue et les nuances culturelles, telles que les règles de notation et les différences de signification de symboles, d'associations de couleurs et d'options de paiement.

NIC. Centre d'information de réseau *(Network Information Center)*. Organisation technique chargée de superviser le fonctionnement technique d'Internet sur le plan régional ou local. On retrouve trois principaux Centres d'information de réseau régionaux : ARIN, RIPE et APNIC, respectivement pour l'Amérique du Nord, l'Europe et l'Asie.

Protocole. Ensemble de messages et de règles standardisées d'échange de messages entre ordinateurs en réseau. Les protocoles sont complexes et sont généralement mentionnés en terme de « couches » : la couche d'application, la couche de liaison, etc.

Protocole d'application. Protocole de réseau habituellement employé par un usager de l'ordinateur. Les protocoles d'application servent généralement à des fins particulières sur le réseau, notamment l'échange de fichiers ou de courrier entre les ordinateurs.

Protocole IP. Protocole Internet *(Internet Protocol)*. Voir TCP/IP.

RIPE. Réseaux IP Européens. Centre d'information de réseau supervisant le fonctionnement technique d'Internet en Europe.

SGML. Langage général de balisage *(Standard Generalized Markup Language)*. Langage de définition du langage de balisage et normalisé dans le domaine de l'imprimé. Le HTML fut développé à l'origine en tant qu'application SGML.

TCP/IP. Protocole de gestion de transmission/protocole Internet *(Transmission Control Protocol/Internet Protocol)*. Principal ensemble de protocoles servant au fonctionnement d'Internet. TCP et IP sont des « couches » indépendantes de protocoles de réseautage Internet qui concernent différents aspects du fonctionnement du réseau, mais utilisées conjointement le plus souvent.

Vitalité ethnolinguistique. Le potentiel de survie d'une communauté ethnolinguistique.

Vitalité technolinguistique. Potentiel d'une communauté ethnolinguistique à profiter des technologies, surtout celles liées à l'information, et d'utiliser sa langue avec ces technologies. En analogie à la vitalité ethnolinguistique.

TIC (ICT). Technologie d'information et de communication. Toute technologie servant à traiter ou à transmettre l'information.

TLD. Domaine de premier niveau *(Top-Level Domain)*. Nom de domaine directement attribué par ICANN à un registre de nom de domaine regroupant plusieurs hôtes reliés, généralement par pays ou à des fins organisationnelles.

Consortium Unicode. Consortium supervisant le développement de Unicode.

Unicode. Codage de caractères de 64 bits actuellement en développement, et visant à fournir un outil technique standard pour représenter les caractères de toutes les langues écrites au monde. Unicode est développé en collaboration avec l'Organisation internationale de normalisation (ISO) et le consortium W3C, afin d'assurer que les normes de ces trois organisations seront compatibles.

Réseau **Usenet (nouvelles).** Application d'échange de messages (« nouvelles ») à affichage public et à grande diffusion parmi les internautes en réseaux. Définit

aussi toutes les nouvelles ou l'ensemble des nouvelles échangées de cette façon. Usenet est important pour Internet, puisqu'il s'agit d'un protocole à faible coût, facilement implanté, pouvant servir au courrier électronique et ne requérant aucune connexion de réseau à cette fin. En ce sens, c'est souvent la première application Internet à atteindre un nouvel emplacement.

UTF-8, UTF-16, UTF-32. Codages de caractères Unicode recourant à des unités de 8, 16 et 32 caractères respectivement. UTF-8 et UTF-16 sont des codes de largeur variable, en ce sens que certains caractères exigent plus qu'une unité de 8 ou 16 bits pour le codage. UTF-32 est un code de largeur fixe, en ce sens que tous les caractères permettent le codage à 32 bits.

W3C. Consortium World-Wide Web. Consortium supervisant le développement de protocoles, langages de balisage et autres normes techniques se rapportant au Web.

World-Wide Web. (« le Web ») Application servant à échanger des documents, programmes et contenus multimédias formatés sur Internet. Définit aussi l'ensemble des documents et le contenu disponible par le truchement de la Toile. Le Web est l'application la plus connue d'Internet, en raison de la facilité avec laquelle le navigateur Web effectue des recherches de documents et autres contenus.

XML. Langage de balisage extensible *(Extensible Markup Language)*. Langage de définition du langage de balisage, une version simplifiée de SGML, visant à fournir de l'information sur la Toile plus adaptée que HTML, et permettant de définir plusieurs types de balisages. Les langages de balisage actuels définis dans XML incluent ceux pour le contenu Web (XHTML), les graphiques (Scalable Vector Graphics [SVG]), les équations mathématiques (MathML), la musique (MML, MusicML) et beaucoup d'autres applications.

REFERENCES

Adam, A. 1998. *Artificial Knowing: Gender & the Thinking Machine.* London: Routledge.

Anis, J. 1997. A Linguistic Approach to Programming.Arob@se, 1.2. http://www.liane.net/arobase

Androutsopoulos, J. 1998. Orthographic variation in Greek e- mails: a first approach. *Glossa* 46, S. pp. 49-67.

Barrera-Bassols, N. and Zinck, J.A. 2002. Ethnopedological research : a worldwide review. In *17th World congress of soil science CD-ROM proceedings: Confronting new realities in the 21st century.* 590.1-590.12. Bangkok: Kasetsart University. (http://www.itc.nl/library/Papers/arti_conf_pr/barrera.pdf).

Block, D. 2004. Globalization, transnational communication and the Internet. *International Journal on Multicultural Societies,* Vol. 6,No.1, pp.13-28.

Climent, S., J. Moré, A. Oliver, M Salvatierra, I Sànchez, M. Taulé and L. Vallmanya. 2004. Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation. *Journal of Computer-Mediated Communication,* Vol. 9, No.1. http://www.ascusc.org/jcmc/

Crystal, D. 2000. *Language Death.* Cambridge: Cambridge University Press.

—. 2001. *Language and the Internet.* Cambridge: Cambridge University Press.

—. 2003. *English as a Global Language, Second Edition.* Cambridge: Cambridge University Press.

Dalby, A. 2003. *Language in Danger.* New York: Columbia University Press.

Dunker, E. 2002. Cross-cultural usability of the library metaphor. *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries.* Portland, OR.

Durham, M. 2004. Language Choice on a Swiss Mailing List. *Journal of Computer-Mediated Communication* 9.1. http://www.ascusc.org/jcmc/

Fellbaum, C., and G. Miller. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Ferguson, C. A. 1959. Diglossia. *Word,* 15, pp.325-340.

Friedman, B. and H. Nissenbaum. 1995. Minimizing bias in computer systems. *Conference companion on Human factors in computing systems,* 444. ACM Press.

Friedman, B. and H. Nissenbaum. 1997. Bias in computer systems. In Friedman, B., ed. *Human Values and the Design of Computer Technology,* pp.21-40. Stanford, California. Cambridge ; New York, CSLI Publications; Cambridge University Press.

—. 1997. Self-presentation and interactional alliances in e-mail discourse: the style- and code-switches of Greek messages, *International Journal of Applied Linguistics* 7: pp.141-164.

Georgakopolou, A. (Forthcoming). On for drinkies? E-mail cues of participant alignments. In S. Herring (ed.), *Computer-Mediated Conversation.*

Global Reach. 1999-2005. Global internet statistics by language. Online marketing information. http://global-reach.biz/globstats/index.php3

Greenberg, J. 1956. The measurement of linguistic diversity. *Language,* Vol. 32, No.2, pp.109-115.

Grefenstette, Gregory. 1999. The WWW as a resource for example-based MT tasks. Paper presented at ASLIB "Translating and the Computer" conference, London.

Grimes, J. E. 1986. "Area norms of language size." In B.F. Elson, ed., *Language in global perspective: Papers in honor of the 50th anniversary of the Summer Institute of Linguistics*, 1935-1985, pp.5-19. Dallas: Summer Institute of Linguistics.

Hafner, K., and Lyon, M. 1996. *Where Wizards Stay Up Late: The Origins of the Internet.* New York: Simon and Schuster.

Hård af Segerstad, Y. 2002. Effects of Mobile Text Messaging on Swedish Written Language — human adaptability made visible. *International Conference on Cultural Attitudes towards Technology and Communication, The Net(s) of Power: Language, Culture and Technology*, Montréal.

Holmes, H. K. 2004. An analysis of the language repertoires of students in higher education and their language choices on the Internet (Ukraine, Poland, Macedonia, Italy, France, Tanzania, Oman and Indonesia). *International Journal on Multicultural Societies*, Vol. 6, No.1, pp. 29-52.

Ifrah, G. 1999. *The Universal History of Numbers: From Prehistory to the Invention of the Computer.* New York: John Wiley and Sons.

Information Sciences Institute. 2003. USC Researchers Build Machine Translation System — and More — For Hindi in Less Than a Month. http://www.usc.edu/isinews/stories/98.html

Kay, Martin, Jean-Mark Gawron, and Peter Norvig. 1993. *Verbmobil : A Translation System for Face-to-Face Dialog.* Stanford , CA: CSLI Publications.

Krauss, Michael. 1992. The world's languages in crisis. *Language* Vol. 68, No.1, pp. 4-10.

Koutsogiannis, D., and B.. Mitsikopolou. 2004. Greeklish and Greekness: Trends and Discourses of "Glocalness". *Journal of Computer-Mediated Communication* 9.1. http://www.ascusc.org/jcmc/

Lavoie, B. F. and E. T. O'Neill. 1999. How "World Wide" is the Web? Annual Review of OCLC Research 1999. 2003.

Lévénez, Eric. 2003. Computer languages timeline. http://www.levenez.com/lang/

Lieberson, S. 1964. An extension of Greenberg's linguistic di versity measures. Language, 40, pp.526-531.

Mafu, S. 2004. From oral tradition to the information era: The case of Tanzania. *International Journal on Multicultural Societies*, Vol.6, No.1, pp. 53-78.

Muhlhäusler, P. 1996. *Linguistic Ecology: Language Change & Linguistic Imperialism in the Pacific Rim.* London: Routledge.

Nettle, D. 1999. *Linguistic Diversity.* Oxford: Oxford University Press.

Nettle, D., and S. Romaine. 2000. *Vanishing Voices: The Extinction of the World's Languages.* Oxford: Oxford University Press.

Nunberg, Geoffrey. 1998. Languages in the Wired World. Paper presented at *La politique de la langue et la formation des nations modernes*, Centre d'Etudes et Recherches Internationales de Paris.

O'Neill, Edward T, Brian F. Lavoie, and Rick Bennett. 2003. Trends in the Evolution of the Public Web: 1998 - 2002. *D-Lib Magazine,* 9.4. http://www.dlib.org/dlib/april03/ lavoie/04lavoie.html

O'Neil, E.T. ; P.D. McClain; and B.F. Lavoie 1997. A methodology for sampling the World-Wide Web. Technical report, *OCLC Annual Review of Research.* http://www.oclc.org/oclc/research/publications/review97/oneill/o'neilla%r980213.html

Paolillo, J. C. 1996. Language Choice on soc.culture.Punjab. *Electronic Journal of Communication/ Revue Electronique de Communication,* 6(3). http://www.cios.org/

Paolillo, J. C. 2001. Language Variation in the Virtual Speech Community: A Social Network Appoach. *Journal of Sociolinguistics,* 5.2.

Paolillo, J. C. 2002. Finite-state transliteration of South Asian text encodings. In *Recent Advances in Natural Language Processing: Proceedings of the ICON International Conference on Natural Language Processing.* New Delhi: Vikas Publishing House, Ltd.

Paolillo, J. C. To appear, 2006. 'Conversational' code switching on Usenet and Internet Relay Chat. To appear in S. Herring, ed., *Computer-Mediated Conversation.* Cresskill, NJ: Hampton Press.

Peel, R. 2004. The Internet and language use: A case study in the United Arab Emirates. *International Journal on Multicultural Societies,* Vol. 6, No. 1, pp.79-91.

Phillipson, R. 1992. *Linguistic Imperialism.* Oxford: Oxford University Press.

Phillipson, R. 2003. *English-Only Europe?* London: Routledge.

Pimienta, D.; and B. Lamey. 2001. Lengua española y cultural hispanicas en la Internet: Comparació con el ingles y el frances. II Congreso Internacional de la Lengua Espanola, Valladolid, 16-19 October 2001.

Pimienta, D.; et al. 2001. L5: The fifth study of languages on the Internet. http://funredes.org/LC/english/L5/L5tendencies.html

Reed, S. L., and D. B. Lenat. 2002. Mapping Ontologies onto Cyc. American Association for Artificial Intelligence. http://www.aaai.org/

Resnik, P. 1999. Mining the Web for Bilingual Text. *37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), College Park, Maryland.

Rheingold, H. 2000. *The Virtual Community: Homesteading on the Electronic Frontier,* revised edition. Cambridge, MA: MIT Press.

Skutnabb-Kangas, T., and R.. Phillipson. 1995. *Linguistic Human Rights: Overcoming Linguistic Discrimination.* Berlin: Mouton de Gruyter.

Smith, E. A. 2001. On the co-evolution of linguistic, cultural and biological diversity. In L. Maffi, ed. *On Biocultural Diversity,* 95-117. Washington DC: Smithsonian Institution Press.

Smith, M. 1999. Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of USENET. In M. Smith and P. Kollock, eds., *Communities in Cyberspace.* London: Routledge Press.

Spencer, H. and Lawrence, D. 1998. *Managing Usenet.* Sebastopol, CA: O'Reilly.

Su, H.-Y. 2004. The Multilingual and Multi-Orthographic Taiwan-Based Internet: Creative Uses of Writing Systems on College-Affiliated BBSs. *Journal of Computer-mediated Communication* 9.1. http://www.ascusc.org/jcmc/

Torres i Vilatarsana, Marta. 2001. Funciones pragmáticas de los emoticonos en los chats. Interlingüística 11.

Torres i Vilatarsana, Marta. 1999. Els xats: entre l'oralitat i l'escriptura. Article publicat a la revista Els Marges, 65 (desembre, 1999). Publicat a Internet (gener, 2001) amb el consentiment d'aquesta revista.

UNESCO. 2003. *Cultural and Linguistic Diversity in the Information Society.* UNESCO publications for the World Summit on the Information Society. CI.2003/WS/07 http://unesdoc.UNESCO.org/images/0013/ 001329/132965e.pdf

Unicode Consortium. 1991. *The Unicode Standard: Worldwide Character Encoding.* Reading, Mass., Addison-Wesley Pub.

Unicode Consortium. 1996. *The Unicode Standard, Version 2.0.* Reading, Mass., Addison-Wesley Developers Press.

Unicode Consortium. 2000. *The Unicode Standard, Version 3.0.* Reading, Mass., Addison-Wesley.

Unicode Consortium. 2003. *The Unicode Standard, Version 4.0.* Reading, Mass., Addison-Wesley.

Warschauer, M., G. R. El Said and A. Zohry. 2002. Language Choice Online: Globalization and Identity in Egypt. *Journal of Computer-Mediated Communication* (JCMC), 7.4. http://www.ascusc.org/jcmc/

Wasserman, Herman. 2002. Between the local and the global: Souoth African languages and the Internet. *Litnet Seminar Room.* http://www.litnet.co.za/seminarroom/11wasserman.asp

Wright, S. 2004. Introduction. *International Journal on Multicultural Societies,* Vol.6,No.1, pp. 3-11.

Wurm, S. A.. 1991. Language death and disappearance: causes and circumstances. In R. H. Robbins and E. M. Uhlenbeck, eds., *Endangered Languages,* 1-18. Oxford: Berg.

Wurm, S. A., ed. 1996. *Atlas of the World's Languages in Danger of Disappearing.* Paris: UNESCO Publishing/Pacific Linguistics.

4.

Perspectives alternatives

a. Diversité linguistique sur Internet : une perspective asiatique

Yoshiki Mikami[*], Ahamed Zaki abu Bakar[●], Virach Sonlertlamvanich[○], Om Vikas[■], Zavarsky Pavol[*], Mohd Zaidi abdul Rozan[*], Göndri Nagy János[▲], Tomoe Takahashi[*]

(Membres du Projet d'observatoire des langues (LOP), Agence de la science et de la technologie du Japon)

« Avant de terminer cette lettre, j'aimerais souligner respectueusement à Son Éminence le fait que durant plusieurs années, j'ai voulu consulter dans cette Province des livres imprimés dans la langue et l'alphabet du pays, comme c'est le cas à Malabar et étant d'un grand intérêt pour la communauté chrétienne. Malheureusement, ce fut impossible pour deux raisons : tout d'abord parce qu'il semblait impossible de couler autant de moules, plus de six cents en tout, comparativement

* Université de la technologie de Nagaoka, JAPON : ● Université de la technologie de la Malaisie, MALAISIE : ○ Laboratoire de linguistique informatique thaï, THAÏLANDE : ■ Service de la technologie des langues indiennes (TDIL), Ministère des technologies de l'information, INDE : ▲ Université de Miskolc, HONGRIE. On peut contacter les auteurs à l'adresse de courriel : mikami@kjs.nagaokaut.ac.jp.

à seulement vingt-quatre comme c'est le cas en Europe »… Lettre d'un jésuite à Rome (Priolkar, 1958).

« Lorsque Gutenberg imprima sa fameuse Bible à Mainz il y a plus de 500 cents ans, il n'eut besoin que d'un caractère de base pour chaque lettre de l'alphabet. En comparaison, quand la mission américaine imprima la bible arabe à Beyrouth en 1849, au moins 900 caractères furent requis – et même ce nombre s'avéra insuffisant »… John M. Munro, 1981 (Lunde, 1981).

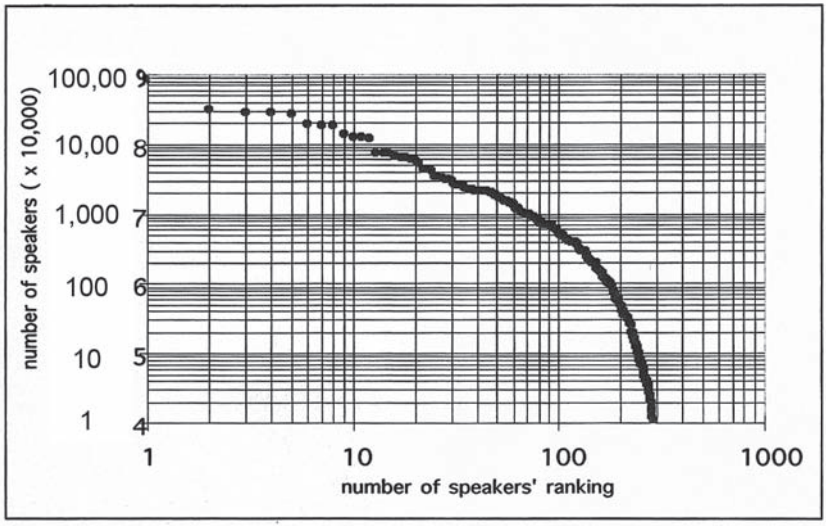
Diversité linguistique et écrite en Asie

Les experts linguistiques estiment qu'environ 7 000 langues sont parlées à travers le monde à l'heure actuelle (Gordon, 2005). Le nombre de langues officielles reste élevé et pourrait se chiffrer à plus de 300. Le Haut Commissariat des Nations Unies aux droits de l'homme (UNHCHR) a traduit un texte d'une valeur universelle, la Déclaration universelle des droits de l'homme (UDHR), dans plus de 328 langues différentes (UNHCHR, 2005).

De toutes les langues apparaissant sur ce site, le chinois est la plus parlée (presque un milliard de personnes), suivi de l'anglais, du russe, de l'arabe, de l'espagnol, du bengali, de l'hindi, du portugais, de l'indonésien et du japonais. La liste des langues inclut celles parlées par moins de cent mille personnes. Les langues asiatiques sont parmi les plus parlées au monde (six des dix langues les plus parlées) et constituent presque la moitié (48) des cent langues les plus parlées.

Le site de l'UNHCHR fournit aussi le nombre approximatif de groupes linguistiques pour chaque langue. Si l'on répartit les langues selon la population et que l'on fait le relevé de chaque langue sur un tableau à échelle logarithmique, le rapport entre la population et son rang ressemble à une courbe de la Loi de Zip comme l'indique la Figure 1, avec un écart d'au moins un dixième à un centième.

Figure 1: Quasi courbe de la Loi de Zip pour les groupes linguistiques



La diversité en Asie est plus évidente si l'on observe les multiples caractères d'écriture servant à représenter la langue. Cette diversité des caractères d'écriture s'avère problématique dès que l'on s'interroge sur la complexité de la localisation linguistique. La réponse à la question « Combien de caractères d'écriture sont utilisés à travers le monde ? » est difficile, puisqu'elle dépend d'un comptage très fragmenté. Aux fins de simplifier le présent article, on traitera en une seule catégorie tous les caractères d'écriture et alphabets dérivés du latin de même que leurs développements dans différentes langues européennes, le vietnamien, le philippin, etc. De même, les caractères d'écriture cyrilliques et arabes constituent une seule catégorie. C'est aussi le cas pour les idéogrammes chinois, les syllabiques japonais et le caractère hangul coréen. Les autres caractères d'écriture sont très diversifiés. Prenons le cas des « caractères d'écriture indic » qui forment la cinquième catégorie. Celle-ci comprend non seulement les caractères de langues indiennes telles le devanagari, bengali, tamoul, gujarâtî, etc., mais aussi quatre autres caractères d'écriture majeurs de l'Asie du Sud-est, c'est-à-dire le thaï, le lao, le cambodgien (khmer) et le myanmar. En dépit de leurs formes différentes, ces caractères d'écriture ont tous la même origine (l'ancien caractère brahmi) et le même type de formulation. Le regroupement de la population parlant chacune de ces langues

est présenté au Tableau 1. Les caractères d'écriture utilisés en Asie s'étendent à ces cinq catégories, tandis que ceux utilisés ailleurs dans le monde sont surtout d'origine latine, cyrillique, arabe et plusieurs autres.

Tableau 1. Distribution de la population selon les principales catégories de caractères d'écriture

Caractère d'écriture	Latin	Cyrillique	Arabe	Hanzi	Indic	Autres*
Nombre d'usagers (millions)	2 238	451	462	1 085	807	129
[% du total]	[43,28%]	[8,71%]	[8,93%]	[20,98%]	[15,61%]	[2,49%]

* Autres inclut grec, géorgien, arménien, amharique, dhivehi, hébreu, etc.

Statut actuel de la couverture linguistique – le cas de Windows

Depuis une décennie, les produits des technologies de l'information et de la communication (TIC) actuels peuvent accomplir des tâches multilingues dans une certaine mesure. Grâce à l'émergence d'une norme de code de caractères multilingues sous forme d'ISO/IEC 10646, aussi utilisée pour la norme Unicode, de même que pour le déploiement international de logiciels, le nombre de langues supportées par des plateformes majeures de bureau TIC a augmenté au cours de la dernière décennie. La couverture linguistique de ces plateformes majeures reste cependant limitée. La plus récente version de Windows XP (Professional SP2) est en mesure de traiter une liste longue de 123 langues. Toutefois, l'examen attentif de cette liste indique qu'il s'agit pour la plupart de langues européennes et très peu sont asiatiques ou africaines. La couverture linguistique est résumée au Tableau 2. Ce tableau fait voir les langues catégorisées selon le regroupement des caractères d'imprimerie présentés dans la première partie de cet article. En ce sens, la couverture par Windows XP est d'environ 83,72 % de la population globale. Même si ce chiffre semble acceptable, il apparaît être une estimation exagérée ne reflétant pas la réalité, comme on le constatera dans le présent article.

Tableau 2. Couverture linguistique de Windows XP SP2 pour les principales catégories de caractères d'écriture

Région linguis.	Latin	Cyrillique	Arabe	Hanzi	Indic	Autres
Europe	langues europ.* et slaves**	langues russe, macédonienne et slaves***	—	—	—	grec géorgien arménien
Asie	azéri vietnamien malaisien indonésien uzbek turc	mongolien azéri kazakh kirghiz uzbek	arabe urdu perse	chinois japonais coréen	gujarâtî tamoul telugu kannara bengali malayalam punjabi hindi marathe sanskrit konkani oriya thaï	assyrien dhivehi hébreu

* Inclut : albanais, allemand, anglais, basque, catalan, danois, espagnol, estonien, féringien, finlandais, français, galicien, gallois, hollandais, hongrois, islandais, italien, letton, lithuanien, maltais, norvégien, portugais, roumain, sami et suédois.

** Inclut : bosniaque, croate, polonais, serbe, slovaque, slovène, tchèque.

*** Inclut : bélarusse, bosniaque, bulgare, serbe et ukrainien.

Le cas de Google

Les moteurs de recherche sont des composantes indispensables de la société d'information globale. Ils permettent d'accéder à une vaste réserve de connaissances. Si l'on examine la couverture linguistique des moteurs de recherche les plus populaires, on constate que la situation est encore plus déplorable que pour la couverture linguistique de Windows. En date d'avril 2005, Google, qui est l'un des moteurs de recherche multilingue utilisé à l'échelle internationale, avait indexé plus de huit milliards de pages rédigées dans différentes langues. Toutefois, les langues recensées jusqu'ici se limitent à environ 35. Parmi celles-ci, on ne retrouve que sept langues asiatiques, notamment l'indonésien, l'arabe, le chinois classique, le chinois simplifié, le japonais, le coréen et l'hébreu (Tableau 3). En termes de

couverture linguistique, cette proportion diminue à 61,37 % surtout parce que les pages en langues asiatiques et africaines ne peuvent faire l'objet de recherches.

Tableau 3. Couverture linguistique de Google pour les principales catégories de caractères d'écriture

Région linguist.	Latin	Cyrillique	Arabe	Hanzi	Indic	Autre
Europe	langues europ.* et slaves**	russe bulgare serbe	—	—	—	grec
Asie	indonésien		arabe	chinois tradition-nel et simpl., japonais coréen		hébreu turc

* Inclut : allemand, anglais, catalan, danois, espagnol, estonien, finlandais, français, hollandais, hongrois, islandais, italien, letton, lithuanien, norvégien, portugais, roumain, suédois.
** Inclut : croate, polonais, slovaque, slovène et tchèque.

Le cas du corpus multilingue UDHR

Voici un autre exemple. Tel que mentionné dans la première partie de cet article, la Déclaration universelle des droits de l'homme (UDHR) sur le site Web du Haut Commissariat des Nations Unies aux droits de l'homme (UNHCHR) est affichée dans plus de 300 langues différentes, commençant par l'abkhaze et se terminant par le zoulou. Malheureusement, on constate aussi que beaucoup de ces traductions, surtout celles dans les langues à caractères d'écriture ne dérivant pas du latin, sont affichées en formats « GIF » ou « PDF », plutôt que sous la forme de texte codé. À nouveau, nous résumons la situation dans le Tableau 4 en regroupant les principaux caractères d'imprimerie comme dans les tableaux précédents. Ce tableau indique clairement que les langues dérivées des caractères d'imprimerie latins sont surtout présentées sous la forme de textes codés. Les langues ayant recours aux caractères d'imprimerie non dérivées du latin, surtout l'indic et autres, peuvent difficilement être présentées sous forme codée. Quand le caractère d'imprimerie n'est pas présenté sous l'une des trois formes indiquées, il

est regroupé dans la colonne « Non disponible ». De plus, il faut absolument télécharger des fontes spéciales pour lire correctement ces caractères d'imprimerie. Cette situation difficile peut être décrite comme une fracture numérique parmi les langues, ou qualifiée de « fracture numérique linguistique ».

Tableau 4. Formes de représentation du corpus multilingue UDHR par regroupement des principaux caractères d'imprimerie

Forme de présentation	Latin	Cyril.	Arabe	Hanzi	Indic	Autres
Codé	253	10	1	3	0	1
PDF	2	4	2	0	7	10
Image (GIF)	1	3	7	0	12	7
Non disponible	0	0	0	0	1*	1*

* Les langues non disponibles sont le magadi et le bhojpuri.

Localisation linguistique des technologies de l'information – un regard historique

Retournons cinq siècles en arrière, à l'époque de l'invention de la presse à imprimer. Cette technologie fut inventée séparément dans les pays de l'Est et de l'Ouest. À l'Est, des artisans coréens découvrirent cette technologie au 13e siècle et furent suivis des Chinois. Mais la technologie ne devint pas populaire et fut remplacée par la xylographie. L'origine directe des technologies d'impression maintenant répandues en Asie remonte donc à celle inventée par Gutenberg au milieu du 15e siècle.

La première presse à imprimer fut amenée à Goa en 1556. On croit aussi qu'il s'agit de la première presse en Asie. Cet appareil se retrouva par la suite dans d'autres régions asiatiques, notamment Manille, Malacca, Macau, etc. À l'origine, ces machines servaient surtout à imprimer des textes religieux traduits ou translittérés à l'aide du latin, mais servirent plus tard à imprimer différents textes à l'aide de types de caractères locaux. Selon un historien indien, Doctrina Christiana fut le premier texte imprimé en Asie comportant des caractères locaux tamouls.

La deuxième page du texte nous indique l'approche retenue pour la localisation linguistique de la technologie en caractères d'impression tamoule. Bien que cette langue comporte environ 246 syllabes en tout, des échantillons de types de caractères relevés seulement à la deuxième page du livre en contiennent plus de cent cinquante. Un jésuite en mission sur la côte du Malabar au 17e siècle écrivit une lettre à Rome, dans laquelle il déplorait « j'essaie depuis longtemps d'imprimer des textes dans la langue et les caractères d'imprimerie locaux, mais sans succès. Ceci est attribuable au fait que nous devons forger plus de 600 types de caractères différents ici sur les côtes du Malabar, plutôt que seulement 24 comme c'est le cas à Rome » (Priolkar, 1958).

Doctrina fut traduite en langue tagalog en 1593 à Manille, alors le centre des activités coloniales espagnoles de l'époque. Il semble toutefois que cette traduction s'accompagna aussi de translittération. La version actuelle de la Doctrina en tagalog s'appuya sur trois approches : la langue tagalog avec caractères d'imprimerie tagalog ; la langue tagalog avec caractères d'imprimerie latins ; et la langue espagnole avec caractères d'imprimerie latins. Dans le siècle qui suivit l'introduction de la technologie d'impression à Manille, la première approche a complètement disparu au profit des deux dernières approches. Et finalement, les caractères d'imprimerie tagalog furent complètement oubliés, même parmi la population locale (Hernandez, 1996). Un timbre-poste émis par le service postal des Philippines en 1995 représente le caractère d'imprimerie tagalog comme un motif de leur héritage culturel maintenant disparu.

Ces deux faits historiques nous enseignent que lorsque la localisation linguistique n'est pas effectuée de manière convenable, l'émergence de la nouvelle technologie risque de détruire le système d'écriture de la culture elle-même.

Normes de codage comme pierre angulaire de la localisation linguistique

Cette division est certainement attribuable à plusieurs facteurs, de nature économique, politique, sociale, etc. Mais d'un point de vue technique, la localisation linguistique devrait être le principal facteur. Tel que clairement énoncé dans la lettre du jésuite à Rome, écrite il y a quatre siècles (et citée en exergue à la première page de cet article), les pionniers des technologies de l'information à l'ère de la typographie devaient surmonter des difficultés semblables par nature

à celles rencontrées de nos jours par les ingénieurs en informatique qui doivent effectuer la localisation linguistique des technologies pour différents caractères d'écriture. Le principal obstacle des langues utilisant des caractères d'écriture non latins est certes le manque (ou l'absence) de disponibilité des normes de codage appropriées. C'est la raison pour laquelle les créateurs du site Web UDHR doivent convertir le texte non encodable en format PDF ou en images. Si l'on se réfère aux répertoires internationaux reconnus de séquences de codages, comme le IANA Registry of character codes (IANA, 2005) ou le ISO International Registry of Escape Sequences (IPSJ/ITSCJ, 2004), on ne peut trouver aucune séquence de codage pour ces langues pouvant avoir « passé à travers les mailles du filet ». Il est à noter que beaucoup de normes de codage de caractères établies au niveau national se retrouvent aussi dans plusieurs langues. Ces normes sont identifiées comme étant nationales. Concernant la famille de systèmes d'écriture indiens, la première norme nationale indienne fut annoncée en 1983 et appelée Indian Standard Script Code pour l' Information Interchange (ISSCII). Par la suite en 1991, elle fut amendée et devint la deuxième version (norme nationale IS 13194) utilisée à l'heure actuelle en Inde. Cependant, bien qu'il existe des normes nationales, des vendeurs de matériel informatique, des développeurs de polices de caractères et même des usagers ont créé leurs propres tableaux de codes de caractères, ce qui entraîne inévitablement une situation chaotique. La création de ces supposées séquences de codage exotique ou de codage interne local fut particulièrement favorisée par la popularité des outils de développement conviviaux de polices de caractères. Bien que les systèmes d'application dans ces domaines ne soient pas autonomes et soient largement diffusés sur le Web, la nécessité d'une standardisation n'a pas fait l'objet d'une attention sérieuse de la part des usagers, vendeurs et développeurs de polices de caractères. Cette situation chaotique s'explique aussi par l'absence d'associations professionnelles et d'organismes de réglementation gouvernementale. Aruna Rohra et Ananda of Saora Inc., ont préparé une étude intéressante (voir : http://www.gse.uci.edu/markw/languages.html), qui a recueilli des documents linguistiques de langues indiennes. L'étude a découvert 15 séquences de codage différentes sur les 49 sites Web tamoul visités (Aruna et Ananda, 2005).

UCS/Unicode

La première version du Universal Multiple-Octet Coded Character Set (UCS, ISO/IEC 10646) fut publiée en 1993. L'Unicode, initialement mis au point à

titre de consortium industriel, est maintenant synchronisé à la révision de UCS. Il s'agit réellement d'un effort valable pour éliminer les situations chaotiques. Mais il n'a pas encore acquis un statut dominant, du moins en Asie. Notre plus récente étude révèle que la pénétration du codage UTF-8 est limitée à seulement 8,35 % de toutes les pages Web sous ccTLD asiatique (Mikami et autres, 2005). Les dix premiers et les dix derniers ccTLDs sont indiqués au Tableau 5. Même si l'on prévoit que la vitesse de migration sera élevée, le processus doit être étroitement surveillé.

Tableau 5. Ratio d'usage UTF-8 des pages Web par ccTLD

CcTLD	nom	ratio	ccTLD	nom	ratio
Tj	Tadjikistan	92,75 %	uz	Ouzbékistan	0,00 %
Vn	Vietnam	72,58 %	tm	Turkménistan	0,00 %
Np	Népal	70,33 %	sy	Syrie	0,00 %
Ir	Iran	51,30 %	mv	Maldives	0,00 %
Tp	Timor oriental	49,40 %	la	Lao	0,01 %
Bd	Bangladesh	46,54 %	ye	Yémen	0,05 %
Kw	Koweït	36,82 %	mm	Myanmar	0,07 %
Ae	États Arabes Unis	35,66 %	ps	Palestine	0,12 %
Lk	Sri Lanka	34,79 %	bn	Brunei	0,36 %
Ph	Philippines	20,72 %	kg	Kirghizstan	0,37 %

Source : Projet d'observatoire des langues.

Projet d'observatoire des langues - Objectifs

Le Projet d'observatoire des langues (LOP) fut créé en 2003 (UNESCO, 2004) afin de reconnaître l'importance de surveiller le niveau d'activité linguistique dans l'espace cybernétique. On prévoit que le Projet d'observatoire des langues sera un outil pour évaluer le niveau d'usage de chaque langue sur le Web. De

façon plus spécifique, le projet devrait fournir périodiquement un profil statistique des langues, caractères d'écriture et séquences de codage dans l'espace cybernétique. Lorsque cet observatoire sera pleinement fonctionnel, on sera en mesure de répondre aux questions suivantes : combien de langues différentes retrouve-t-on dans l'univers virtuel ? Quelles langues sont absentes de cet univers virtuel ? Combien de pages Web sont rédigées dans une langue donnée, par exemple le pashto ? Combien de pages Web sont rédigées en caractères d'écriture tamoule ? Quels types de séquences de codage sont utilisés pour le codage d'une langue donnée, par exemple le berbère ? À quelle vitesse Unicode remplace-t-il les séquences de codage conventionnelles et développées localement sur Internet ? En plus de recueillir ces informations, on prévoit que le projet fera une proposition pour corriger la situation actuelle, tant au niveau technique que des politiques.

Projet Alliance

À l'heure actuelle, plusieurs groupes d'experts collaborent à l'Observatoire des langues à l'échelle internationale. Les organisations fondatrices incluent : l'Université de la technologie de Nagaoka au Japon ; l'Université des études étrangères de Tokyo au Japon ; l'Université Keio au Japon ; l'Université de la technologie de la Malaisie, en Malaisie ; l'Université Miskolc en Hongrie ; le projet de développement technologique des langues indiennes relevant du ministère indien des technologies de l'information ; ainsi que le Laboratoire de recherche en communications de la Thaïlande. Le projet est financé par l'Agence japonaise de science et de technologie, en vertu du programme RISTEX (RISTEX, 2005). L'UNESCO appuie officiellement le projet depuis sa création. Parmi les principales composantes techniques de l'Observatoire des langues, on retrouve une puissante technologie de robot Web *(Web crawler)* ainsi qu'une technologie d'identification des propriétés linguistiques (Suzuki et autres, 2002). La technologie de robot Web, appelée UbiCrawler (Boldi et autres, 2004), est extensible et entièrement distribuée grâce aux efforts conjoints de développement du département des sciences informatiques de l'Université de Milan ainsi que de l'Institut d'informatique et de télématique du Conseil de recherche national italien. Cette technologie constitue un puissant moteur de collecte de données pour l'observatoire des langues. Pour de brèves descriptions des efforts conjoints du LOP et de l'équipe UbiCrawler, voir UNESCO WebWorld News, 23 fév. 2004 (UNESCO, 2004).

Conclusion

Dans cet article, nous avons souligné l'importance de surveiller dans l'espace cybernétique le comportement et les activités des différentes langues parlées à travers le monde. Le Projet d'observatoire des langues (LOP) permet une méthode perfectionnée pour comprendre et surveiller les langues. Le consortium LOP veut contribuer à sensibiliser davantage le monde entier aux langues existantes et celles en voie de disparition, et appliquer aussi des mesures préventives dans ce dernier cas. Pour que ces efforts réussissent, l'Observatoire se veut aussi le point central de développement du capital humain, tout autant que le dépositaire des différentes ressources linguistiques. L'accumulation de ces ressources numériques par la recherche et le développement aidera les pays en voie de développement ainsi que les communautés régionales à acquérir la capacité et l'habileté requises pour faire migrer leurs langues autochtones dans l'espace cybernétique, en vue d'éviter la disparition de leur héritage national.

Références

Aruna, R. & Ananda, P. 2005. Collecting Language Corpora: Indian Languages. *The Second Language Observatory Work Shop Proceedings.* Tokyo University of Foreign Studies, Tokyo.

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. 2004. UbiCrawler: A scalable fully distributed web crawler. *Software: Practice & Experience,* Vol. 34, No. 8, pp.711-726.

Gordon, R. 2005. *Ethnologue: Languages of the World 15th Edition.* (http://www.ethnologue.com/)

Hernandez, Vincente S. 1996. *History of Books and Libraries in the Philippines:* Manila, The National Commission for Culture and the Arts, pp. 24-31.

IANA. 2005. *Character Sets.* (http://www.iana.org/assignments/character-sets)

IPSJ/ITSCJ. 2004. *International Register of Coded Character ets to be used with Escape Sequences. (*http://www.itscj.ipsj.or.jp/ISO-IR/)

Mikami, Y., Zavarsky, P., Zaidi, M., Rozan, A., Suzuki, I., ?akahashi, M., Maki, T., Ayob, I.N., Boldi, P., Santini, M. & Vigna, S. 2005. The Language Observatory Project (LOP). P*roceedings of the Fourteenth International World Wide eb Conference,* May 2005. Chiba, Japan. pp.990-991.

Lunde. P. 1981. *Arabic and the Art of Printing.* Saudi, Aramco World.

Priolkar, A. K. 1958. *The Printing Press in India - Its Beginning and Early Development.* Bombay, Marathi Samshodhana Mandala. pp.13-14.

RISTEX. 2005. (http://www.ristex.jp/english/top_e.html)

Suzuki, I., Mikami, Y., Ohsato, A. & Chubachi, Y. 2002. A language and character set determination method based on N-gram statistics, ACM *Transactions on Asian Language Information Processing*, Vol. 1, No. 3, pp.270-279.

UNESCO. 2004. Parcourir le cyberespace à la recherche de la diversité linguistique. UNESCO WebWorld News, 23rd Feb. 2004. (http://portal.UNESCO.org/ci/en/ev.php-URL_ID=14480&URL_DO=DO_TOPIC&URL_SECTION=201.html)

UNHCHR. 2005. *Universal Declaration of Human Rights.* (http://www.unhchr.ch/udhr/navigate/alpha.htm)

b. Une note sur les langues africaines sur la Toile mondiale

Xavier Fantognan

Aperçu

Les Cahiers du RFAL n° 23 « Traitement informatique des langues africaines » soulignent que le nombre de langues africaines est estimé à environ 2000, qui représente un tiers des langues du monde. C'est donc un patrimoine et une richesse qui méritent qu'on y prête attention. Aujourd'hui, le cyberespace peut permettre à toutes les langues de participer d'être de véritables instruments de communication à grande échelle. Cependant, toutes les langues du monde ne font pas usage et ne profitent pas de l'opportunité que représente cet espace. Bien évidemment pour y accéder, il faut avoir fait l'objet d'un traitement informatique, traitement qui relève de l'aménagement linguistique. Dès lors, la première question que l'on se pose ici se rapporte à l'utilisation des langues africaines dans le cyberespace. Marcel Diki-Kidiri et Edema Atibakwa, dans « Les langues africaines sur la Toile », explorent plus de 3 000 sites pour ne retenir que ceux qui traitent des langues africaines. De leur analyse, on retient qu'il existe bien une abondante documentation sur les langues africaines sur la Toile, mais très peu de sites utilisent une langue africaine comme langue de communication. Bien que de nombreux facteurs puissent être pris en compte pour expliquer cet état des faits, deux facteurs dominants seraient l'inexistence de cybercommunautés linguistiques capables d'intensifier leurs échanges dans leurs langues via la Toile et l'absence d'un traitement informatique concluant des langues africaines.

Cette conclusion sera modérée, nuancée, voire corrigée par une étude différente faite par Gilles Maurice de Schryver et Anneleen Van der Veken, « Les langues africaines sur la Toile : étude des cas haoussa, somali, lingala et isixhosa ». Ces auteurs ont exploré plutôt les forums de discussion pour y découvrir un taux d'utilisation tout à fait satisfaisant de trois langues africaines largement diffusées : le kiswahili, le hausa et le lingala.

Les principaux enseignements qu'on peut retenir de l'étude du RIFAL sont les suivants :

— Les langues africaines apparaissent sur la Toile beaucoup plus comme des objets d'étude (mention, documentation, description, échantillons, textes, cours) que comme des véhicules de communication ;

— La langue de communication utilisée pour parler des langues africaines est très largement l'anglais, même pour les langues en zone francophone ;

— Les cours de langues africaines sont beaucoup trop rares sur la Toile. Ce qui entrave la possibilité de développer des cybercommunautés de locuteurs utilisant les langues africaines comme véhicules de communication via l'Internet ;

— Les produits logiciels ou les solutions informatiques intégrant en standard des polices de caractères pour toutes les langues africaines sont rarement proposés sur les sites.

Pour corriger cette situation, il y a donc lieu de promouvoir :

— la multiplication des sites bilingues (ou multilingues) comportant le français ou l'anglais et au moins une langue africaine comme langues de communication ;

— une plus grande diffusion de la documentation sur les langues africaines, car cette documentation existe mais n'est pas systématiquement diffusée sur la Toile ;

— les cours de langues africaines de qualité à diffuser sur la Toile ;

— le développement et la diffusion de produits logiciels ou de solutions informatiques facilitant l'écriture des langues africaines et leur utilisation normale et courante dans le cyberespace.

Nous ne pouvons plus dire aujourd'hui que les langues africaines ne sont pas présentes sur la Toile mondiale. Il existe beaucoup de documentations sur les

langues africaines sur la Toile mais très peu de textes sont écrits en langues africaines et pourquoi ? Le manque de motivations parmi les Africains à écrire dans leur propre langue est une des raisons que l'on peut citer pour expliquer le relatif insuccès des langues africaines sur la Toile. Le cybernaute qui s'exprime sur la Toile veut être lu et compris, il va donc écrire dans une langue connue par le plus grand nombre de gens.

En effet, une grande partie des textes en langues africaines trouvés sur la Toile n'a pas été écrit par des Africains, comme nombre de documents religieux ou de textes destinés à l'enseignement. Des forums où des Africains communiquent avec d'autres Africains, en langues africaines, sont l'exception et non la règle.

Microsoft a annoncé que Windows et Office seront prochainement traduits en langage Swahili. Le Kiswahili est sans doute la langue la plus parlée d'Afrique. Près de 100 millions de personnes parlent cette langue, en Afrique et dans les îles de l'Océan Indien. Avant de passer à la traduction proprement dite, les linguistes de Microsoft devront établir un glossaire commun aux différents dialectes issus du Kiswahili. Microsoft prévoit aussi de traduire ses logiciels dans d'autres langues africaines, notamment les langues Hausa et Yoruba.

Si les intentions de Microsoft semblent bonnes, il est tout de même inquiétant de constater que les logiciels de Microsoft seront la seule alternative des Swahili qui ne parlent pas d'autres langues. En effet, les logiciels libres traduits en Kiswahili ne sont pas légions. Espérons que les efforts de Microsoft pour la standardisation des langues africaines profiteront aussi à Linux et aux logiciels libres.

Dans ce dernier cas, celui des logiciels libres, un travail considérable est en cours en Afrique. Au Burkina-Faso, les langues comme le mooré, le dioula connaissent une localisation avec Open Office. Le même travail est en cours au Mali avec le bambara, au Bénin avec le fongbé, le yoruba, le mina et le dendi. Le formidable travail élaboré avec l'amharique et son alphabet illustre de la possibilité de rendre plus efficace la recherche sur l'informatisation des langues africaines. La démarche de UNICODE pour la standardisation de l'alphabet N'ko réconforte plus d'un.

Cependant, de véritables questions restent posées à savoir que les questions orthographiques et la normalisation des langues africaines ne sont pas encore

résolues. Beaucoup de langues sont toujours transcrites phonétiquement et le risque de voir chaque langue disposer de son alphabet n'est plus à écarter.

Si l'Afrique dispose de 2000 langues environ, seulement 400 environ d'entre elles ont été décrites. Il en reste 1600 qui n'ont pas bénéficié d'études sérieuses. Aucune de ces langues aujourd'hui n'a d'audience sur le Web pas plus les 400 qui ont connu une description mais qui souffrent d'enrichissement en vue de devenir de véritables langues vivantes sur la Toile mondiale.

Références

Diki-Kidiri M., Don D., Dimo-Lexis, Dictionnaires monolingues et Lexiques spécialisés, Outils logiciels pour linguiste, CNRS-LACITO, Paris.

Meloni H. ; 1996. Fondements et Perspectives en traitement automatique de la parole. AUPELF/UREF.

Morvan P. ; 2000. Dictionnaire de l'Informatique : Acteurs concepts, réseaux, Larousse, Paris.

Peek J., Lui C., et al ; 1997. Système d'information sur Internet : Installation et mise en œuvre, Editions O'Reilly International Thomson.

Rint-Riofil, C., Chanard, et Diki-Kidiri, M. (hors date) Stage de formation niveau1 et 3, Document de travail : Introduction aux inforoutes par le développement de la terminologie et des contenus textuels pour le français et les langues partenaires, Lumigny, Marseilles.

Gilles Maurice de Schryver et Anneleen Van der Veken ; 2003. Le traitement informatique des langues africaines, Cahiers du RIFAL, Revue coéditée par l'Agence de la francophonie et la Communauté française de Belgique.

Présentation des Auteurs

Xavier Fantognon est un étudiant en linguistique togolais de l'Université du Bénin (xavier@bj.refer.org) qui a décidé de se consacrer à la mise en valeur des langues africaines sur l'Internet. Il a traduit l'interface de la plate forme libre SPIP en langue Fongbé (http://www.spip.net/fon) et s'engage également sur le front des activités culturelles traditionnelles ou en forme de multimédia.

Yoshiki Mikami est Professeur des Sciences du Management et de l'Information à l'Université Technologique de Nagaoka. Il a occupé des postes de direction au MITI (standards et politiques d'information). Il est responsable du projet d'Observatoire des Langues dans l'Internet (http://www.language-observatory.org/ - http://gii.nagaokaut.ac.jp/gii/- http://kjs.nagaokaut.ac.jp/mikami/).

John Paolillo est professeur associé en science de l'information et en techniques informatiques; Professeur associé adjoint en linguistique, School of Library and Information Science. Ph.D., Linguistics, Stanford University, 1992, B.A., Linguistics, Cornell University, 1986. Domaines de recherche : linguistique informatique, recherche d'information, communication assistée par ordinateur, modèles statistiques et méthodes quantitatives de recherche, sociolinguistique et acquis de langues, acquis en langues étrangères, langues de l'Asie du Sud.

Daniel Pimienta, français d'origine marocaine qui vit à Saint Domingue, est le Président de l'Association Réseaux & Développement (FUNREDES – http://funredes.org), une ONG qui travaille sur le terrain des TIC et développement depuis 1988. Funredes a conduit un certain nombre d'expérimentations sur le terrain en ce qui concerne les langues et les cultures, dans certains cas en collaboration avec l'Union Latine et/ou avec le soutien de l'Agence de la Francophonie. (http://funredes.org/tradauto/index.htm/bamaktxt - http://funredes.org/lc).

Daniel Prado, un argentin qui vit à Paris, est le Directeur du Programme de Terminologie et Industries de la Langue de l'Union Latine (http://unilat.org/dtil/), un organisme inter-gouvernemental de promotion des langues néolatines. Il gère des statistiques sur la réalité dynamique des langues dans notre société et des informations sur les politiques linguistiques et terminologiques.