Measuring Linguistic Diversity on the Internet

a technical and strategic debate

Papers presented by
John Paolillo
and
Daniel Pimienta

Edited with an introduction by
UNESCO Institute for Statistics

Contents

I - INTRODUCTION (UIS)

Background

Importance of languages

Origins of these papers. Sources of contributions and how they were brought together.

WSIS Tunis

2. MODELS AND APPROACHES

Linguistic Diversity in Cyberspace; models for development and measurement

Daniel Pimienta, Funredes

1 - 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 l'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 faites 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 assises sur l'expérience de terrain et avec l'intention de déranger pour ouvrir les esprits, pas pour le plaisir de déranger.

2 – 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[1] 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

1 Voir "La fracture numérique, un concept boiteux", Daniel Pimienta, 2002,
http://funredes.org/mistica/francais/cyberotheque/thematique/fra_doc_wsis1.html
et, "Travailler l'Internet avec une vision sociale", Communauté virtuelle MISTICA, 2002
http://funredes.org/mistica/francais/cyberotheque/thematique/fra_doc_olist2.html

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 sociale et que l'ordre des facteurs dépende des contextes. Néanmoins, 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.

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 détenir un accès sont multiples et peuvent également se présenter en forme de couches progressives: **- 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. 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).	**- 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. 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 : 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

2 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.

3 http://www.webopedia.com/TERM/A/ADSL.html

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.

5 http://www.itu.int/newsarchive/press_releases/2003/33.html

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
		(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 encouragement doit encore être réalisé. Un aspect linguistique, qui est maintenu considéré par l'ICANN[6], est celui des noms de domaine Internet dans toutes les langues[7].
	- 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[2] 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[3] (entre 10 y 50 US$ 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 du 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 mensuel pour les classes moyennes des pays en développement et une valeur de l'ordre de 1% pour les	**- 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.

6 http://www.webopedia.com/TERM/I/icann.html
7 http://en.wikipedia.org/wiki/IDNA
8 http://en.wikipedia.org/wiki/Unicode

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	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[5], " **l'accès universel** ». Mais, s'il s'agit d'une condition nécessaire pour résoudre la fracture numérique, et elle est très loin d'être une condition suffisante… **- alphabétisation fonctionnelle** Que la personne qui utilise l'infrastructure ait la capacité fonctionnelle de lire et écrire dans sa langue. Il s'agit probablement de la seconde fracture qu'il faut résoudre quand on prétend offrir, par exemple, "l'Internet pour tous". **- numérisation de l'alphabet** Que la langue maternelle de la personne qui va utiliser l'infrastructure puisse se prêter à un traitement informatique. Pour cela il faut qu'elle existe en forme écrite et que les caractères de son alphabet soient convenablement codifiés. Ce n'est malheureusement pas le cas pour la majorité des langues encore en usage.	**- alphabétisation fonctionnelle** Il n'est certes pas exclu de tirer parti de la possibilité multimédia des TIC pour adapter des interfaces permettant un certain nombre de possibilités aux personnes analphabètes. Cependant, il faut se rendre à l'évidence si c'est d'accès à la connaissance qu'il s'agit et non d'accès simplement 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** 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[8] et doivent être maintenus et amplifiés.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
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 et cela nous conduit au concept d'**alphabétisation numérique** (en anglais, "digital literacy"). Une éducation à l'espace numérique qui ne soit pas un simple entraînement à l'utilisation de certains programmes d'ordinateurs mais qui inclut une vision holistique des considérations et impacts sociétaux[9] de l'utilisations des TIC pour le développement est sans aucun doute le nœud le plus difficile à résoudre et l'élément à la fois le plus importante 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…	**alphabétisation numérique** L'effort formidable nécessaire pour une éducation numérique 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.
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 lunette, une technologie optique	Comment rendre transparente la technologie si son accès

9 Impact politique, économique, social, culturel, linguistique, organisationnel, éthique, info-écologique, biologique, psychologique…

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	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 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.	demande de passer par une langue autre que la langue maternelle? Ce niveau renforce clairement les arguments avancés pour les niveaux précédents.
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 **production** (de **contenus** par exemple) et de **création** (de **communautés virtuelles** par exemple). Il est clair qu'une conscience d'une finalité de développement est demandée et doit être motivée par les efforts d'éducation.	Le thème linguistique est essentiel dans ce niveau et renvoi à la possibilité et la motivation à produire des contenus et des communautés virtuelles localisés. Il pose aussi clairement la question du multilinguisme et de la nécessité de dispositif de passerelles entre les langues.
APPROPRIATION	*Quand la personne qui utilise est suffisamment habile pour*	

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
SOCIALE	*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.
"**EMPOWERMENT**"[10]	*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	Que signifie "l'innovation" et comment se manifeste-t-elle dans chaque culture ?

10 Ce mot anglais rassemble à la fois les sens de recevoir et prendre la capacité et la notion de prise de pouvoir à travers la capacité.

Niveau d'usage	Description des usages et des obstacles	Questions concernant les langues
	de l'innovation, en particulier sociale (nouvelles formes d'organisation, réponses nouvelles à problèmes connus…).	
DEVELOPPEMENT HUMAIN	*Quand les options de libertés individuelles et collectives s'ouvrent à la personne ou la communauté et peuvent s'exercer en forme de "**capacités**".*[11] 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 communautés a été une réalité dans tous 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 ?

11 "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. ", Amartya Sen - http://www.fas.harvard.edu/~freedoms/.

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[12]. Le croisement de cette discipline avec la question de la diversité culturelle et linguistique ouvre des perspectives et des réflexions tout à fait relevantes de notre débat. Un congrès a été consacré à ce thème en 2004[13] par l'ICIE (International Center on 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.

Parmi celles-ci, Charles Ess[14] 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[15]. 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

12 http://www.unesco.org/webworld/news/infoethics.shtm

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

14 Ess, Charles. 2004. Moral Imperatives for Life in an Intercultural Global Village. In Robert Cavalier (ed.), The Internet and Our Moral Lives, 161-193. Albany, New York: State University of New York Press.

_____. 2005a. Can the Local Reshape the Global? Ethical Imperatives for Humane Intercultural Communication Online. In J. Frühbauer, R. Capurro and T. Hausmanninger (Eds.), Localizing the Internet. Ethical Aspects in an Intercultural Perspective.

_____. 2006b. 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, Charles and Fay Sudweeks. 2005. Introduction: Culture and Computer-Mediated Communication – Toward New Understandings, Journal of Computer-Mediated Communication Vol. 11, No. 1: October, 2005. < http://jcmc.indiana.edu/>

15 L'exemple cité par Charles Ess est celui des centres d'apprentissage d'Afrique du Sud prévus pour enseigner aux peuples indigènes une utilisation effective des TIC : Postma, Louise. 2001. A Theoretical Argumentation And Evaluation of South African Learners' Orientation towards and Perceptions of the Empowering Use of Information. New Media and Society 3 (3: September), 315-28.)

pistes pour structurer un design des interactions homme-machine qui réponde à ce critère[16].

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 fondamentale 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 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[17] ("la langue qui était un trésor"[18]), le spécialiste espagnol du thème des langues et l'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 rechercher 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 et 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[19] ou similaires) des ressources linguistiques produites avec des fonds publiques ou qui font simplement partie du domaine publique.

En tout état de cause, les logiciels libres qui, par leur nature, devraient jouer un rôle particulièrement important dans le secteur linguistique n'y sont 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[20], comme

16 En particulier le travail d'Edward Hall et en particulier sa distinction entre des styles de communication de type "fort contexte/faible contenu" ou de type "faible contexte/fort contrenu": Hall, Edward T. 1976. Beyond Culture. New York: Anchor Books.
17 Voir le site http://jamillan.com, d'une très grande richesse et qualité esthétique.
18 Voir "La lengua que era un tesoro", http://jamillan.com/tesoro.htm et version en anglais résumé: http://jamillan.com/worth.htm
19 http://en.wikipedia.org/wiki/GNU_General_Public_License
20 Voir en particulier le travail de Jean Claude Guédon, "La bibliothèque virtuelle : une antinomie ? " http://sophia.univ-lyon2.fr/francophonie/doc/nlm-fr.html

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"[21] et la déclaration de Berlin sur l'accès ouvert au savoir dans les Sciences[22]. La diversité linguistique a tout a gagné 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.

21 http://plos.org
22 http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

3 - 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% !

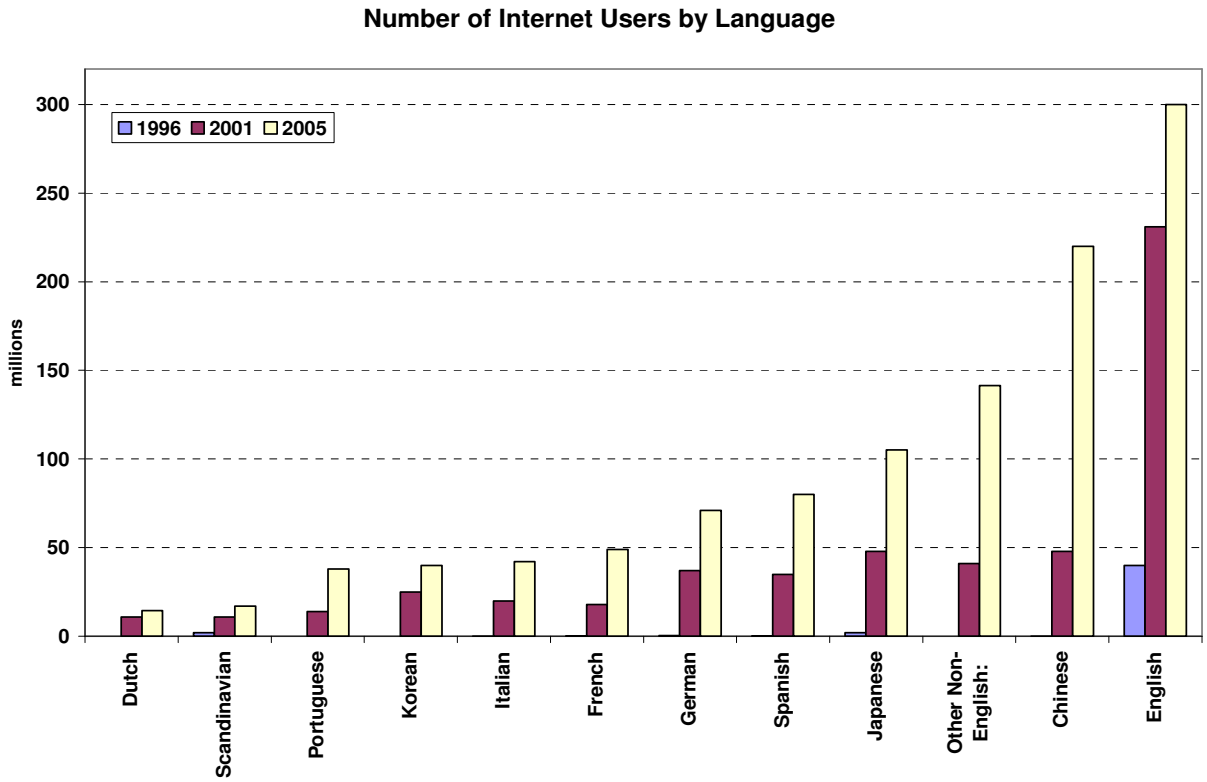
Il est urgent que l'académie reprenne son rôle dans cette affaire (ainsi que les institutions gouvernementales nationales et internationales) et les signes sont clairs que cette évolution est en cours, enfin! Pour s'en rendre compte, il faut consulter les présentations en ligne de la réunion organisée par l'Unesco (avec l'ACALAN et l'AIF) à Bamako sur le multilinguisme dans le cyberespace[23].

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

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[24] 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. 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.

23 http://portal.unesco.org/ci/en/ev.php-URL_ID=17688&URL_DO=DO_TOPIC&URL_SECTION=201.html
24 http://global-reach.biz/globstats/index.php3.

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



Source : Global Reach 2005

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

1) Celle qui 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) Celle qui avait été lancée par une des premières études sur le sujet, qu'Alis Technologies a réalisé 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 qui semble être la référence sur laquelle s'appuie de nombreux auteurs et médias pour continuer d'avancer une valeur de plus de 70% pour les pages web en anglais. La méthode consiste à créer un échantillon de quelques milliers de site web par le jeu du hasard sur les adresses IP[25], 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.

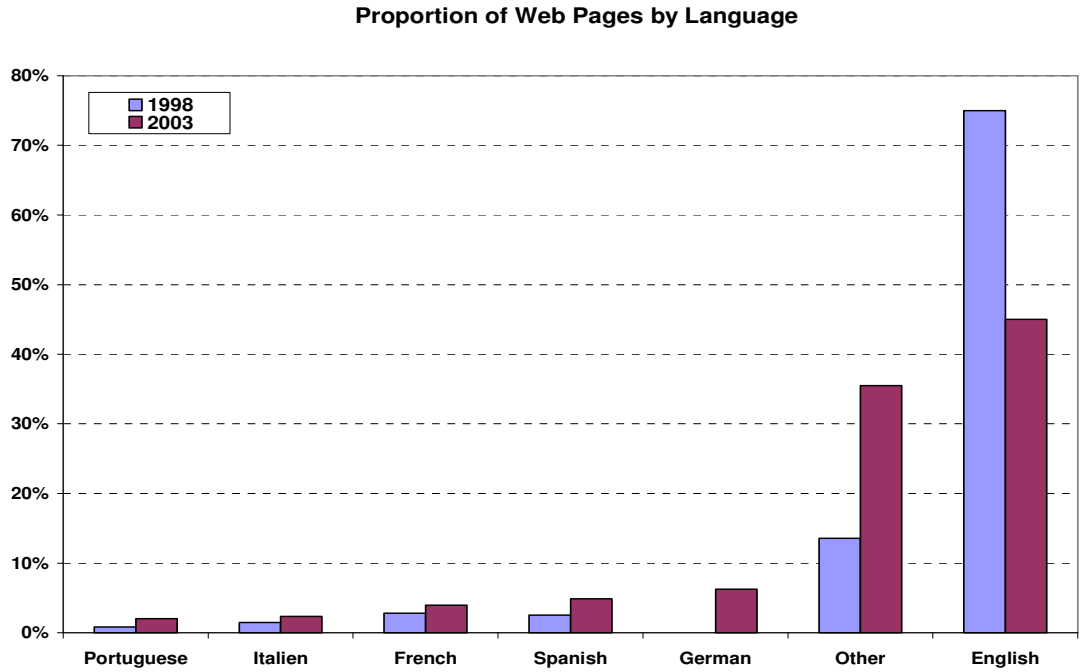
25 http://en.wikipedia.org/wiki/IP_address

La seconde limitation nous préoccupe beaucoup plus et 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 8,000 millions de pages indexées par Google ?). A travers le peu de documentation publié il semble bien 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[26]). Parmi elles, l'approche que FUNREDES et l'Union Latine ont utilisée[27] depuis 1996 (voir http://funredes.org/lc) .

Figure 2 : Proportion de Pages web composées dans une langue donnée



Source: FUNREDES 2003

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

26 http://arxiv.org/ftp/cs/papers/0006/0006032.pdf
27 A deux reprises avec le concours de l'Agence de la Francophonie.

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 considérées 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 page web dans une des langues travaillées[28] par rapport à l'anglais mais pas de valeur absolue. Pour les 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[29] et ne prend pas en compte le web invisible[30].
- Mais surtout elle est très dépendante des possibilités de comptage fiable qu'offrent les moteurs de recherche[31], ce qui à terme et très bientôt[32] risque de la disqualifier puisque aussi bien les moteurs prennent de plus en plus de liberté avec le traitement de la recherche par mot.[33]

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[34] et surtout, en bénéficiant des techniques de recherche par pays et par domaine, de produire une série d'indicateurs originaux et très signifiants[35].

PERSPECTIVES POUR DE NOUVELLES APPROCHES

28 Allemand, espagnol, français, italien, portugais et roumain.
29 Mais quelle "existence" ont réellement les pages non indexées ?
30 Voir http://www.brightplanet.com/technology/deepweb.asp
31 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.
32 Une dernière étude complète est sur le point d'être publiée à http://funredes.org/lc et il se pourrait bien que cela soit la dernière, en tous cas avec cette méthode.
33 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.
34 Elle a également permis une première approximation certes grossière mais intéressante sur le plan des évolutions de la présence des cultures dans l'Internet.
35 Voir, pour l'espagnol : http://www.funredes.org/LC/L5/valladolid.html
et pour le français : http://smsi.francophonie.org/IMG/pdf/lc-franco2003.pdf

Le projet maintenant avancé d'Observatoire des Langues (voir l'article de Yoshiki Mikami, plus loin) porte de nombreux espoirs pour occuper ce vide et apporter les réponses dont les faiseurs de politiques 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 leur apporter de précieux renseignements[36]. 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 de relevance pour les indicateurs comme : langues de lecture et écriture des courriels, langues des sites visités, etc.

36 Voir par exemple http://www.alexa.com/data/details/traffic_details?q=&url=unesco.org

Références démo-linguistiques et sur la mesure de la diversité linguistique

- Étude d'Alis Technologies - 1997: http://babel.alis.com/palmares.html
- OCLC Web characterization project - 1998: http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm
- Étude "webmap" de Inktomi – 2001 : http://www.inktomi.com/webmap/[37]
- Etude Vilaweb citée par eMarketer - 2001: http://www.emarketer.com/analysis/edemographics/ 20010227_edemo.htm[38]
- "Concordancing the Web with KWiCFinder", William H. Fletcher - 2001 http://miniappolis.com/KWiCFinder/FletcherCLLT2001.pdf
- Estimation of English and non-English Language Use on the WWW, Gregory Grefenstette & Julien Nioche – 2001 : http://arxiv.org/ftp/cs/papers/0006/0006032.pdf
- "Trends in the Evolution of the Public Web: 1998 - 2002", Edward O'Neill T, Brian F. Lavoie, and Rick Bennett - 2003: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html .
- Cybermetrics – 2005 : http://www.cindoc.csic.es/cybermetrics/cybermetrics.html
- Caslon Analytics – 2005 : http://caslon.com.au/metricsguide6.htm#stats
- Études de Funredes - 1996-2005: http://funredes.org/LC

Données démo-linguistiques et démographiques au sujet de l'Internet

- Sondage de l'Université Georgia Tech - 1998 : http://www.gvu.gatech.edu/user_surveys/
- Sondages Internet de NUA - 2003: http://www.nua.com/surveys/
- Ethnologue, les langues du monde - 2005: http://www.ethnologue.com/
- Network Wizards Internet Domain Surveys – 2005 : http://www.isc.org/index.pl?/ops/ds/
- Global Internet Statistics (by Language) - 2005: http://global-reach.biz/globstats/index.php3
- Internet World Stats - 2005: http://www.internetworldstats.com/stats7.htm

Références linguistiquement diverses sur la diversité linguistique dans l'Internet

En français

- "Le désir de France, la présence internationale de la France et la francophonie dans la société de l'information", Député Patrick Bloche, Rapport au Premier ministre - 1998: http://www.internet.gouv.fr/francais/textesref/rapbloche98/accueil.htm
- "L'inforoute en français, un portrait québécois", Réjean Roy- 1998: http://obelix.uqss.uquebec.ca/est/9906-15a.pdf
- "La défense de la Francophonie et de la Langue Française sur Internet.", Sachs F. - 1998: http://perso.club-internet.fr/fsachs/memoire.html
- "Le multilinguisme sur le Web", Marie Lebert - 1999 :

37 La page n'existant plus ; il est possible de la retrouver dans les archives de l'Internet : http://web.archive.org/web/20010124051100/http://www.inktomi.com/webmap/

38 Idem que pour la note 24: http://web.archive.org/web/20020608210415/http://www.emarketer.com/analysis/edemographics/20010227_edemo.html

http://www.cefrio.qc.ca/projets/Documents/multi0.htm

- "L´occitan sur Internet: signe des temps, champ du cygne ou pied de nez? ", Alén Mª Carmen y Henri Boyer, Lengas. Revue de Sociolinguistique, Université Paul-Valéry, MontpellierIII, nº 46, 1999, p. 21-31.
- "Les langues officielles sur Internet", Commissariat aux Langues Officielles - 2002 http://www.ocol-lo.gc.ca/archives/sst_es/2002/lang_internet/lang_internet_2002_f.htm
- "Le développement des TI en francophonie: où en sommes-nous cinq ans après la conférence de Montréal? - État des lieux, bilan et prospective", Affaires étrangères du Canada – 2002 http://www.dfait-maeci.gc.ca/foreign_policy/francophonie/rapport_ti_francophonie_2002-fr.asp
- "Le français sur Internet au coeur de l'identité canadienne et de l'économie du savoir", Alain Clavet - 2002 : http://www.ocol-clo.gc.ca/archives/sst_es/2002/internet_id_can/internet_2002_f.htm

En espagnol et autres langues d'Espagne

- "Novas tecnoloxías e usos lingüísticos", Daniel Romero, Isabel Vaquero - 1999 http://www.galego21.org/nos/redada/ticeusoslgcos.pdf
- "Lingua e informática. O galego na Rede." - 1999 http://www.galego21.org/nos/redada/linguainformatica.pdf
- "As linguas na Rede", Daniel Romero, Isabel Vaquero: - 1999 http://www.galego21.org/nos/redada/amesa99.pdf
- "Caracterizando la web chilena", Ricardo Baeza Yates, Carlos Castillo, 2000: http://www.todocl.cl/stats/jun2000/wcl2000.pdf
- "El libro de mil millón de páginas: la ecología lingüística de la Web", José Antonio Millán – 2000: http://jamillan.com/ecoling.htm
- "El español en la sociedad de la información", Daniel Martín Mayorga – 2000: http://cvc.cervantes.es/obref/anuario/anuario_00/martin/
- "La lengua española en internet", Francisco A. Marcos Marín - 2000 http://cvc.cervantes.es/obref/anuario/anuario_00/marcos/
- 10 retos para la red en español", José Antonio Millán - 2002 http://jamillan.com/conclusio.htm
- "Las lenguas románicas en Internet: propuestas para el diseño de un itinerario básico", Juana Castaño Ruiz – 2002: http://www.um.es/tonosdigital/znum4/corpora/indicecorpora.htm
- "Internet, en el centro de la estrategia lingüística", José Antonio Millán – 2002: http://jamillan.com/estratlin.htm
- "La salut del català a Internet", Jordi Mas i Hernàndez – 2003 : http://www.softcatala.org/articles/article26.htm
- Llengua catalana: Informe de política lingüística, la Secretaria de Política Lingüística, 2003: http://www6.gencat.net/llengcat/informe/[39]
- "A lingua galega en Internet", Gómez Guinovart, X. – 2003: http://webs.uvigo.es/sli/arquivos/internet.pdf
- "Euskararen presentzia Interneten neurtu nahian", Iñaki Alegria, M Jesus Rodríguez – 2003:

39 Voir spécialement le chapitre XI LAS NUEVAS TECNOLOGÍAS E INTERNET

http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1070449758/publikoak/BAT.pdf
- Baròmetre de l'ús del català a Internet" – 2005: http://www.wiccac.org/webscat.html

En portugais
- "Diversidade cultural e direito à comunicação", Tadao Takahashi, 2004: http://www.campus-oei.org/pensariberoamerica/ric06a05.htm
- "Characterizing a National Community Web", Daniel Gomes, Mario J. Silva – 2005: http://xldb.fc.ul.pt/data/Publications_attach/gomesCharacterizing.pdf

Autres espaces
- "Language and the Internet", David Crystal, Cambridge University Press, 2001, reprint 2002 ISBN 052180212 1
- "The Latvian Language in the Internet and Resources of Computer Linguistics", A. Spektors. - In: The Latvian Language-Existence, Environment, Context, 1997, Riga: PBLA, pp. 46-53 (in Latvian).
- "Domain Names: when is it going to be in Arabic? ", Abdulaziz H. Al-Zoman, Raed I. Al-Fayez, Anas M. Asiri, SaudiNIC – 2001: http://www.saudinic.net.sa/arabicdomain/PAPERAMAN2.pps
- "Language choice online: Globalization and identity in Egypt", Warschauer, M., El Said, G., & Zohry, A. 2002: http://www.ascusc.org/jcmc/vol7/issue4/warschauer.html
- "Indigenous Language Presence on the Web – the Maori Example", Keegan, Te Taka, Cunningham, Dr. Sally Jo – 2002: http://www.cs.waikato.ac.nz/~tetaka/IndLangPres.pdf
- "Languages.com: The Internet and linguistic pluralism", Warschauer, M. – 2002: http://www.gse.uci.edu/markw/languages.html
- Das Internet spricht Englisch ... und neuerdings auch Deutsch, 2002: http://www.netz-tipp.de/sprachen.html
- "The Language of the Internet: English Dominance or Heteroglossia? ", Susan C. Herring - 2002: http://ella.slis.indiana.edu/%7Eherring/CATaC.ppt
- "Globalization and Indian languages", B. Mallikarjun – 2003: http://www.languageinindia.com/feb2003/globalization.html
- "Linguistic Diversity and the Digital Divide", Jens Allwood- 2003: http://computing.open.ac.uk/sites/SCALLA2004/Abstracts/Allwood.pdf\
- The Multilingual Internet: Language, Culture and Communication in Instant Messaging, E-mail and Chat, Journal of CMC, Special Issue – 2003: http://jcmc.indiana.edu/vol9/issue1/
- “La Nueva Geografía y las cifras de la Sociedad de la Información”, Sebastián Cáceres, Fundación Auna, 2004: http://www.fundacionauna.com/areas/28_observatorio/obser_01_10.asp
- Wikipedia: WikiProyecto Lenguas del Mundo – 2004: http://es.wikipedia.org/wiki/Wikipedia%3AWikiProyecto_Lenguas_del_mundo
- "Working Group on Internet Governance Draft Issue Paper on Cultural and Linguistic Diversity" – 2005: http://www.globalcn.org/en/article.ntd?id=2195&sort=1.7
- "Local Languages, RSS and the Digital Divide", Andy Carvin – 2005:

http://www.andycarvin.com/archives/digital_divide/index.html

- "Languages on the Internet", Sue Wright – 2005
  http://portal.unesco.org/ci/en/ev.php-URL_ID=18255&URL_DO=DO_TOPIC&URL_SECTION=201.html
- "Multilingualism on the Internet", Brenda Danet, Susan Herring - 2005
  http://pluto.mscc.huji.ac.il/~msdanet/papers/multiling.pdf
- Une page sur les langues mayas – 2005 : http://www.enlacequiche.org.gt/

Pages de liens vers des sites sur la diversité linguistique dans l'Internet

http://www.bisharat.net/links.htm
http://www.dfki.de/~gor/bookmarks/hypertextmediawww_multilingualityinternationalisation.html
http://www.portalingua.info/es/statistiques/langues_internet/1/10/index.php?
http://www.portalingua.info/es/statistiques/demographie/1/index.php
http://www.europarl.eu.int/stoa/publi/99-12-01/part1_en.htm
http://www.terralingua.org/EndLangResources.html
http://portal.unesco.org/ci/en/ev.php-URL_ID=6628&URL_DO=DO_TOPIC&URL_SECTION=201.html
http://faculty.ed.umuc.edu/~jmatthew/websites.html
http://www.terralingua.org/AddResourcesmainpage.htm

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 (voir Hamel[40]).

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, dans l'enseignement, etc.

En novembre 2002, le premier Congrès international sur la place des langues néolatines dans la communication spécialisée[41] 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 tiers des pays de la planète (27,53% selon Calvet[42]) 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 donnés internationales les plus importantes[43]. 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. Selon le Cindoc[44], les trois langues néolatines les mieux 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

40 Hamel, Rainer Enrique. "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. (http://unilat.org/dtil/cong_com_esp/comunicaciones_es/hamel.htm#a)

41 Congrès international sur les langues néolatines dans la communication spécialisée (http://www.unilat.org/dtil/cong_com_esp/es/index.htm)

42 Calvet, Louis-Jean (2002). Le marché aux langues. París: Plon.

43 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. A voir….

44 CINDOC 1998, 1999 ADD FULL REFERENCES. Citado por Hamel

porte bien, avec 18,9 % de la production mondiale[45], mais c'est la littérature qui est concernée majoritairement par ce chiffre [46].

Bien entendu, si l'on compare à la situation de 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[47] en langues latines[48] ; le français est la deuxième langue d'usage international ; l'espagnol prend une confortable 3e 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 2 fois et demie moins de locuteurs que l'ensemble des locuteurs latins a 2 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 1 publication scientifique sur 10.

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[49]. 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 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 édités en au moins une langue latine, près de 45 % le sont en anglais. Même l'allemand, avec 10 fois moins de locuteurs, avait à peine 2 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

45 Graddol, Informe del British Council, 1997. Cité par Louis-Jean Rousseau

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

47 Funredes / Union latine /Agence intergouvernementale de la Francophonie. Etude sur la présence des langues latines sur l'Internet (http://www.unilat.org/dtil/LI/2003_2005.htm)

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

49 Nilsson, Henrik. "Perte de domaine, perte de fonctionnalité : indicateurs et enjeux" au Lexipraxi 2005 (http://www.ailf.asso.fr/presentation.htm)

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 pour le SMSI[50] tombe à point nommé. En effet, dans le 1er 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 développées 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

50 Unesco. Plan d'action du SMSI (http://portal.unesco.org/ci/fr/ev.php-URL_ID=15897&URL_DO=DO_TOPIC&URL_SECTION=201.html)

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.[51]

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[52], l'Université des Nations Unies[53] 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 (voir Millán[54]) 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 étant 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

51 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.

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

53 Projet UNL (http://www-clips.imag.fr/projets/unl/)

54 Millán, José Antonio. "A quoi bon un projet européen concurrent ?" dans Courrier International (http://www.courrierint.com/article.asp?obj_id=51004&provenance=hebdo), traduction française d'un article paru dans El País

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.

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 en 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[55]. 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[56] et il est d'espérer que des solutions aux problèmes soulevés seront trouvées.

55 Recommandations du Congrès international sur les langues néolatines dans la communication spécialisée (http://unilat.org/dtil/cong_com_esp/es/recomendacion.htm)

56 Initiative réunissant quatre organisations internationales (CPLP, OIF, OEI et l'Union latine) afin de proposer des actions conjointes pour la francophonie, l'hispanophonie et la lusophonie (http://www.3el.org)

3. Language Diversity on the Internet: Examining Linguistic Bias

John Paolillo, School of Informatics, Indiana University

With

Elijah Wright and Hong Zhang, School Of Library And Information Science, Indiana University

S. Baskaran, S. Ramanan, S. Rameshkumar, L. Shiba Nair, J. Vinoshabu and S. Viswanathan, Natural Language Processing Group, KBC Research Center, Anna University, Chennai (Madras), India.

4 – Alternative perspectives

Language Diversity On The Internet: An Asian View

Yoshiki Mikami*, Ahamed Zaki abu Bakar**, Virach Sonlertlamvanich***, Om Vikas****
Zavarsky Pavol*, Mohd Zaidi abdul Rozan*, Göndri Nagy János*****, Tomoe Takahashi*

Members of the Language Observatory Project (LOP), Japan Science and Technology Agency (JST)

> "Before I end this letter I wish to bring before Your Paternity's mind the fact that for many years I very strongly desired to see in this Province some books printed in the language and alphabet of the land, as there are in Malabar with great benefit for that Christian community. And this could not be achieved for two reasons; the first because it looked impossible to cast so many moulds amounting to six hundred, whilst as our twenty-four in Europe." ... A Jesuit Friar's letter to Rome, 1608 [1]
>
> "Gutenberg, when he set his famous Bible in Mainz more than 500 years ago, only needed one basic piece of type for each letter of the alphabet, while in 1849, when the American Mission Press in Beirut printed an Arabic Bible, no less than 900 characters were used - and even this number was felt to be insufficient." ... John M. Munro, 1981 [2]

Language and Script Diversity in Asia

Language experts estimate that nearly 7,000 languages are spoken the globe today [3]. In terms of official languages, the number of languages is still large and could be more than three hundred. The United Nations Higher Commission for Human Rights (UNHCHR) has translated a text of universal value, the Universal Declaration of Human Rights (UDHR), into as many as 328 different languages. These translated texts can be viewed by visiting the UNHCHR website [4].

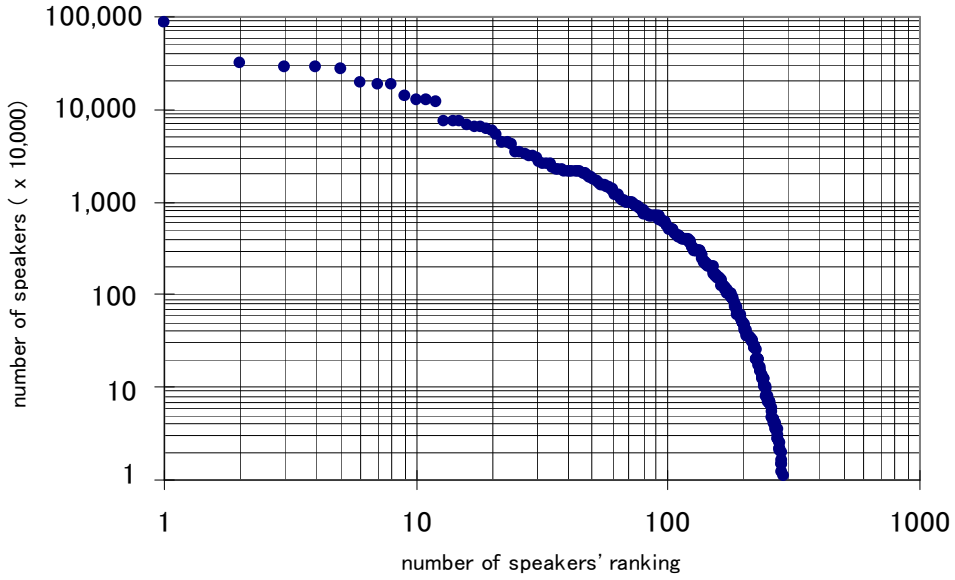
Among all the languages appearing in this site, Chinese has the biggest speaking population of almost a billion, and is followed by English, Russian, Arabic, Spanish, Bengali, Hindi, Portuguese, Indonesian and Japanese. The language list continues until those languages with less than a hundred thousand speakers. Asian languages occupy 6 out of the top ten languages, and almost a half (48) of the top hundred languages.

The UNHCHR site also provides the estimated speaking population of each language. When we sort out languages by speaking population as the key and plot each language

∗Nagaoka University of Technology, JAPAN: **Universiti Tekmologi Malaysia, MALAYSIA: ***Thai Computational Linguistic Laboratory, THAILAND: ****Technology Department of Indian Languages (TDIL), Ministry of Information Technology, INDIA: *****Miskolc University, HUNGARY. Authors can be contacted by mail to mikami@kjs.nagaokaut.ac.jp

in a logarithmic scale chart, then the relationship between speaker population and its ranking emerges as something like a Zipf's-Law curve as shown in Figure 1 with at least at range between tenth to hundredth.

Figure 1: Quasi Zipf's Law Curve of Language Speakers



The diversity in Asia is more evident when we look at the diversity of scripts used to represent languages. From the viewpoint of complexity in localization, diversity of scripts is more problematic issue. "How many scripts are used in the world" is a difficult question to answer as it depends on granule size of counting. In this paper, for the sake of simplicity, we treat all Latin based scripts, alphabets plus its extensions used for various European languages, Vietnamese, Pilipino, etc. as one category. We will also treat Cyrillic based scripts as one and so on for Arabic based script. In the same nature, we will treat Chinese ideograms, Japanese syllabics and Korean Hangul script as one. The remaining scripts are comprised of various kinds of scripts. Here, we will take the "Indic script" to be in the fifth category. This category includes not only Indian language scripts such as Devanagari, Bengali, Tamil, Gujarati, etc. but also four Southeast Asian major language scripts; Thai, Lao, Cambodian (Khmer) and Myanmar. In spite of the differences in their shapes, these scripts have the same origin (the ancient Brahmi script) and have the same type of behaviors in formulation. When we summed up the speaking population of each language by this script grouping, the number of users of each script is summarized in Table 1. Then scripts used in Asia extend to all five categories of scripts, while scripts used in the rest of the world is mostly Latin, Cyrillic, Arabic and several others.

Table 1. Distribution of User Population by Major Script Categories

Script	Latin	Cyrillic	Arabic	Hanzi	Indic	Others*
Number of users in million	2,238 [43.28%]	451 [8.71%]	462 [8.93%]	1,085 [20.98%]	807 [15.61%]	129 [2.49%]

[% of total]

*Others include Greek, Georgian, Armenian, Amharic, Dhivehi, Hebrew, etc.

Current Status of Language Coverage

A case of Windows

Compared to a decade ago, current ICT products are capable of handling multilingualism to a certain degree. Thanks to the emergence of multilingual character code standard in the form of ISO/IEC 10646 which is also used for the Unicode standard, as well as sophisticated internationalization of software, the number of languages being supported by major ICT desktop platforms have increased during the last decade. The language coverage of those major platforms, however, is still limited. The most recent version of Windows XP (Professional SP2) is able to handle a long list of 123 languages. However, if we look at the list more closely, most of the languages are for European languages and very few of which are Asian and African languages. The language coverage is summarized in Table 2. In this table, languages are categorized by the script grouping introduced in the first section of this paper. Hence, the population-based coverage of Windows XP is calculated to be around 83.72% against the global population. This is not a bad figure, but as we will discussed later in this paper, this figure seems to be an overestimated figure which does not tally with reality.

Table 2. Windows XP SP 2 Coverage on Language by Major Script Categories

Script Region	Latin	Cyril	Arabic	Hanzi	Indic	Other
Europe	**European* & Slavic Languages****	**Russian, Macedonian & Slavic languages*****	**---**	**---**	**----**	**Greece Georgia Armenia**
Asia	**Azeri Vietnamese Malay Indonesian Uzbek Turkish**	**Mongolian Azeri Kazakh Kyrgyz Uzbek**	**Arabic Urdu Persian**	**Chinese Japanese Korean**	**Gujarati Tamil Telugu Kannada Bengali Malayalam Punjabi Hindi Marathi Sanskrit Konkani Thai**	**Assyrian Dhivehi Hebrew**

**Includes: Albanian, Basque, Catalan, Danish, Dutch, English, Estonian, Faroese, Finnish, French, Galician, German, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Romanian, Sami, Spanish, Swedish and Welsh.*

***Includes: Serbian, Czech, Croatian, Slovak, Bosnian, Polish & Slovenian*

****Includes: Belarusian, Bulgarian, Serbian, Bosnian & Ukrainian*

A Case of Google

Search engines are indispensable components of the global information society. Vast pool of knowledge can be made accessible through the function of search engines. When we investigate the language coverage of popular search engines, the situation is far worse compared to the case of the Windows' language coverage. One of the globally used multilingual search engine, Google, is found as of April 2005, to have indexed more than 8 billion pages written in various languages. However, the languages covered so far is limited to only some 35 languages. Among these, the Asian languages covered by Google are only seven: Indonesian, Arabic, Chinese Traditional, Chinese Simplified, Japanese, Korean and Hebrew (Table 3). If we calculate the population-based coverage, it will decreases to 61.37% mainly because Asian and African language pages are not searchable.

Table 3. Google Coverage on Language by Major Script Categories

Script / Region	Latin	Cyril	Arabic	Hanzi	Indic	Other
Europe	**European* & Slavic Languages** ***	**Russian Bulgarian Serbian**	**---**	**---**	**---**	**Greece**
Asia	**Indonesian**		**Arabic**	**Traditional &Simplified Chinese Japanese Korean**		**Hebrew Turkish**

**Includes: Catalan, Danish, Dutch, English, Estonian, Finnish, French, German, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Norwegian, Portuguese, Romanian, Spanish and Swedish*

***Includes: Croatian, Czech, Polish, Slovak & Slovenian*

A Case of UDHR Multilingual Corpus

Let us present one more example. As mentioned in the first section of the paper, if we visit the website of the Office of the Higher Commissioner for Human Rights of the United Nations, we will find more than 300 different language versions of the Universal Declaration of Human Rights (UDHR) starting from Abkhaz and ending with Zulu. Unfortunately, we will also find many of the language translations, especially for non-Latin script based languages, are just posted as "GIF" or "PDF" files, not in the form of encoded text. We again summarized the situation by major script grouping as matching as the previous tables (Table 4). The table clearly shows that languages which use Latin scripts are mostly represented in the form of encoded texts. Languages which use non-Latin script especially Indic and other scripts on the other hand, are difficult to be represented in encoded form. When the script is not represented by any of the three foremost forms provided, they are grouped as not available. Moreover, it is compulsory to download special fonts to properly view these scripts. This difficult situation can be described as a digital divide among languages or termed as the 'language digital divide'.

Table 4. Form of Representation of the UDHR Multilingual Corpus by Major Script Grouping

Script Form of Presentation	Latin	Cyril	Arabic	Hanzi	Indic	Other
Encoded	253	10	1	3	0	1
PDF	2	4	2	0	7	10
Image (GIF)	1	3	7	0	12	7
Not available	0	0	0	0	1*	1*

*Not available languages are Magahi and Bhojpuri

IT Localization

A Historic Flashback

Let us look back five hundred years ago, when an epoch-making printing technology emerged. Type- printing technology was invented in the East and the West independently. In the East, the technology was first created by Korean artisans in the 13th century, and followed by Chinese. But the technology did not flourish later and was gradually replaced by xylography. The direct root of type-printing technologies now prevailing through Asia can be traced back to the one invented by Gutenberg in mid 15th century.

The first printing press machine was brought to Goa in 1556. This was believed to be the first printing machine brought to Asia as well. Later the machine was brought to other places in Asia, like Manila, Malacca, Macau, etc. Initially these machines were primarily used to print translated or transliterated religious texts using Latin letters but later they were used to print various text using local script typefaces. According to one Indian historian, the first printed text using local letters in Asia is Doctrina Christiana in Tamil. The second page of which tells us what kind of approach was employed in the localization of type-printing technology into Tamil language. Although Tamil language has some ten vowels and thirty plus consonants, sample typefaces shown on the second page of the book are more than hundred fifty in number. A Jesuit father stationed somewhere in Malabar coast in the 17th century wrote a letter to Rome and complained "I have long been trying to print texts by using local languages and scripts here, but have not succeeded yet. The main reason is that we must forge more than 600 typefaces here in Malabar coasts, instead of just 24 at home in Rome". [1]

In Manila, the central place of Spanish colonial activities at that time, Doctrina was translated into Tagalog language in 1593. But it happened that translation was accompanied by transliteration. Actually Tagalog version of Doctrina employed three approaches; Tagalog language by Tagalog script, Tagalog language by Latin script, and Spanish language by Latin script. And in the space of one hundred years after the introduction of type-printing technology into Manila, the latter two approaches had completely replaced the first one. Finally Tagalog script was totally forgotten even among local populations.[5] A mailing stamp issued by Philippines' post in 1995, depicts Tagalog script as a motif of their now lost cultural heritage.

Two historic episodes give us a lesson. When localization was not successfully done, emergence of new technology would even destroy the writing system or culture itself.

Encoding Standards as a Cornerstone of Localization

There are certainly many factors behind this divide; economical, political, social and etc. But among those, from technical viewpoint, localization should be the main factor. As is clearly stated in the Jesuit Friar's letter to Rome written four hundred years ago, quoted in the first page of this paper, even from the era of type-printing, pioneers of information technology had to overcome difficulty of similar nature when localizing technologies into different script users like as today's computer engineers do. Especially the lack or non-availability of appropriate encoding standards is the major obstacles especially in non-Latin script user communities. Due to this fact, the UDHR website creators have to put the text not able to be encoded but in the form of PDF or images. If we look at internationally recognized directories of encoding schemes, like the IANA Registry of character codes [6] or ISO International Registry of Escape Sequences [7] (ISO-IR), we can not find any encoding schemes for these languages which we termed as have fallen through the net. We must note that many character encoding standards that were established at the national level are also present for many languages. These standards are identified as National Standard. In the case of the family of Indian writing systems, the first national Indian standard was announced in 1983. It was named the Indian Standard Script Code for the Information Interchange (ISSCII). Later in 1991, it was amended to become the second version, national standard IS 13194 which is currently in use in India. However, although there exist national standards, hardware vendors, font developers and even end-users have been creating their own character code tables which inevitably lead to a chaotic situation. The creations of so called exotic encoding scheme or local internal encoding have been accelerated particularly through the introduction of user-friendly font development tools. Although the application systems working in these areas are not stand-alone systems and are published widely via the web, the necessity for standardization has not been given serious attentions by users, vendors and font developers. The non-existence of professional association and government standard bodies is another reason for this chaotic situation. Aruna Rohra of Saora Inc. has produced an interesting report while doing a study to collect the language corpora of Indian languages. It found 15 different encoding schemes from 49 Tamil web sites visited [8].

UCS/Unicode

The first version of the Universal Multiple-Octet Coded Character Set (UCS, ISO/IEC 10646) was published in 1993. The Unicode, initially born as an industrial consortium effort, has been now synchronized to the revision of UCS. It is really a strong drive to eliminate the chaotic situations. But still has not acquired a prevailing status at least in Asian part of the world. Our most recent study has disclosed that penetration of UTF-8 encoding is limited to only 8.35% of whole web pages under Asian ccTLDs [9]. Top ten ccTLDs and the least ten ccTLDs are shown in Table 5. Although migration speed is expected to be high, we need to monitor carefully the process.

Table 5. UTF-8 Usage Ratio of Web Pages by ccTLD

ccTLD	name	ratio	ccTLD	name	ratio
tj	Tajikistan	92.75%	uz	Uzbekistan	0.00%
vn	Viet Nam	72.58%	tm	Turkmenistan	0.00%
np	Nepal	70.33%	sy	Syria	0.00%

ir	Iran	51.30%	mv	Maldives	0.00%
tp	Timor East	49.40%	la	Lao	0.01%
bd	Bangladesh	46.54%	ye	Yemen	0.05%
kw	Kuwait	36.82%	mm	Myanmar	0.07%
ae	UAE	35.66%	ps	Palestine	0.12%
lk	Sri Lanka	34.79%	bn	Brunei	0.36%
ph	Philippines	20.72%	kg	Kyrgyzstan	0.37%

source: Language Observatory Project, [9]

The Language Observatory Project

Objectives

Recognizing the importance of monitoring language activities level in cyberspace, the Language Observatory Project (LOP) was launched in 2003 [10]. The Language Observatory Project is planned to provide means for assessing the usage level of each language in cyberspace. More specifically, the project is expected to periodically produce a statistical profile of language, scripts, encoding scheme usage in cyberspace. Once the observatory is fully functional, the following questions can be answered: How many different languages are found in the virtual universe? Which languages are missing in the virtual universe? How many web pages are written in any given language, say Pashto? How many web pages are written using the Tamil script? What kinds of character encoding schemes are employed to encode a given language, say Berber? How quickly is Unicode replacing the conventional and locally developed encoding schemes on the net? Along with such a survey, the project is expected to work on developing a proposal to overcome this situation both at a technical level and at a policy level.

Project Alliance

Currently, several groups of experts are collaborating on the world language observatory. Founding organizations include: Nagaoka University of Technology, Japan; Tokyo University of Foreign Studies, Japan; Keio University, Japan; Universiti Teknologi Malaysia, Malaysia; Miskolc University, Hungary; Technology Development of Indian Languages project under Indian Ministry of Information Technology and Communication Research Laboratory, Thailand. The project was funded by Japan Science and Technology Agency under RISTEX [11] program. UNESCO has given an official support to the project since its inception. Major technical components of the Language Observatory are basically powerful crawler technology and language property identification technology [12]. As for crawler technology, the UbiCrawler [13], a scalable, fully distributed web crawler developed by the joint efforts of the Dipartimento di Scienze dell'Informazione of the Università degli Studi di Milano and the Istituto di Informatica e Telematica of the Italian National Council of Research, is working as a powerful data collecting engine for the language observatory. Brief descriptions of the joint efforts of LOP and UbiCrawler team can be found in [10].

Conclusion

In this paper, we stress the importance of monitoring the behavior and activities of world languages in cyberspace. By having a monitoring body such as that performed by the

Language Observatory Project, a smarter method to understand the language scenario can be realized. The LOP consortium hope to make the world more aware of its living and dying languages. Steps to assist endangered languages can then be made before its extinction. For this effort to bear fruits, the observatory is also designed to be the focal point for human capital development as well as serves to accumulate various language resources. These digital resources accumulated through research and development as well as through other means will be the bridge to lessen the digital divide. They will assist developing countries and communities in the region to have the ability and capacity to get their indigenous language into cyberspace and hence preserves a national heritage from extinction.

References

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

[2] Paul Lunde, Arabic and the Art of Printing, Saudi Aramco World, March/April 1981.

[3] Ethnologue: Languages of the World 15th Edition, http://www.ethnologue.com/

[4] http://www.unhchr.ch/udhr/navigate/alpha.htm [accessed on 15/04/2005]

[5] Vincente S. Hernandez, History of Books and Libraries in the Philippines 1521-1900, pp.24-31, The National Commission for Culture and the arts, Manila, 1996.

[6] http://www.iana.org/assignments/character-sets

[7] http://www.itscj.ipsj.or.jp/ISO-IR/

[8] Aruna Rohra, et al (2005). Collecting Language Corpora: Indian Languages. The Second Language Observatory Work Shop Proceedings 21-25 Feb. 2005. Tokyo University of Foreign Studies: Japan

[9] Yoshiki Mikami, Pavol Zavarsky, Mohd Zaidi Abd Rozan, Izumi Suzuki, Masayuki Takahashi, Tomohide Maki, Irwan Nizan Ayob, Paolo Boldi, Massimo Santini, Sebastiano Vigna, The Language Observatory Project (LOP). Proceedings of the Fourteenth International World Wide Web Conference, pp.990-991, 10-14 May 2005, Chiba, JAPAN

[10] UNESCO WebWorld News, February 23, 2004, “Parcourir le cyberespace à la recherche de la diversité linguistique”, http://portal.unesco.org/ci/en/ev.php-URL_ID=14480&URL_DO= DO_TOPIC&URL_SECTION=201.html

[11] http://www.ristex.jp/english/top_e.html

[12] Izumi Suzuki, Yoshiki Mikami, Ario Ohsato, Yoshihide Chubachi, A language and character set determination method based on N-gram statistics, ACM Transactions on Asian Language Information Processing, 1(3), pp.270-279, 2002.

[13] Paolo Boldi, Bruno Codenotti, Massimo Santini and Sebastiano Vigna, UbiCrawler: A scalable fully distributed web crawler. Software: Practice & Experience, 34(8), pp.711-726, 2004.

Une Note Sur Les Langues Africaines Sur La Toile Mondial

Xavier Fantognan

Aperçu

Les Cahiers du RFAL n°23 « Traitement informatique des langues africaines » soulignent que le nombre des langues africaines est estimé à environ 2000, et cela 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 des 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'autre langue. 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 Fongbe, le Yoruba , le Mina et le Dendi. Le formidable travail élaboré avec l'Amharic 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 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

1- MARCEL DIKI-KIDIRI, DIDIER DON, Dimo-Lexis, Dictionnaires monolingues et Lexiques spécialisés, Outils logiciels pour linguiste, CNRS-LACITO,paris.
2- MELONI Henri ; 1996 : Fondements et Perspectives en traitement automatique de la parole. AUPELF/UREF, 168p.
3- MORVAN Pierre ; 2000 : Dictionnaire de l'Informatique : Acteurs concepts, réseaux, Larousse, 21 rue du Montparnasse 75006, Paris, 323p.
4- PEEK Jerry, LUI Cricket et al ; 1997 : Système d'information sur Internet : Installation et mise en œuvre, Editions O'REILLY INTERNATIONAL THOMSON, 730p.
5- RINT-RIOFIL, C. Chanard et M. Kiki-Kidiri, 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, Marseil-Luminy.
6- http://www.uji.es/serveis/slt/triam/triam15.html
7- http://www.wsis2005.org/bamako2002/
8- http://sango.free.fr/
9- http://fr.wikipedia.org
10- http://www.geocities.com/fon_is_fun/
11- http://portal.unesco.org/culture
12- http://www.unicode.org/
13- http://www.tavultesoft.com/keyman/
14- http://www.sil.org/
15- http://www.laterminologie.net/
16- http://www.systransoft.com/

I.2 - PRÉSENTATION DES AUTEURS

Adama Samassekou est le Président de l'Académie Africaine des Langues (ACALAN - http://acalan.org). Il a été Ministre de l'Education du Mali et Président de la première phase du processus du Sommet Mondial pour la Société de l'Information (http://itu.int/wsis).

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 Fongbè (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 directions 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/)

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 intergouvernemental 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.

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).